Étude de la symbiose dans le plancton marin par une approche transcriptome et méta-transcriptome

Arnaud Meng

Université Pierre et Marie Curie

École n° 515 : Complexité du Vivant et

École n° 227 : Sciences de la Nature & de l'Homme : écologie & évolution

Évolution Paris Seine - UMR 7138

Équipe Analyse des données à haut-débit en génomique

Étude de la symbiose dans le plancton marin par une approche transcriptome et méta-transcriptome

---

présentée et soutenue publiquement par

Arnaud Meng

le 15 décembre 2017

Directeurs de thèse : **Stéphane Le Crom** et **Fabrice Not**

Co-encadrante de thèse : **Lucie Bittner**

Composition du jury

Mme. Hélène Chiapello,	IR INRA, INRA Toulouse	Rapporteur
M. Didier Debroas,	PR, Université Clermont-Auvergne	Rapporteur
Mme. Laure Guillou,	DR2 CNRS, Station Biologique de Roscoff	Examinateur
Mme. Eve Toulza,	MCU, Université de Perpignan	Examinateur
Mme. Claire Lemaitre,	CR2 INIRIA, INRIA Rennes Bretagne Atlantique	Examinateur
M. Marc-André Sélosse,	PR MNHN, Muséum National d'Histoire Naturelle	Examinateur
M. Stéphane Le Crom,	PR, Université Pierre et Marie Curie	Co-directeur
M. Fabrice Not,	CR CNRS, Station Biologique de Roscoff	Co-directeur

"If there's one thing the history of evolution has taught us,
it's that life will not be contained.
Life breaks free.
It expands to new territories."

Jeff Goldblum, Jurassic Park (1993), scripted by Michael Crichton, David Koepp

Remerciements

Dans un premier temps, je souhaite présenter mes remerciements à l'Université Pierre et Marie Curie qui a financé ce travail de thèse au travers du programme interdisciplinaire Interface pour le Vivant (IPV).

Je remercie tout particulièrement les trois personnes qui m'ont accompagné de près dans cette longue (mais courte à la fois) et difficile (mais tellement enrichissante) expérience qu'est la thèse. Alors un grand MERCI à Lucie, Stéphane et Fabrice sans qui cette aventure n'aurait pu démarrer ! Merci de m'avoir fait confiance, en premier lieu dès notre entretien pour mon stage de M2 et puis tout au long de ces 3 dernières années. Je suis conscient de la chance que j'ai eu d'avoir pu travailler à vos côtés, ne serait-ce que pour tous les congrès et déplacements que vous m'avez encouragé à faire et qui ont très largement contribués à mon épanouissement scientifique. Je vous remercie également de m'avoir formé au monde de la recherche scientifique et pour toutes les connaissances que vous m'avez transmises avec patience !

Je tiens également à remercier tous les étudiants qui ont fait partie et qui font partie de l'équipe et avec qui j'ai partagé toutes ces journées dans la même pièce qu'est notre petit bureau. Merci à Jade, à Anita et Quentin que j'ai encadré du mieux que j'ai pu. MERCI à Anne-Sophie et Émile qui vont maintenant prendre le relais et devenir les « vieux étudiants » du labo ;) À vous deux en particulier, je vous souhaite tout le meilleur pour la suite de votre thèse. Je suis bien placé pour savoir que vous êtes parfaitement encadrés et que vous ferez de superbes thèses !

A tous les autres doctorants/Post-doc de l'unité, je vous remercie pour les super moments qu'on a partagé ! Raph, Jananan, Chloé, Romain, Andrews, Marguerite, Thomas, Gab, Camille et Juliette je n'oublierai pas les soirées SCEP, ni les midis d'organisation de la JDD, ni les nombreux regroupements au palace Auvinet :) Je garderai longtemps les nombreux souvenirs et débats qui ont animé nos discussions : « mange-t-on des araignées

en dormant ? », « et si on ouvrait un bar pour vendre des bières au plancton ?», sans oublier quelques mots-clefs qui font encore écho à des expériences étranges comme par exemple « gloast ».

Je tiens à remercier toutes les personnes que j'ai eu la chance de côtoyer que ce soit à l'UPMC ou en dehors. En particulier je pense à Gaëlle Lelandais et à Éric Pelletier qui ont accepté de participer à mes comités de thèse et qui m'ont aidé dans mes travaux grâce à leurs conseils !

Un petit coucou aussi aux gens de la station biologique de Roscoff que j'ai rencontré et qui m'ont chaleureusement accueilli au cours de mes séjours en Bretagne ! Je garde des super souvenirs des journées de travail chez vous au grand air et des journées qui se finissent au Ty Pierre.

Mention spéciale à Erwan qui est devenu au cours de ces 3 années et demi, mon mentor bioinfo avec qui j'ai adoré travailler et discuter et sans qui toutes ces analyses auraient pu être beaucoup plus difficiles ! Merci pour tout !

Merci aussi aux membres de la communauté de l'ann... des radiolaires ! J'ai énormément appris au cours de nos meetings annuels et j'ai pu découvrir une foule de choses sur les radiolaires grâce à vous !

Merci à mes amis bioinfo, les anciens du Master BI qui m'ont encouragé et avec qui nous nous sommes toujours serrés les coudes depuis le Master. Merci Valou, Alex, Thibault, Caro et Fabrice !

Un gros MERCI aux amis badistes qui m'ont permis de me vider la tête au bad et de garder le moral toutes ces années ! Merci Ninie, Sam, Gaëlle, Félix, Stéphanie, Kiti, David, Sophie, Sylvie et tous les autres bien sûr !

Je voudrais également dire un grand MERCI à ma famille qui m'a soutenue toutes ces années !

Et finalement, je remercie les membres du jury qui ont accepté de m'évaluer et à qui ces prochaines pages sont destinées !

Table des matières

Table des figures

Chapitre 1

Introduction

1.1 Le plancton

Le plancton se définit comme l'ensemble des organismes aquatiques à la dérive, ne pouvant lutter contre les courants. Le mot « plancton » vient du grec ancien πλανκτος qui signifie « errer » ou « dériver ». Cette dénomination a été utilisée pour la première fois en 1887 par le zoologiste allemand Christian Andreas Viktor Hensen pour désigner l'ensemble des « plus petits » organismes vivants ou morts flottants à la surface ou en profondeur d'eau douce ou salée. Sa définition a été rapidement adoptée par l'ensemble des naturalistes de son temps, pour évoluer ensuite en excluant de cette définition les organismes morts [Bone et L. Noble 2016].

Il existe une très grande diversité d'organismes planctoniques dans la colonne d'eau. Ces communautés, extrêmement diversifiées tant par leur morphologie, leur physiologie ou encore leur type trophique, sont composées d'organismes unicellulaires eucaryotes (protistes) et procaryotes (bactéries, archées), mais également par des éléments génétiques mobiles (*e.g.* virus, plasmides) et des eucaryotes multicellulaires (*e.g.* méduses, mollusques et crustacés). Le spectre de taille du plancton varie de plusieurs dizaines de centimètres (mégaplankton) à quelques dixièmes de micromètres (femtoplankton) lorsque l'on considère notamment les virus marins (Figure 1.1).

	taille	exemple
mégaplancton	20 - 200 cm	méduses (*Chrysaora hysoscella*)
macroplancton	2 - 20 cm	mollusques (*Conus amadis*)
mésoplancton	0,2 mm - 2 cm	copépodes (*Calanus finmarchicus*)
microplancton	20-200 µm	dinoflagellés (*Pyrodinium bahamense*)
nanoplancton	2 - 20 µm	coccolithophores (*Emiliania huxleyi*)
picoplanton	0,2 - 2 µm	bactéries, virus géants (ou girus)
femtoplancton	< 0,2 µm	virus

Figure 1.1 – Échelle des tailles des organismes planctoniques marins

La diversité taxonomique et la structuration des communautés au sein des écosystèmes marins planctoniques, sont influencées par les conditions physico-chimiques telles que la température, la salinité ou encore l'accès aux nutriments [Richardson

et Schoeman 2004]. Dans certaines conditions, par exemple lorsque des individus d'une communauté rencontrent des conditions qui leurs sont favorables, ils peuvent former des efflorescences planctoniques (*blooms*). Ces *blooms* sont fréquemment observées chez certaines lignées de protistes photosynthétiques comme les diatomées, les dinoflagellés ou encore les coccolithophores [Hallegraeff 1993] (Figure 1.2).

Figure 1.2 – Photographie satellite prise par la NASA (*National Aeronautics and Space Administration*) d'une efflorescence (ou *bloom*) de plancton, ici du phytoplancton : coccolithophores (Haptophyta, micro-algues unicellulaires eucaryotes), dans l'océan Atlantique au niveau de la côte ouest française le 20 avril 2013

De manière classique, deux grandes catégories de plancton peuvent être distinguées : le « plancton végétal » ou phytoplancton (du grec πηψτον : « plante ») et le « plancton animal » ou zooplancton (du grec ζοο : « animal »). Le phyto-

plancton correspond aux organismes planctoniques capables de produire leur propre matière organique par la réaction de photosynthèse, ils sont dits autotrophes* vis à vis du carbone. Grâce à ce processus, le phytoplancton contribue à environ 50% de la production de matière organique globale (production primaire) sur la planète alors que l'autre moitié est produite par les producteurs primaires terrestres [Field et al. 1998 ; Guidi et al. 2016]. Au sein du phytoplancton on compte de nombreuses lignées de micro-algues unicellulaires eucaryotes (*e.g.* Dinophyta, Cryptophyta ou Bacillariophyta) et procaryotes (*i.e.* les cyanobactéries). Le phytoplancton est à la base de la plupart des chaînes alimentaires dans les écosystèmes marins planctoniques. Il est consommé par d'autres organismes planctoniques appartenant au zooplancton comme les Radiolaria, les Cnidaria ou les Crustacea [Berglund et al. 2007]. Le zooplancton correspond aux organismes hétérotrophes* nécessitant une source externe de matière organique pour se nourrir. Enfin certaines lignées, dites mixotrophes*, sont capables de s'approvisionner en matière organique nécessaire à leur survie en combinant hétérotrophie et autotrophie. Dans la mixotrophie, deux cas de figures se présentent : soit les organismes originellement hétérotrophes* peuvent séquestrer les chloroplastes* de leur proies photosynthétiques préalablement ingérées ou tirer des bénéfices d'une association symbiotique avec ces dernières ; soit certains organismes originellement autotrophes* sont capables de prédation comme c'est le cas pour certaines espèces de dinoflagellés [Flynn et al. 2013 ; Stoecker 1999 ; Stoecker, P. J. Hansen et al. 2017].

La diversité du plancton s'exprime également en fonction de la répartition verticale le long de la colonne d'eau. Une partie des lignées planctoniques vit dans la zone photique qui correspond à la couche superficielle des océans où les rayons lumineux pénètrent, permettant donc l'accès à l'énergie lumineuse. Dans cette zone, les lignées planctoniques capables de phototrophie utilisent l'énergie lumineuse pour synthétiser la matière organique via la photosynthèse. Les lignées hétérotrophes* strictes ou mixotrophes*, se nourrissant de lignées photosynthétiques peuvent également vivre ou migrer dans la zone photique pour se nourrir d'organismes phytoplanctoniques [Levine et al. 1999].

De nombreuses lignées planctoniques synthétisent des squelettes en carbonate de

calcium (*i.e.* coccolithophores, foraminifères) ou en silice (*i.e.* diatomées, radiolaires polycystines) qui sédimentent vers le fond des océans lorsque ces organismes meurent et ainsi ils participent aux grands cycles biogéochimiques et particulièrement à celui du carbone [Falkowski 2012 ; Herndl et Reinthaler 2013]. Par exemple, dans le phénomène de la pompe biologique à carbone, le dioxyde de carbone dissous dans la zone photique est converti en matière organique par le biais de la photosynthèse, puis est transporté par sédimentation vers les profondeurs sous la forme de particules sous l'effet de la gravité ou par transport actif par le zooplancton, c'est ce que l'on appelle l'export de carbone. Ce carbone absorbé par la pompe biologique est pour une grande partie stocké dans les sédiments marins durant des milliers d'années. La pompe biologique est le mécanisme biologique le plus important sur Terre permettant de séquestrer le carbone à partir du CO2 sur des temps géologiques (plus de $10x10^{12}$ kg de carbone par an sont exportés de la zone photique par la pompe à carbone biologique, [Buesseler et Boyd 2009]) et ainsi a un impact direct sur le climat de la planète. Cet exemple met en perspective l'importance de l'étude des acteurs, des fonctions ainsi que des interactions qui composent et structurent les communautés planctoniques [Benoiston et al. 2017 ; Guidi et al. 2016 ; Worden et al. 2015].

1.2 Plancton marin et génomique

Nos connaissances des organismes planctoniques sont principalement limitées aux quelques lignées suffisamment robustes pour être prélevées sans dommage, observées au microscope, puis éventuellement cultivées en laboratoire. C'est le cas notamment pour les copépodes (Crustacea) et les micro-algues telles que certaines diatomées (Bacillariophyta) ou dinoflagellés (Dinophyta) [Tristan Biard, Stemmann et al. 2016 ; Bittner et al. 2013]. D'autre part, les séquences des gènes codant pour la petite sous unité de l'ARN ribosomique (18S rDNA et 16S rDNA pour les lignées eucaryotes et procaryotes) sont aujourd'hui les standards pour l'étude de la diversité des lignées microbiennes dans l'environnement [Bittner et al. 2013 ; Forster et al. 2015 ; Piganeau et al. 2011]. De récentes initiatives scientifiques à large échelle visant à étudier le plancton marin dans l'environnement telle que l'expédition *Tara*

Océans [Karsenti et al. 2011 ; Pesant et al. 2015] ou encore à partir d'espèces en culture telle que le « *Marine Microbial Eukaryote Transcriptome Sequencing Project* » (MMETSP) [Keeling, Burki et al. 2014] ont permis d'accumuler de nombreuses données génomiques provenant de milliers d'échantillons d'espèces planctoniques (*e.g.* 35 000 échantillons d'eau de mer filtrée pour *Tara* Océans) ainsi que de relever des métadonnées environnementales associées. En complément des études environnementales, les collections de culture de microorganismes marins (*e.g. Roscoff Culture Collection*, RCC [Vaulot et al. 2004]) constituent des ressources précieuses pour les études de génomique du plancton marin.

1.2.1 *Marine Microbial Eukaryotic Sequencing Project* (MMETSP)

Le projet MMETSP (https://www.imicrobe.us/project/view/104) représente un effort international qui a pour objectif d'augmenter le nombre de données de séquences génomiques sur les protistes marins afin de compléter les banques de données et de fournir davantage de références [Keeling, Burki et al. 2014] (Figure 1.3). Or, comme pour les bactéries et archées, la proportion estimée de lignées cultivables de protistes représente en moyenne moins de 1% du nombre total de lignées dans un échantillon d'une communauté naturelle [Bittner et al. 2013 ; Forster et al. 2015 ; Shi et al. 2009]. La quantité et la nature des données produites par ce projet restent néanmoins inédites et permettent notamment de nouvelles études phylogénomiques*, affinant la compréhension de l'évolution de la diversité de ces organismes [Burki, Kaplan et al. 2016 ; Burki, Kudryavtsev et al. 2010 ; Janouškovec et al. 2016]. Les études comparatives de méta-génomique* et de méta-transcriptomique* environnementales sont également facilitées grâce à ce jeux de données car elles bénéficient désormais de références génomiques dont les séquences sont produites et traitées de façon homogènes [CARRADEC et al in press]. Ce projet a impliqué plus de 70 laboratoires, et 200 collaborateurs ont participé aux cultures et au séquençage (Figure 1.4). Un total de 678 transcriptomes (ensemble des ARNs d'un organisme dans des conditions données et à un instant donné) incluant 210 genres, répartis en

305 espèces et 396 souches de protistes ont été générés à ce jour et sont disponibles sur le site iMicrobe (https://www.imicrobe.us/project/view/104).

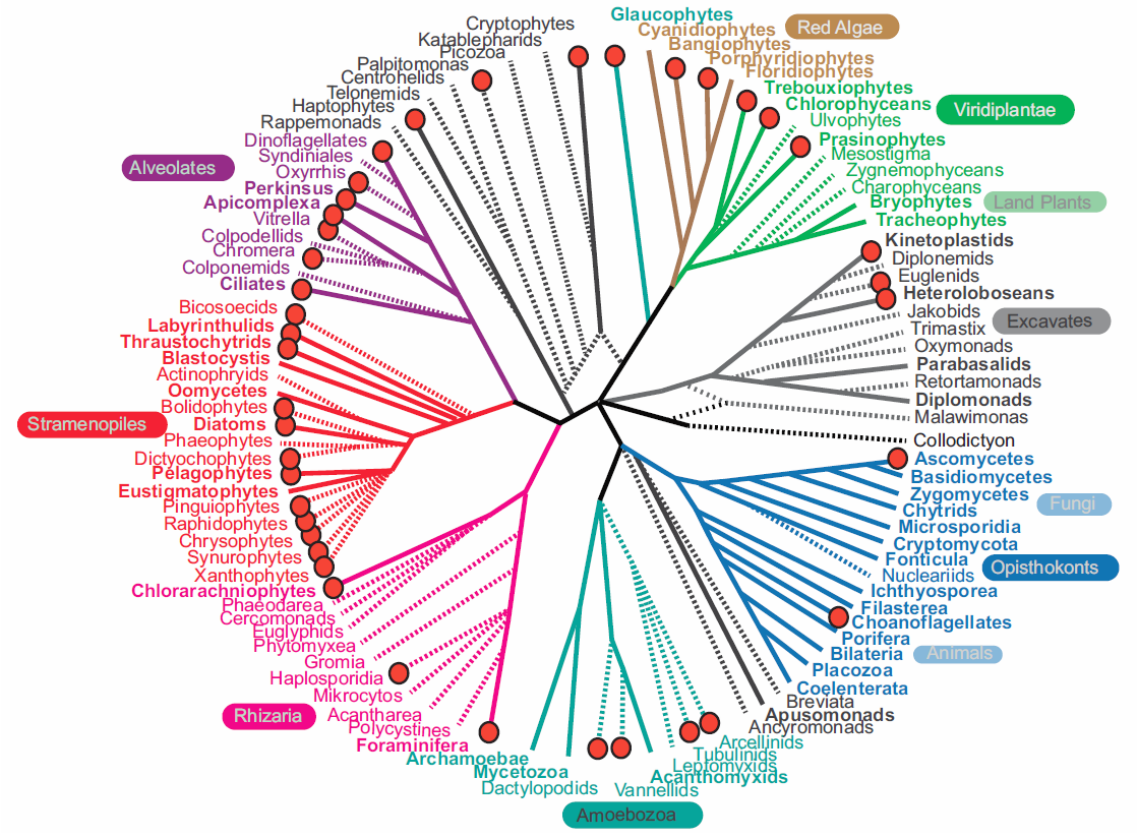


FIGURE 1.3 – [KEELING, BURKI et al. 2014] Représentation schématique des relations de parenté entre les lignées majeures d'eucaryotes. Les lignes pleines représentent les lignées dont le génome est disponible sur la base de données du *Joint Genome Institute* (JGI), alors que les lignes en pointillés représentent les lignées dont le génome n'est pas disponible. Les lignées dont le transcriptome a été assemblé au cours du projet MMETSP sont indiquées avec un point rouge

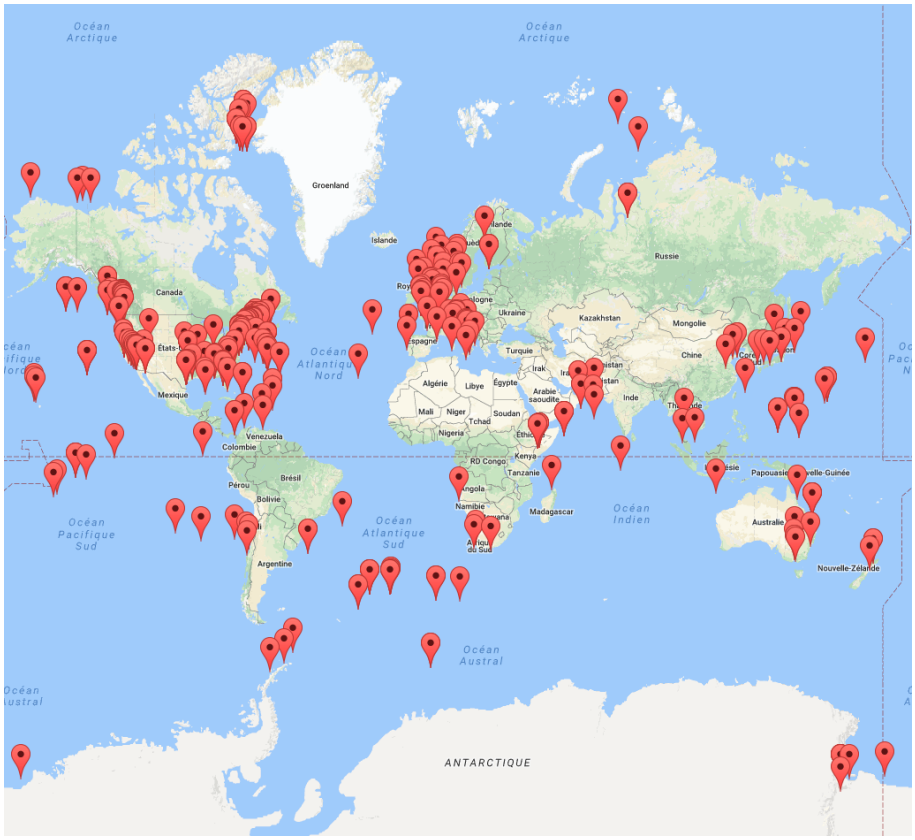


FIGURE 1.4 – Origines géographiques des échantillons de protistes du projet MMETSP (https://www.imicrobe.us/project/view/104)

1.2.2 La *Roscoff culture collection* (RCC)

La Roscoff Culture Collection (http://roscoff-culture-collection.org/) [VAULOT et al. 2004] est une plateforme de cultures de plancton marin située à la Station Biologique de Roscoff et comptant à ce jour 4 027 souches de micro-algues marines, bactéries et virus (octobre 2017). Cette banque d'organismes inclut 542 souches de Dinophyceae dont une majorité d'espèces du genre Alexandrium (421 genres) (ttp://roscoff-culture-collection.org/about-rcc/rcc-stats). Parmi ces dinoflagellés, 3 souches symbiotiques, cultivées ici dans leur phase libre et dont le transcriptome a été séquencé, ont été sélectionnées pour mes travaux de thèse : RCC3507 (*Gymnoxanthella radiolariae*, Gymnodiniales) [YUASA, HORIGUCHI et al. 2016] et RCC3468 (*Brandtodinium nutricula*, Peridiniales) [PROBERT et al. 2014], toutes deux décrites comme photosymbiontes de radiolaires (Rhizaria) du groupe des polycystines, ainsi que RCC1491 (*Pelagodinium beii*, Suessiales) [SIANO

et al. 2010] connue comme symbionte de foraminifères (Rhizaria). Une souche supplémentaire de dinoflagellés non-symbiotique a également été sélectionnée : RCC1516 (*Heterocapsa* sp., Peridiniales) (Figure 1.5). Les transcriptomes de chacune des ces 4 souches ont été séquencés par la plateforme de séquençage du Genoscope (Evry, France) en 2014.

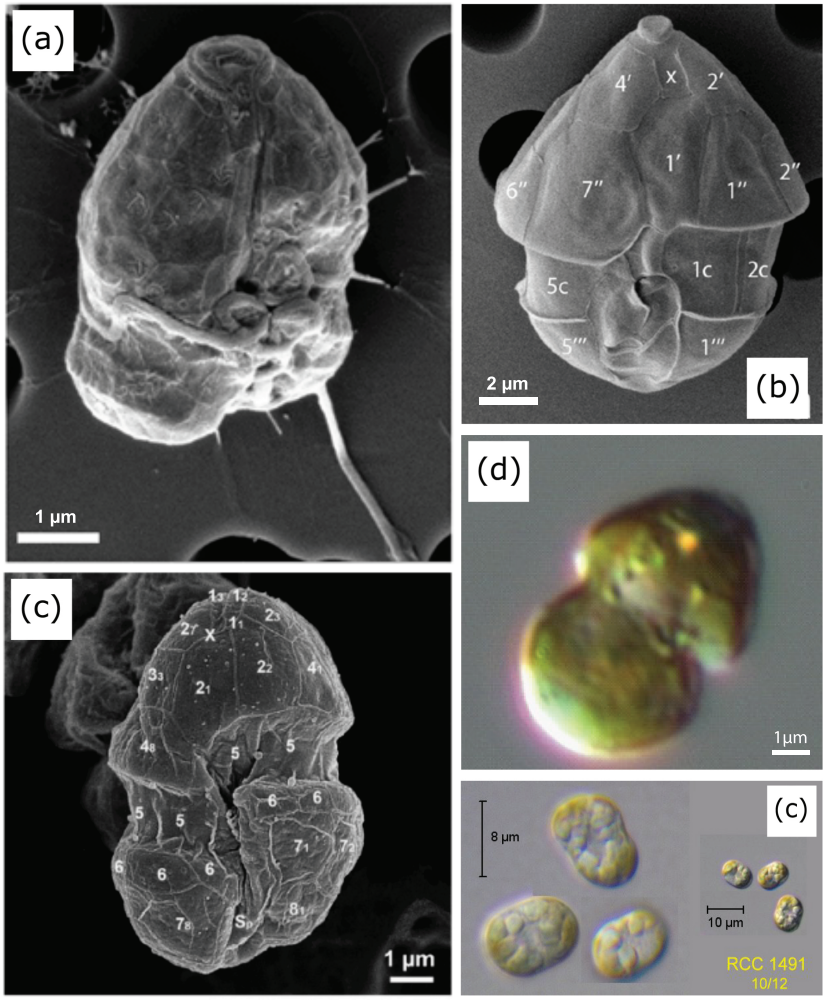


FIGURE 1.5 – Images par microscopie à balayage électronique (a, b et c) et microscopie optique (c et d) des souches de dinoflagellés provenant de la RCC. (a) RCC3507, (b) RCC3468, (c) RCC1491 et (d) RCC1516.

1.2.3 État des lieux de nos connaissances actuelles en génomique chez les protistes

La génomique permet l'étude de l'ensemble du matériel génétique d'un être vivant grâce à sa description par séquençage. Le séquençage est un procédé permettant

de déterminer précisément l'enchaînement des nucléotides (bases) d'une molécule d'ADN ou d'ARN. L'idée de « lire » l'ADN est apparue entre 1970 et 1973 lorsque Ray Wu (biologiste à l'université de Cambridge, Royaume-Uni) et ses collaborateurs suggèrent l'utilisation d'amorces composées d'oligonucléotides pour amplifier des courts fragments de l'ADN. En 1977, il y a tout juste 40 ans, Frederick Sanger (biochimiste au Medical Research Council, Royaume-Uni, devenu double prix Nobel par la suite) publie la première méthode de séquençage d'ADN dans l'article intitulé « *DNA sequencing with chain-terminating inhibitors* » [Sanger, Nicklen et Coulson 1977]. Grâce à ces découvertes, de nombreux domaines scientifiques ont trouvé de formidables opportunités d'applications telles que la médecine personnalisée ou encore les études de génomique environnementale [Pettersson, Lundeberg et Ahmadian 2009]. La génomique a connu un essor important durant ces 20 dernières années grâce à l'évolution des méthodes de séquençage et au développement des technologies dites à haut-débit. En 2005 apparaissent les premières méthodes de séquençage haut-débit (HTS pour *High-Throughput Sequencing* ou NGS pour *Next-Generation Sequencing*) [Reuter, Spacek et M. P. Snyder 2015]. Ces méthodes, dont la technologie Illumina est la plus utilisée aujourd'hui, permettent de séquencer des millions de fragments d'ADN (ou ARN) simultanément, réduisant significativement le temps nécessaire au traitement d'un génome complet. En 2015, une expérience de séquençage permet de lire jusqu'à 500 Gb (milliards de bases) par jour contre 1 Mb (million de bases) par jour en 1996 [Reuter, Spacek et M. P. Snyder 2015]. Si les méthodes sont plus rapides, leur coût diminue également. En effet, la lecture d'une seule base d'ADN coûtait 10$ en 1985 contre 0,05$ le million de bases aujourd'hui [Pettersson, Lundeberg et Ahmadian 2009]. Cette productivité grandissante des méthodes de séquençage couplées à des analyses bioinformatiques a permis à des projets de grande ampleur de voir le jour dès 2005. En génomique, on peut notamment citer le projet « 1 000 génomes humains » (initié en 2008, http://www.internationalgenome.org, [Consortium 2010 ; Siva 2008], le projet TCGA (*The Cancer Genome Atlas*, initié en 2005, https://cancergenome.nih.gov/, Tomczak, Czerwińska et Wiznerowicz 2015), ou encore le projet *1 000 Plant genomes* (annoncé dès 2008, initié en

2012, www.onekp.com, [Matasci et al. 2014]. En outre, les projets de séquençage de génome individuel, ou à plus petite échelle, voient le jour régulièrement permettant la création exponentielle de génomes de référence depuis 2005 (Figure 1.6). Les technologies de séquençage à haut-débit représentent donc une opportunité pour étendre nos connaissances à des lignées encore peu explorées du vivant et difficilement cultivables, dans la mesure où celles-ci peuvent être ponctuellement isolées du reste de la communauté dans laquelle elles vivent [Bork et al. 2015]. Dans le domaine de la méta-génomique*, les grands projets se multiplient aussi au fil des ans [Knight et al. 2012] : on peut citer notamment le HMP / *meta-Hit project* (*Metagenomic of Human intestinal tract*, initié en 2008, http://www.metahit.eu/, [Qin et al. 2010]), l'expédition *Tara* Océans dès 2009 (https://www.embl.de/tara-oceans/, [Bork et al. 2015 ; Karsenti et al. 2011] ou encore le EMP *project* (*Earth Microbial Project*, www.earthmicrobiome.org/, Gilbert, Jansson et Knight 2014, THOMPSON et al. *in press*]. En 2015, un premier article issu de l'analyse de 68 stations et 246 échantillons de l'expédition *Tara* Oceans a notamment fourni plus de 7.2 tera-bases de données et un premier catalogue exhaustif de gènes de référence d'organismes picoplanctoniques, essentiellement procaryote, marins [Sunagawa et al. 2015]. Pour un organisme, ou une communauté d'organismes dans un milieu donné, le séquençage du matériel génétique (ADN ou/et ARN) permet donc l'établissement d'un catalogue de gènes, celui-ci constituant alors une référence génomique, laquelle pourra être étudiée, complétée puis comparée à d'autres catalogues.

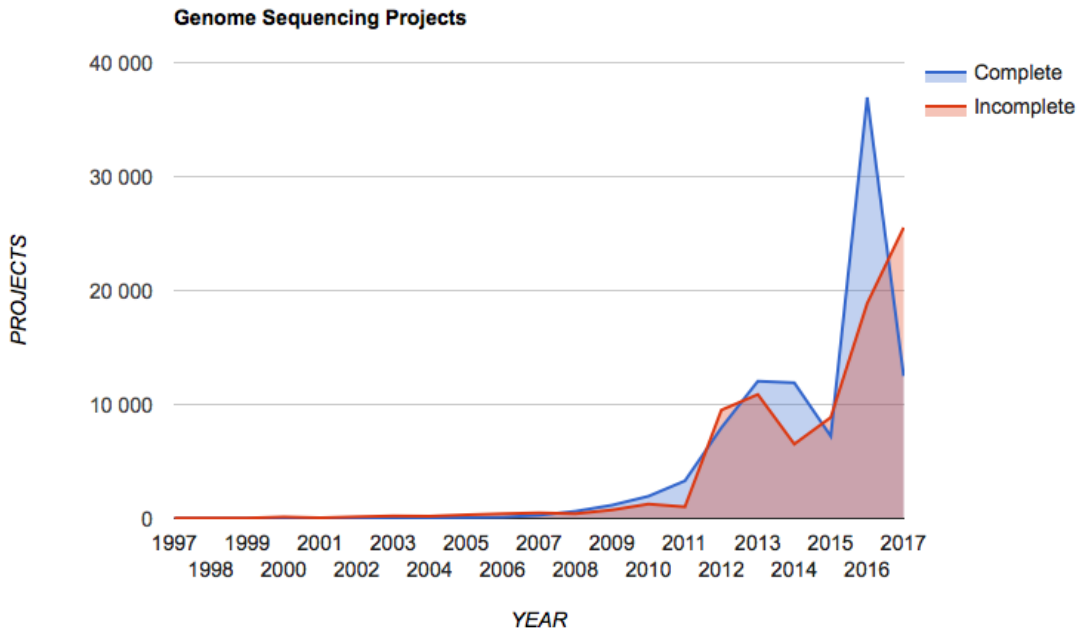


FIGURE 1.6 – Nombre de projets de séquençage de génomes complets (bleu) et incomplets (rouge) depuis 1997 (JGI https://gold.jgi.doe.gov/statistics)

Néanmoins, dans le cas du plancton, il existe très peu de références génomiques (Figure 1.7). En effet, environ 90% des 7 500 génomes référencés d'espèces eucaryotes disponibles dans les banques de données sont des génomes d'animaux (Opisthocontes, Metazoa), de plantes (Viridiplantae et Bryophytes) et de champignons (Opisthochontes, Fungi) [SIBBALD et ARCHIBALD 2017]. Contrairement à la plupart des lignées de bactéries, d'archées et de virus, les génomes de protistes marins ont tendance à être plus grands et plus complexes (Figure 1.8), ce qui constitue un facteur limitant pour les études génomiques, notamment au niveau des étapes d'assemblage et d'annotation [CARON, WORDEN et al. 2008]. Nos connaissances génomiques actuelles sur les protistes sont aussi biaisées par les études sur un nombre restreint d'organismes modèles* actuels (*e.g. Saccharomyces cerevisiae*, Phytophthora) et/ou les espèces parasites causant des maladies chez l'Homme (*e.g. Plasmodium*, Trypanosoma) [ANANTHARAMAN, IYER et ARAVIND 2007 ; MUKHERJEE et al. 2017]. Par conséquent la quantité de références séquencées (*e.g.* génomes, transcriptomes), assemblées et annotées ne constitue qu'une très faible portion de la diversité phylogénétique microbienne eucaryote [SIBBALD et ARCHIBALD 2017] (Figure 1.7). Le déficit de références génomiques pour les lignées de protistes s'explique également

par la difficulté à échantillonner et maintenir ces organismes en dehors de leur habitat naturel [Keeling et Campo 2017 ; Sibbald et Archibald 2017]. On compte à ce jour seulement 200 génomes référencés de protistes, dont 20 sont considérés comme « complets » (c'est-à-dire, dont tous les gènes ont été répertoriés) [Caron, Alexander et al. 2016]. L'ajout de nouveaux génomes (ou de données génomiques au sens large) de référence pour les protistes permettrait d'accélérer le développement de nouveaux modèles eucaryotes et contribuerait à parfaire nos connaissances sur la physiologie et l'écologie des protistes [Sibbald et Archibald 2017].

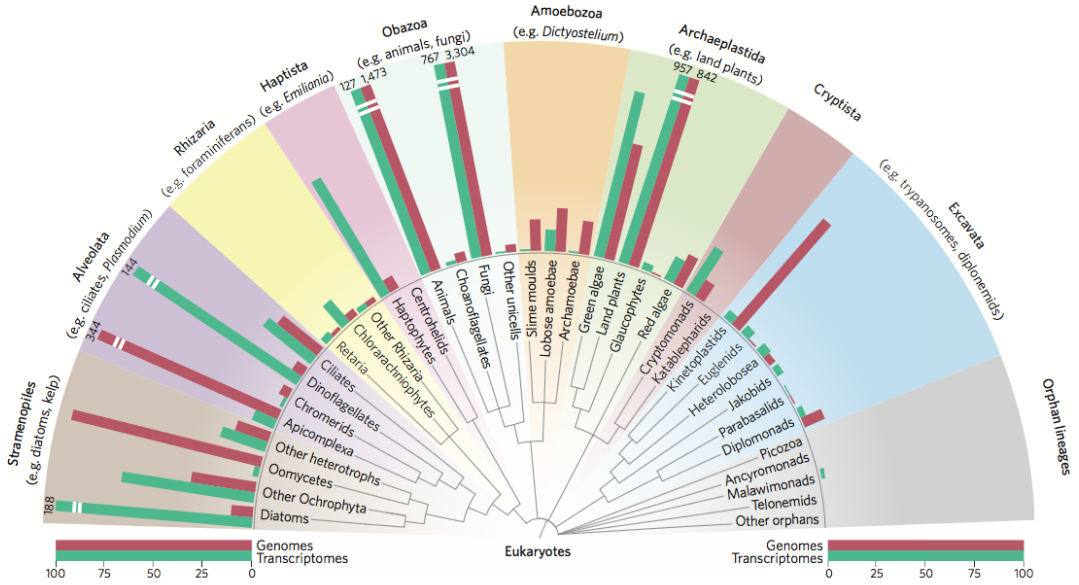


Figure 1.7 – [Sibbald et Archibald 2017] : Distribution des génomes (en rouge) et des transcriptomes (en vert) disponibles sur l'arbre phylogénétique des eucaryotes. Les barres d'histogramme représente le nombre brut de projets d'étude de génomes et transcriptomes référencés comme « complets » dans la banque de données *Genome Online Database*. Les données des transcriptomes incluent celles du projet de la MMETSP

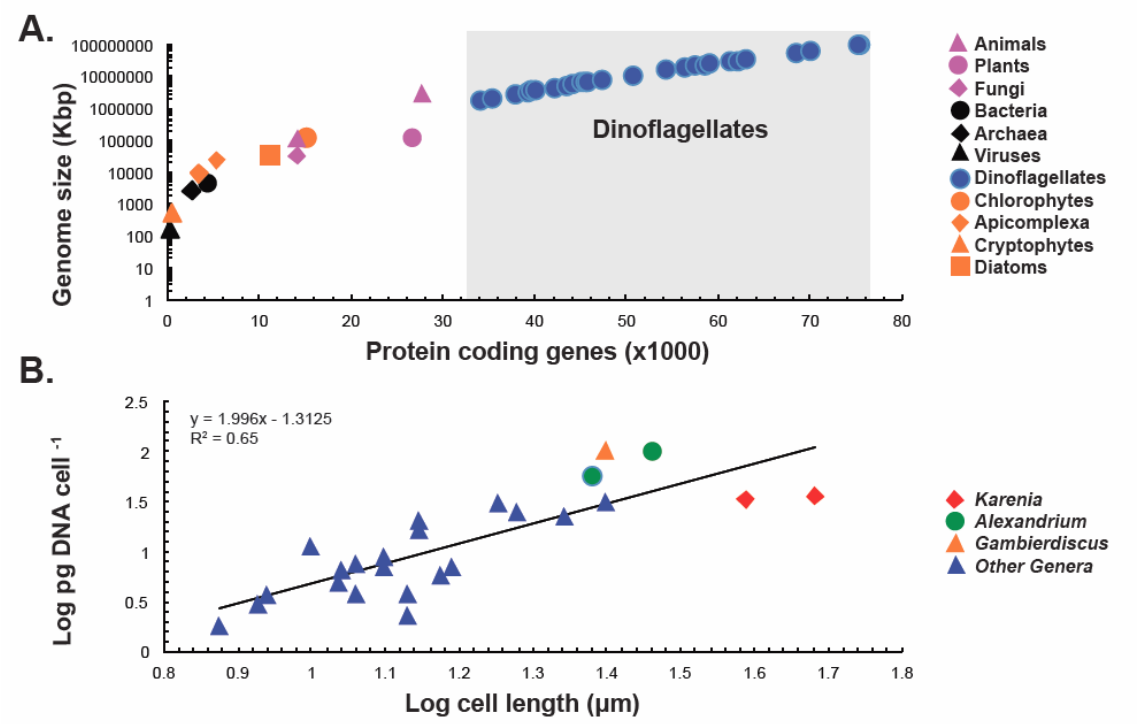


FIGURE 1.8 – [MURRAY et al. 2016] : (A) Taille de génome (haploïde) estimée en fonction du nombre de gènes codant pour des protéines pour diverses lignées d'organismes. Le contenu en ADN des dinoflagellés obtenu par cytométrie en flux [KOHLI et al. 2015 ; T. LAJEUNESSE 2002], a été converti en taille de génome (voir Thèse de G. Kholi 2013) et le nombre de gènes pour chaque génome de dinoflagellé a été calculé par régression [HOU et LIN 2009]. (B) Corrélation entre les dimensions de cellule (longueur ou axe dorso-ventral de la cellule) et la quantité d'ADN nucléaire de 26 espèces de dinoflagellés. Les données de taille de génome proviennent de [T. C. LAJEUNESSE et al. 2005 ; SHOGUCHI et al. 2013 ; VELDHUIS, CUCCI et SIERACKI 1997]. La courbe de tendance a été inclue (y = 1,996x - 1,3125 ; R^2 = 0.65)

1.2.4 La transcriptomique, une méthode adaptée pour l'étude d'organismes non-modèles*

Pour l'ensemble des lignées de protistes que j'ai étudiées dans ce travail de thèse, je disposais de données de séquençage des ARN totaux extraits à partir des cellules d'organismes cultivés ou échantillonnés et isolés directement. Ces données furent acquises par la technique appelée RNA-seq qui est une méthode, basée sur le séquençage à haut débit, qui permet d'obtenir l'ensemble des ARN messagers issues de l'expression des gènes (*i.e.* transcrits) d'un organisme, et ce à partir d'une faible quantité (quelques picogrammes) de matériel génétique de départ [MARTIN et Z. WANG 2011 ;

Nagarajan et Pop 2013]. Avec ce type d'approches il est ainsi possible d'identifier les gènes qui sont exprimés dans des conditions expérimentales ou environnementales données et dans quelles proportions [Oshlack, Robinson et Young 2010], d'identifier de nouveaux gènes ou transcrits [Z. Wang, Gerstein et M. Snyder 2009], d'étudier les transcrits alternatifs, c'est-à-dire des ARN résultants de différents évènements d'épissage [Alamancos, Agirre et Eyras 2013] ou encore d'étudier les remaniements génomiques c'est-à-dire des mutations/insertions/délétions [Piskol, Ramaswami et J. B. Li 2013]. Le RNA-seq est donc une approche largement utilisée pour étudier les organismes modèles* et non-modèles*. Dans le cas de ces derniers, le RNA-seq présente l'avantage de pouvoir produire un jeu de données génomique pour l'ensemble des ARN d'un organisme (*i.e.* un transcriptome) ou une communauté de plusieurs organismes présents dans un échantillon (*i.e.* un méta-transcriptome) dont on ne possède encore aucune information [Z. Wang, Gerstein et M. Snyder 2009]. L'étude d'un méta-transcriptome correspond à l'exploration du profil des gènes exprimés à un instant donné par une communauté d'organismes, dans un environnement/une condition donnée. Les premières études de méta-transcriptomique datent de 2005 [Aguiar-Pulido et al. 2016]. Les protocoles alors utilisés comprenaient des étapes d'amplification de l'ARN environnemental via des amorces aléatoires, suivies de clonage et de séquençage par la méthode Sanger. En 2005, les études de méta-transcriptomique* analysaient les séquences d'environ 400 clones [Poretsky et al. 2005] mais les performances des HTS permettent désormais d'étudier de manière plus exhaustive le profil fonctionnel des échantillons environnementaux ou de communautés microbiennes complexes [Aguiar-Pulido et al. 2016 ; Chistoserdova 2009]. Par exemple, les applications de la méta-transcriptomique* dans le domaine de la santé humaine ont permis des découvertes majeures en reliant les informations de composition microbienne intestinale et leur impact sur certaines pathologies (diabète de type I, obésité par exemple) [Jiang et al. 2016].

Dans cette thèse, je distingue un transcriptome d'holobionte (assemblage de deux organismes appartenant à des espèces distinctes (*i.e.* association symbiotique entre deux organismes)) (cf. chapitre 4), d'un méta-transcriptome. Le terme méta-transcriptome est utilisé pour désigner l'ensemble des informations génomiques liées

aux ARN d'une communauté d'organismes issus d'un échantillon environnemental [Handelsman et al. 1998]. À la différence, un holobionte, défini par son holo-génome (terme défini en 2013 par Eugène Rosenberg dans *The Hologenome Concept : Human, Animal and Plant Microbiota* [Rosenberg et Zilber-Rosenberg 2013] ou holo-transcriptome, désigne le matériel génétique d'entités bien définies : un hôte et ses symbiontes microbiens. D'ailleurs, le préfixe « méta » du grec μέτα signifie « au-delà » alors que le préfixe « holo » du grec ὅλος signifie « entier ». Si l'on considère qu'un méta-transcriptome équivaut à un holo-transcriptome alors on perd la notion de symbiose qui existe dans un holobionte [Bordenstein et Theis 2015]. De fait, mon travail porte sur l'étude les interactions qui existent entre les entités eucaryotes d'un holobionte (l'hôte et ses symbiontes microalgues) au travers de l'holo-transcriptome. Ne tenant pas compte des possibles communautés microbiennes procaryotes qui entourent l'holobionte, il s'agit donc d'une sous-partie de l'holo-transcriptome et je parlerai alors de transcriptome d'holobionte (Figure 1.9) [Theis et al. 2016].

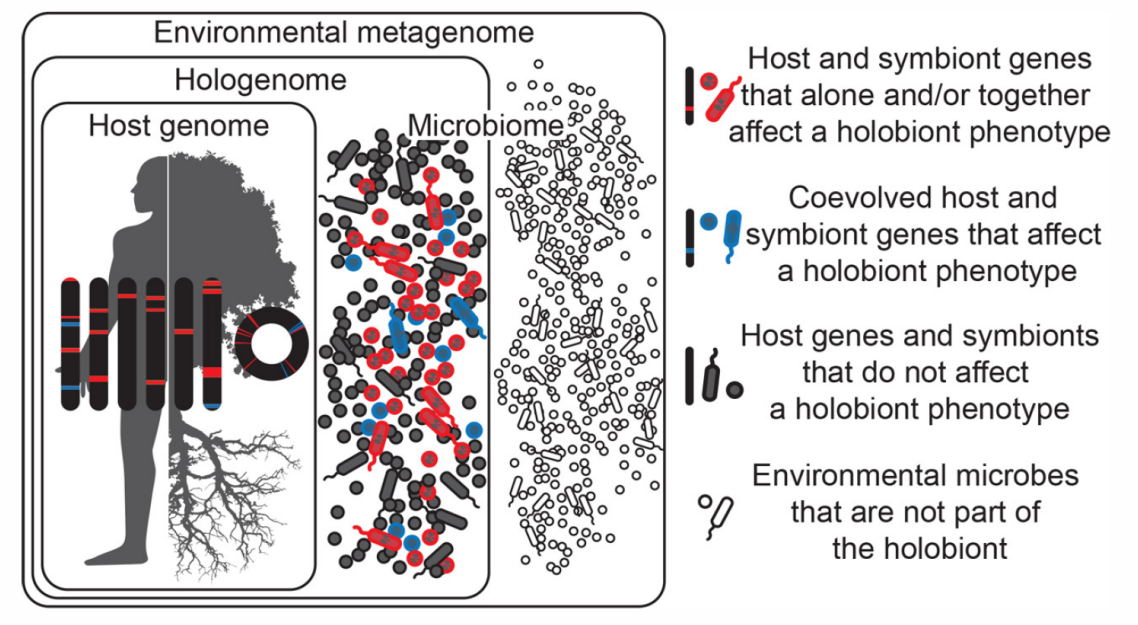


FIGURE 1.9 – [THEIS et al. 2016] : Distinction entre méta-génome environnemental et holo-génome. En bleu et rouge sont représentés les gènes de l'hôte et des symbiontes impactant le phénotype de l'holobionte. Les gènes n'ayant pas d'impact sur l'holobionte sont représentés en gris. D'un point de vue évolutif, seuls les gènes bleus ont co-évolués au sein d'un holobionte. Les gènes blancs représentent les gènes d'organismes microbiens externes à holobionte mais sont inclus dans la définition d'un méta-génome environnemental.

1.2.5 Défis liés aux analyses de transcriptomes et de méta-transcriptomes

Les études de transcriptomique et de méta-transcriptomique* présentent des difficultés inhérentes aux étapes de biologie moléculaire utilisées pour produire les données et aux algorithmes bioinformatiques utilisés pour les analyser. Lors des étapes préliminaires au séquençage des transcriptomes, des difficultés concernant le tri entre ARN ribosomaux (ARNr) et ARN messagers (ARNm) peuvent apparaître. Au cours de la phase d'extraction des ARN, la prédominance des ARNr comparativement aux ARNm entraîne un biais au cours de l'étape de séquençage. Ce biais tend à rendre difficile le séquençage des ARNm qui sont alors confondus dans la masse de séquences d'ARNr [S. HE et al. 2010 ; PASCAULT et al. 2015]. Des efforts ont été fait pour le développement d'outils informatiques (*e.g.* SortMeRNA) visant à séparer les ARNr des échantillons avant l'étape d'assemblage des séquences produites dans le but de

pouvoir les analyser de manière indépendante et de « nettoyer » les ARNm qui seront utilisés pour établir le profil fonctionnel de l'échantillon étudié [Abernathy et Overturf 2016 ; Kopylova, Noé et Touzet 2012]. Une autre difficulté est liée à l'instabilité des molécules d'ARN qui se dégradent relativement vite, de l'ordre de quelques heures en fonction des conditions physico-chimiques de stockage. Cela impacte la qualité du séquençage des molécules d'ARN lorsque celles-ci sont partiellement dégradées [Gallego Romero et al. 2014]. Les séquences complètement ou partiellement dégradées verront alors leur fiabilité diminuer et seront inexploitables. Les problèmes plus spécifiques aux analyses bioinformatiques se posent à la fois pour des échantillons de transcriptomique et de méta-transcriptomique. Le manque de références génomiques pour les organismes présents dans un jeu de données de transcriptomique ou de méta-transcriptomique* limite les possibilités d'assignation des séquences à des informations connues [Aguiar-Pulido et al. 2016]. De plus, les séquences d'organismes taxonomiquement proches vivant dans un même milieu et donc co-présents dans un échantillon méta-génomique* pourront être indiscernables si aucune référence génomique précise n'est disponible, conduisant ainsi à la création d'assemblages fragmentés ou potentiellement chimériques [Toseland et al. 2014]. Ces chimères* impacteront les analyses qui suivent l'étape d'assemblage telles que l'annotation fonctionnelle dont les résultats comporteront des séquences annotées qui seront considérées comme faux-positifs*. Les études de transcriptomes ou méta-transcriptomes chez les organismes dont on ne dispose pas du génome de référence sont actuellement limitées par la phase appelée « assemblage » qui est l'étape de reconstruction des séquences de transcrits à partir des lectures qui sont les fragments de séquences produites par le séquençage. Dans le cadre de l'étude du transcriptome d'un organisme pour lequel les données génomiques sont déjà disponibles dans les banques de séquences publiques (*e.g. The European Bioinformatic Institute* EBI / *National Center for Biotechnology Information* NCBI), ces séquences servent de support d'alignement aux nouvelles lectures. Ces assemblages dit *ab initio* ont l'avantage de mobiliser des moyens informatiques raisonnables de quelques Gigaoctets de mémoire vive, et de 2 à 4 processeurs en moyenne. Cependant, les alignements auront des résultats dont la qualité dépendra directement de celle des

séquences de référence [Martin et Z. Wang 2011]. Par opposition, on distingue les méthodes d'assemblage dites *de novo*. Actuellement les techniques d'assemblage *de novo* utilisent des algorithmes basés sur des graphes dans le but de reconstruire les séquences des transcrits à partir de nouvelles lectures uniquement [Miller, Koren et Sutton 2010]. Il existe plusieurs solutions permettant de réaliser un assemblage *de novo* dont l'efficacité dépend de la longueur initiale des séquences que l'on cherche à assembler [Martin et Z. Wang 2011]. Par exemple, l'approche employant les graphes de de Bruijn [Bruijn 1946] (Figure 1.10) est reconnue comme la plus efficace pour le traitement des séquences courtes (*i.e.* inférieures à 100 pb) et comportant un taux d'erreur inférieur à 1% avec 75% de lectures parfaites telles que celles générées par les technologies Illumina [Nagarajan et Pop 2013]. En revanche l'approche *overlap-layout-consensus* (OLC) est préconisée pour les séquences de taille supérieure à 200 pb incluant un taux plus élevé d'erreurs, qui dépasse les 1%, comme avec les technologies Sanger et Roche 454 [Nagarajan et Pop 2013 ; Robasky, Lewis et Church 2014]. De nombreux outils ou assembleurs existent déjà afin d'effectuer de l'assemblage transcriptomique *de novo*. Parmi les plus couramment utilisés on trouve Trinity [Grabherr et al. 2011], Velvet/Oases [Schulz et al. 2012 ; Zerbino et Birney 2008], Trans-ABySS [Robertson et al. 2010], SOAPdenovo-Trans [Luo et al. 2012], MIRA [Chebreux 2017] ou encore Bridger [Chang et al. 2015] (Figure 1.11). La comparaison de ces outils a montré que les quantités de RAM nécessaires à leur bon fonctionnement sur un jeu de données composé de 106 millions de reads pairés (de *D. Melanogaster*) variaient entre 8,2 Go et 137 Go [Zhao et al. 2011].

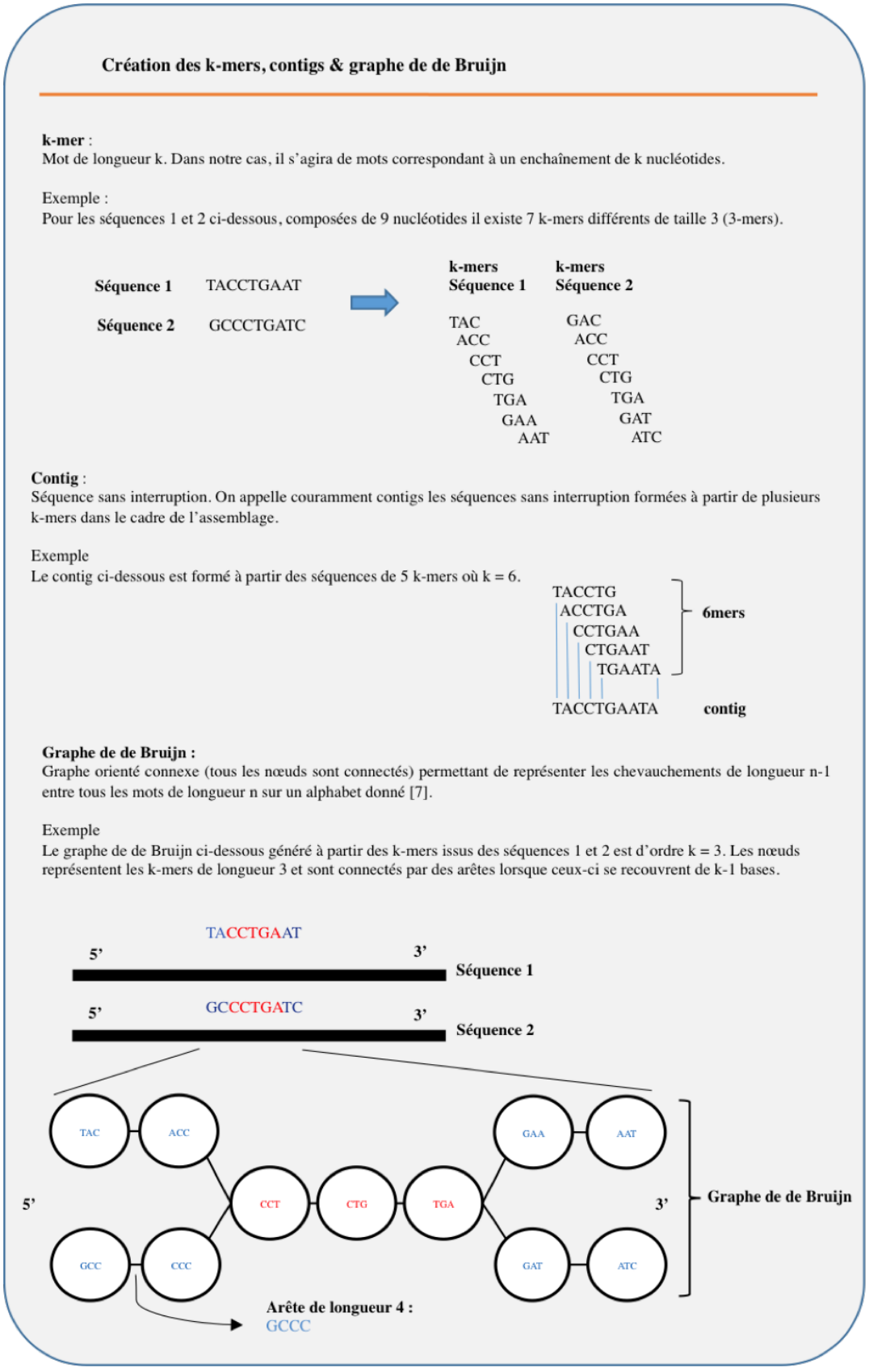


FIGURE 1.10 – Présentation des graphes de de Bruijn. (Extrait de mon rapport de stage de Master 2, 2014)

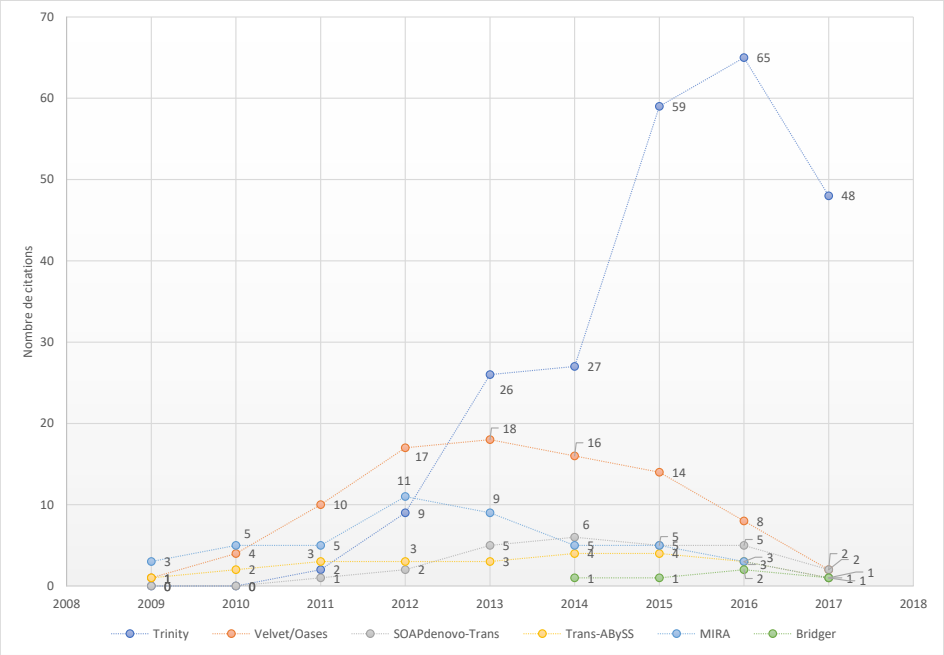


FIGURE 1.11 – Évolution du nombre de citations par année de plusieurs programmes d'assemblage *de novo* de transcriptome (2008-2017). Les citations sont issues de la base PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) avec la recherche de pattern : `((transcriptom*[Title/Abstract]) AND assembl*[Text Word]) AND (assembler_name)[Text Word]`

Les assembleurs cités ci-dessus implémentent tous des algorithmes de graphes de de Bruijn mais diffèrent par certaines étapes de leur fonctionnement. D'après des études comparatives [GALLO et al. 2014 ; ZHAO et al. 2011], on ne peut prédire à ce jour si un assembleur permettra de générer *a priori* le meilleur assemblage possible, ce qui rend d'autant plus difficile le choix de la solution optimale. Dans un premier temps, les assembleurs peuvent être comparés du point de vue de l'optimisation de leurs capacités de traitement. La quantité de RAM utilisée (*Random Access Memory*), le nombre de CPUs (*Central Processing Unit*) nécessaire, leur temps de calcul ainsi que leur capacité à être parallélisable pour traiter des informations de manière simultanée, sont les paramètres qui permettent de les différencier. Dans un deuxième temps, la robustesse d'un assemblage peut être évaluée selon une diversité de métriques* liées aux propriétés des séquences assemblées telles que leur nombre, leur

longueur moyenne ou encore le taux de d'alignement des lectures sur les séquences assemblées, aussi appelé le taux de *mapping** [B. Li et al. 2014]. Les résultats d'assemblages peuvent différer si lors du processus de reconstruction une métrique* est favorisée par rapport aux autres. Par exemple, le logiciel Trinity a tendance à favoriser la reconstruction de longs contigs* alors que Velvet/Oases maximise plutôt leur nombre [Haas et al. 2013].

1.2.6 Explorer l'ensemble des jeux de données / où comment explorer les séquences non-annotées

Une fois assemblées, les données de séquences produites doivent être analysées pour en tirer l'information biologique pertinente à l'étude menée. Les quantités toujours plus importantes de données à traiter dans les domaines de la génomique et de la transcriptomique font de leur analyse comparative et exhaustive un défi toujours plus difficile à relever [Stephens et al. 2015]. Il existe néanmoins des méthodes algorithmiques et bioinformatiques permettant de traiter de gros volumes de données de manière efficace. C'est le cas des méthodes d'analyse de réseaux au sens de la théorie des graphes. Certaines études s'appuient sur la construction et l'exploitation de réseaux de similarité de séquences (*sequence similarity networks*, SSN*) pour visualiser et explorer les relations entre familles de protéines [Alvarez-Ponce et al. 2013 ; Atkinson et al. 2009 ; Corel et al. 2016 ; Lopez, Halary et Bapteste 2015 ; Méheust et al. 2016]. La structure des réseaux de similarité de séquences modélise un ensemble de séquences homologues sous la forme d'un réseau dans lequel les séquences correspondent aux sommets (ou noeuds) et chaque arête représente une relation entre deux séquences comme par exemple un alignement global (Figure 1.12). Des regroupements de noeuds connexes sont définis, appelés composantes connexes (CC). Les composantes connexes définies au sein d'un réseau représentent chacune une sous-communauté du réseau. Chacune de ces sous-communautés peut être traitée de manière indépendante, ce qui accélère considérablement les calculs et offre la possibilité d'explorer des questions spécifiques à certaines populations de séquences. Les alignements permettent de détecter les homologies entre deux ou

plusieurs séquences. Dans le cas de séquences protéiques, une composante connexe peut résulter du regroupement de complexes fonctionnels. En effet les protéines sont composées de domaines qui sont des unités fonctionnelles et/ou structurales donnant sa fonction à une protéine. Sur la base d'une homologie entre domaines de protéines, il est possible d'inférer une homologie de fonction [Marchler-Bauer et al. 2011 ; Pearson 2013]. Ces méthodes sont particulièrement intéressantes car elles facilitent l'exploration de génomes ou transcriptomes ou protéomes* composés de plusieurs milliers à millions de séquences. Par exemple, en 2009 Atkinson et al. explorent 3 super-familles de protéines à partir de la comparaison de 773, 621 et 1330 séquences sous la forme de SSN*. En 2016, Méheust et al. étudient l'évolution réticulée chez des lignées eucaryotes à partir de SSN* comparant 2 192 940 séquences protéiques.

Pour explorer l'ensemble de mes transcriptomes, j'ai donc choisi de construire des réseaux de similarité des séquences (SSN*) des protéines prédites à partir des séquences de contigs* assemblés. Les noeuds et arêtes des réseaux présentent l'avantage de pouvoir ajouter des informations propres à chacune des séquences comme la taxonomie de l'organisme à laquelle appartient la séquence, les traits fonctionnels de l'organisme (*e.g.* photosynthétique, symbiotique, mixotrophe*) ou encore l'annotation fonctionnelle de la séquence elle-même. De même les arêtes (*i.e.* les alignements) portent des informations tel que le score d'alignement, la longueur, etc. La structure des réseaux (avec laquelle on peut jouer par filtration des informations portées par les arêtes) permet alors d'explorer les données en plusieurs dimensions et de rechercher et sélectionner des noeuds (*i.e.* séquences) particuliers ainsi que les composantes connexes auxquelles ils appartiennent (*i.e.* l'ensemble des séquences qui leur sont homologues). Par exemple, il est possible de rechercher les séquences ayant une annotation fonctionnelle A et appartenant au groupe taxonomique B et dont l'individu d'origine présente un trait fonctionnel particulier comme par exemple la production d'une toxine (Figure 1.13).

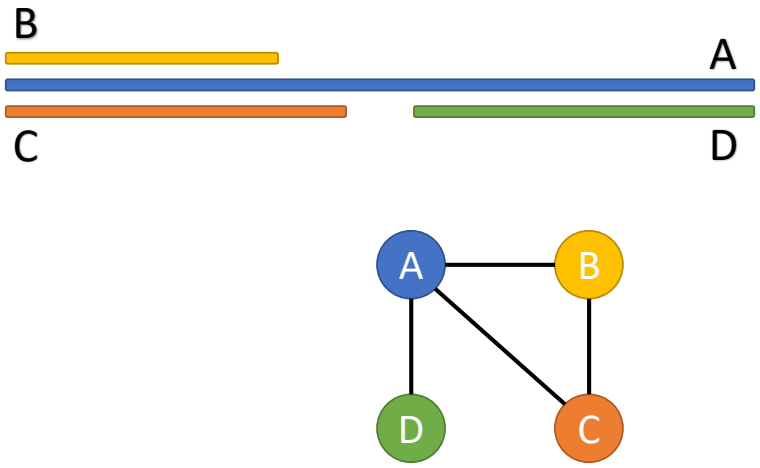


FIGURE 1.12 – Exemple d'un réseau de similarité de séquences. En haut, la représentation des alignements globaux entre quatre séquences A, B, C et D. En bas, Le réseau de similarité de séquence correspondant aux alignements. Les sommets (ou noeuds) représentent les séquences A, B, C et D ; une arête représente un alignement entre deux séquences.

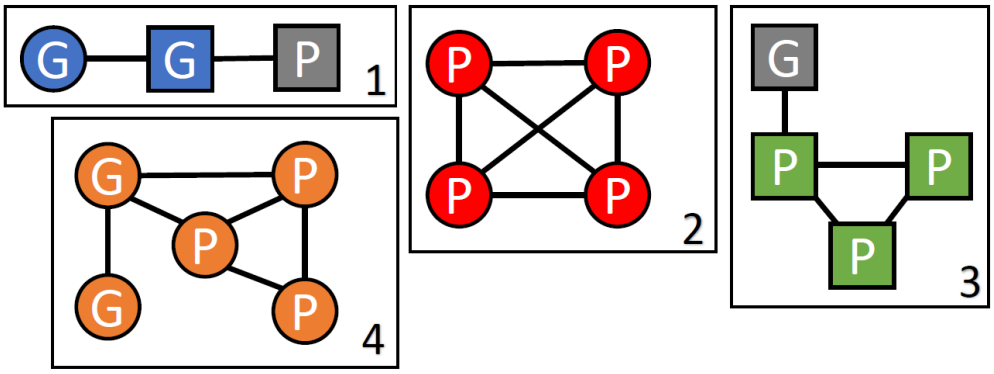


FIGURE 1.13 – Représentation schématique de quatre composantes connexes 1, 2, 3 et 4. Les lettres (P et G) correspondent à deux taxons fictifs. La couleur des noeuds correspond à des fonctions biologiques distinctes (ici quatre fonctions fictives : bleue, rouge, verte et orange) ou à une absence d'annotation fonctionnelle (gris). La forme des noeuds (carré ou rond) correspond au trait fonctionnel de l'individu dont sont originaires chaque séquence : par exemple, un rond indique que la séquence appartient au protéome* d'une espèce toxique et un carré pour une séquence d'un protéome* d'une espèce non-toxique.

Par ailleurs, au sein d'une composante connexe, il est envisageable d'étendre les annotations fonctionnelles des séquences connues aux séquences non-annotées de la même composante. Ainsi des hypothèses quant aux fonctions peuvent êtres faites pour des séquences initialement inconnues, et dans mon cas appartenant à

des espèces qui demeurent inexplorées au niveau de leur génome. Outre l'aspect d'exploration des profils fonctionnels au travers des composantes connexes du réseau, cette approche SSN* permet également de proposer de potentiels nouveaux gènes. Par exemple, la détection d'une composante connexe sans aucune annotation peut témoigner de l'expression de protéines homologues par plusieurs organismes ou échantillons indépendants mais regroupés dans le réseau. Dans le contexte de mon étude, je suggère que la détection d'homologies entre plusieurs séquences obtenues à partir d'assemblages de transcriptomes indépendants et provenant d'échantillons indépendants atteste de la réalité biologique et de l'importance des séquences concernées pour les lignées dont elles sont originaires (*i.e.* ces séquences ne représentent pas des chimères* issues des artefacts de reconstruction de séquences). Dans notre approche SSN*, la possibilité de trouver des groupes de séquences homologues sans annotation au sein de composantes connexes peut correspondre à l'identification de fonctions encore inconnues pour une lignées de protistes (ici les dinoflagellés). Ces groupes de séquences homologues, et plus généralement l'ensemble de ces informations inconnues au sein des transcriptomes et génomes, pourrait potentiellement permettre d'affiner les relations de parenté [Rinke et al. 2013], ainsi que permettre une caractérisation fonctionnelle plus exhaustive des organismes concernés.

1.3 Les symbioses dans le plancton marin

1.3.1 Définition et rôle de la symbiose dans le vivant

Les associations symbiotiques sont omniprésentes dans le monde vivant. Historiquement principalement associées à la description de parasites, la définition de la symbiose évolua lorsque, pour la première fois, Alfred William Bennett en 1877 (botaniste britannique) puis Anton de Bary en 1979 (botaniste, microbiologiste et mycologue allemand, étudiant notamment les lichens) proposèrent une nouvelle façon d'employer le mot « symbiose » pour définir un nouveau type d'interaction entre deux organismes. Leur vision impliquant la notion de « bénéfices mutuels » vint compléter cette définition initiale de la symbiose et la nouvelle définition fut progressivement acceptées et étendues aux lignées animale (Metazoa) à partir de

1881 [Perru 2006]. Finalement, les scientifiques de l'époque s'accordèrent à définir ce phénomène comme un état durable d'associations physiques d'au moins deux organismes appartenant à des espèces différentes parmi lesquelles au moins l'un en tirent des bénéfices. Cette association entre organismes fait intervenir de nouvelles appellations des partenaires symbiotiques. Ainsi, on définit le plus souvent l'organisme de plus grande taille comme « hôte » et les autres comme « symbiontes ». Un hôte peut avoir un ou plusieurs symbionte(s), l'association hôte-symbiontes est alors appelé holobionte*. Il existe donc plusieurs types d'interactions symbiotiques (Figure 1.14). Les associations mutualistes correspondent aux associations dans lesquelles les deux organismes tirent mutuellement bénéfices de leurs interactions. De nombreux exemples d'associations mutualistes existent dont l'un des plus connus est celui des mycorhizes résultant de l'association symbiotique entre des champignons et les racines des plantes [Zwaan, Jorissen et Stigter 1990]. Lorsqu'un seul des deux partenaires de la symbiose, le plus souvent le symbionte, tire des bénéfices de son interaction avec son partenaire, il s'agit de commensalisme. Dans ce type d'interaction, le symbionte peut se nourrir des réserves de son hôte et cohabiter avec ce dernier, sans pour autant lui nuire. Par ailleurs, leur deux organismes formant une interaction commensale demeurent capables de vivre séparément, c'est ce qu'on appelle une symbiose facultative [Bogitsh, Carter et Oeltmann 2013]. C'est notamment le cas pour le commensale *Candida albicans* (Fungi, Saccharomycetales) qui est vit à l'état naturelle au sein des muqueuses de l'Homme mais n'entraine ni symptôme ni maladie habituellement [Pande, Chen et S. M. Noble 2013]. Enfin, le parasitisme désigne les interactions symbiotiques pour lesquelles l'hôte se trouve affecté par la présence du ou des symbiontes. Ces interactions sont le plus souvent obligatoires, le parasite étant largement dépendant de son hôte pour survivre, et parfois tellement fortes, que l'on parle alors du système hôte-parasite comme d'un super-organisme décrit par Combes dans son ouvrage « *Parasitism : The Ecology and Evolution of Intimate Interactions* » de 2001.

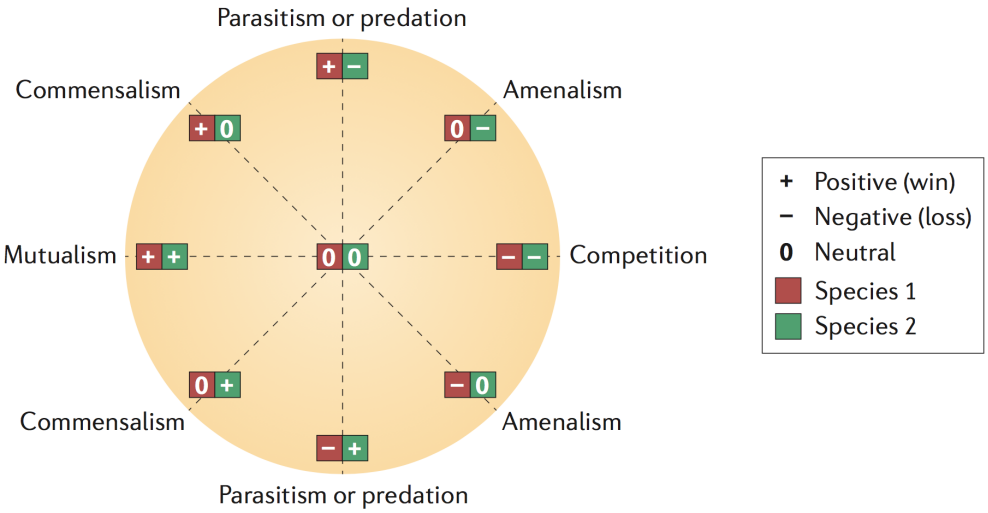


FIGURE 1.14 – [FAUST et RAES 2012] : Formalisation des interactions écologiques entre deux organismes d'espèces différentes. Pour chaque type d'interaction, le résultat de l'association peut être pour chacun des partenaires positif (+), négatif (-) ou neutre (0).

Les relations symbiotiques, qu'il s'agisse d'association mutualiste, de commensalisme ou encore de parasitisme, peuvent encore se différencier en sous catégories. Il est possible de distinguer les ecto-symbioses dans lesquelles le symbionte se trouve à l'extérieur de son hôte, des endo-symbioses où le symbionte est localisé à l'intérieur de son hôte . Dans le cadre d'une endo-symbiose entre deux ou plusieurs protistes, où une ou plusieurs cellules de symbiontes sont présentes dans la cellule de l'hôte, on parle alors d'endo-cyto-symbiose (DECELLE, COLIN et FOSTER 2015 ; SELOSSE, CHARPIN et NOT 2017). Le terme symbiose du grec συμβίωσις (« sumbiōsis »), signifiant littéralement « vivre ensemble », est un phénomène important pour évolution du vivant et reconnu comme un moteur de la biodiversité [ARCHIBALD 2015 ; ZIMORSKI et al. 2014]. Les interactions symbiotiques ont des implications sur la physiologie des partenaires de la relation. Ainsi, certains mammifères herbivores mais aussi certains arthropodes, comme les termites, sont capables, et ce uniquement grâce à leurs symbiontes bactériens et/ou fongiques, de digérer les molécules de cellulose et de lignine constituant des végétaux [RUSSELL et WILSON 1996 ; YUKI et al. 2015]. Les symbioses affectent également la structure et le fonctionnement des écosystèmes. Dans les océans, les sources hydrothermales localisées à proximité des dorsales océaniques sont considérés comme des environnements abyssaux hostiles

avec des variations de température de plusieurs centaines de degrés, des pressions hydrostatiques élevées (de 100 à plusieurs centaines de bars) ainsi que l'absence de lumière empêchant les processus de photosynthèse. Néanmoins, de nombreuses espèces y prolifèrent grâce à des association symbiotiques. En 2011, une étude montre que des bactéries chimiosynthétiques symbiotiques sont capables de produire de la matière organique par oxydation de molécules inorganiques telles que l'hydrogène, le sulfure d'hydrogène ou encore le méthane. Cette matière organique est alors utilisée par leur hôte, une moule du genre Bathymodiolus, ce qui leur permet de proliférer dans cet environnement a priori hostile [Petersen et al. 2011]. Le développement et le maintien des récifs coralliens et leur succès écologique dans des environnements pourtant oligotrophes sont également le fruit d'une association symbiotique entre les coraux (groupe d'espèces appartenant à l'embranchement des Cnidaria) et des micro-algues unicellulaires photosynthétiques du genre *Symbiodinium* qui sont ici les symbiontes localisés à l'intérieur des cellules de leur hôte [Roth 2014]. Grâce à la photosynthèse, les symbiontes fournissent à leur hôte les nutriments leur permettant de survivre malgré un environnement défavorable à leur prolifération. En échange, les micro-algues bénéficient de divers nutriments provenant de leur hôte [Davy, Allemand et Weis 2012]. Il s'agit ainsi d'un exemple d'une association mutualiste pouvant être qualifiée d'endo-photo-cyto-symbiose car elle fait intervenir un symbionte photosynthétique. La symbiose est ainsi à la base de la formation et du maintien d'écosystèmes benthiques majeurs mais c'est aussi un phénomène répandu dans la colonne d'eau au sein du plancton.

1.3.2 La symbiose dans le plancton marin

Les symbioses dans le plancton marin constituent la clef du succès évolutif et écologique de nombreuses lignées. Des symbioses ont été décrites entre des hôtes métazoaires et des symbiontes protistes ou procaryotes (*e.g.* cnidaire-dinoflagellés, porifères-bactéries chimiosynthétiques). Des associations symbiotiques faisant intervenir des hôtes unicellulaires eucaryotes et des endosymbiontes procaryotes comme les cyanobactéries ou même d'autres microalgues eucaryotes ont été décrites par le passé et l'on continue de découvrir régulièrement de nouvelles symbioses dans la zone

photique de tous les océans du monde, où elles jouent des rôles clés dans l'écologie et la biogéochimie des écosystèmes planctoniques [Decelle, Colin et Foster 2015 ; Guidi et al. 2016 ; Stoecker, M. D. Johnson et al. 2009 ; Taylor 1982]. En ce qui concerne les symbioses qui impliquent à la fois des hôtes et symbiontes protistes, les dinoflagellés jouent plusieurs rôles. Ils peuvent êtres hôtes de diatomées, de cryptophytes ou encore de pelagophyceae mais sont également fréquemment rencontrés dans le rôle de symbionte [Decelle, Colin et Foster 2015], en particulier avec des Rhizaria tels que les foraminifères et radiolaires qui sont largement répandus et abondants dans les océans [Tristan Biard, Stemmann et al. 2016]. Les interactions symbiotiques entre radiolaires et dinoflagellés ont été mises en évidence pour la première fois en 1881 par Karl Brandt qui identifia les cellules jaunâtres au sein de cellules de polycystines, comme des microalgues symbiotiques [Brandt 1881 ; Probert et al. 2014]. Consécutivement, des études ont cherché à identifier précisément les acteurs impliqués dans ces symbioses. Les dinoflagellés de l'espèce *Scrippsiella nutricula* furent dans un premier temps décrites comme photosymbiontes dans six espèces de polycystines [Gasol, Giorgio et Duarte 1997]. Puis sur la base de nouvelles phylogénies moléculaires faisant intervenir de nouveaux échantillons, cette espèce de dinoflagellés a été assignée au genre *Brandtodinium* en référence à Karl Brandt qui fût le premier à avoir décrit ces symbiontes [Probert et al. 2014]. Par ailleurs, des photosymbioses entre *Acanthochiasma* sp. (Radiolaria, Acantharia) et des dinoflagellés symbiotiques de plusieurs genres (*Pelagodinium*, *Heterocapsa*, *Azadinium* and *Scrippsiella*) ont été observées et décrites [Decelle, Probert et al. 2012]. Ces symbioses, tout comme dans d'autres modèles d'associations symbiotiques entre hôte hétérotrophe* et microalgues photosynthétiques, profitent aux radiolaires qui peuvent utiliser les produits issus de la photosynthèses de leur symbiontes.

1.3.3 Les radiolaires, un groupe largement répandu et abondant dans les océans

Les radiolaires (Rhizaria, Radiolaria) est un groupe de protistes planctoniques hétérotrophes* apparus au cours du Précambrien, soit il y a plus de 542 millions

d'années [D. M. Anderson, Chisholm et Watras 1983]. Ils possèdent un squelette minérale composé de silice (*i.e.* SiO2), ce qui fait des radiolaires un groupe qui présente un bilan fossile riche et diversifié. Ce groupe taxonomique tire son nom des formes, le plus souvent radiales, prisent par leur squelette minéral (Figure 1.15). En 1887, Ernst Heinrich Philipp August Hæckel, biologiste et philosophe allemand, publie un ouvrage de plus de 1 800 pages et 140 planches dédié à la description des Radiolaria. Son travail se base sur les échantillons de plancton collectés lors d'une expédition océanographique réalisée entre 1873 et 1876 à bord du navire *Challenger*. Il créé une classification de plus de 700 espèces, en s'appuyant notamment sur la forme du squelette des radiolaires (Figure 1.15). Sa classification restera une référence jusque dans les années 1970, lorsque les phylogénies moléculaires basées sur les séquences d'ADN apparaissent [Krabberød, Bråte et al. 2011]. Les radiolaires appartiennent au super groupe des Rhizaria, un des six super groupes d'eucaryotes [O. R. Anderson 2012]. L'embranchement des Radiolaria comprend cinq ordres : les Collodaria, Nassellaria et Spumellaria (qui forment la classe des Polycystinea) et les Taxopodia qui présentent un squelette en silice, et l'ordre des Acantharia qui possèdent un squelette en sulfate de strontium [Tristan Biard, Bigeard et al. 2017 ; Burki et Keeling 2014 ; Burki, Kudryavtsev et al. 2010 ; Decelle, Suzuki et al. 2012 ; Ogane et al. 2010 ; Sierra et al. 2013 ; Suzuki et Aita 2011] (Figure 1.16). Les Phaeodaria, traditionnellement considérés comme appartenant aux Radiolaria, sont maintenant classés dans le phylum des Cercozoa sur la base d'analyses de phylogénie moléculaire des d'ADN ribosomaux [Polet et al. 2004].

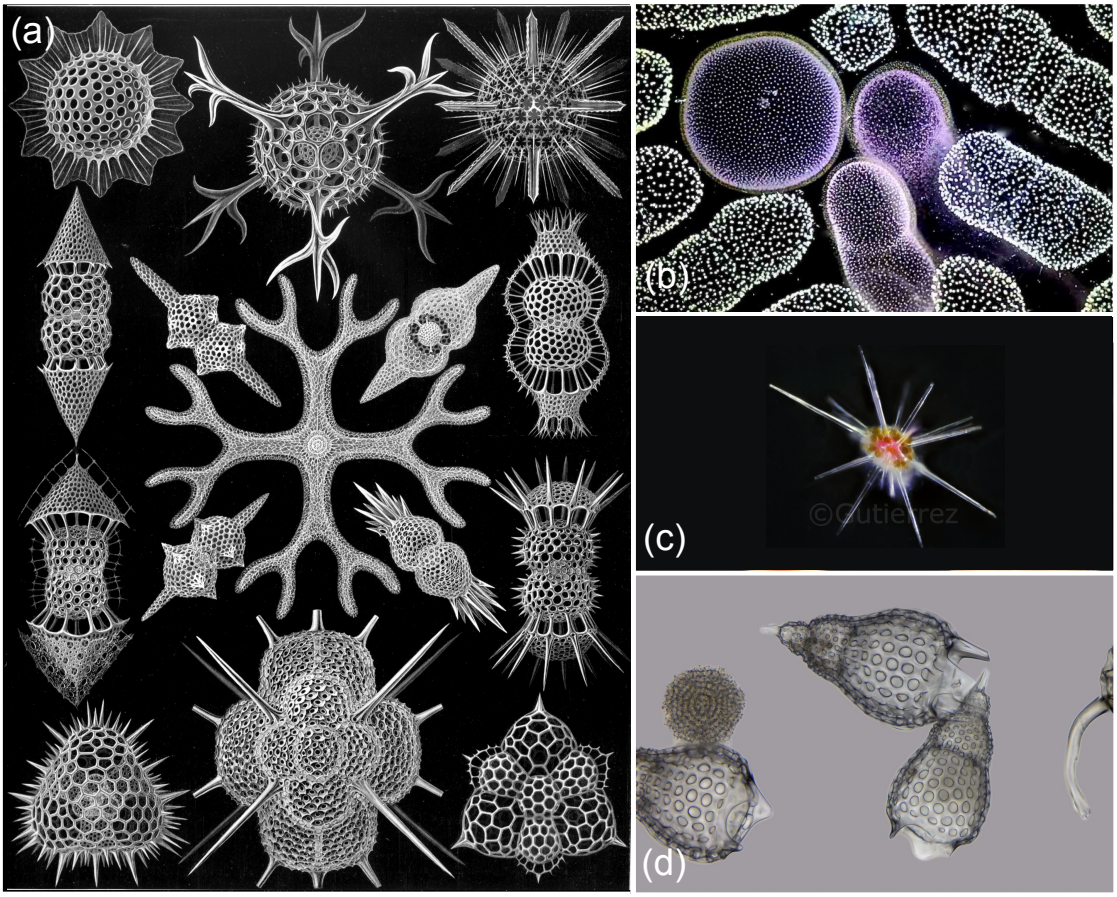


FIGURE 1.15 – Illustrations de Radiolaria. Les cellules de radiolaires ou colonies de radiolaires peuvent avoir des tailles comprises allant du nanoplancton (2 - 20 μm) au macroplancton (2 - 20 cm) (Figure 1.1) [SUZUKI et NOT 2015]. (a) Planche numéro 91 représentant des structures de Spumellaria (Polycystinea) [HAECKEL 1904]. (b) Colonie de radiolaires coloniaux : les Collodaria (Polycystinea) (https://johandecelle.wordpress.com). (c) Cellule d'Acantharia dont les symbiontes sont visibles (cellules plus petites et jaunâtres) (de https://johandecelle.wordpress.com). (d) Cellules de Nasselaria (Polycystinea) http://www.mikro-foto.de)

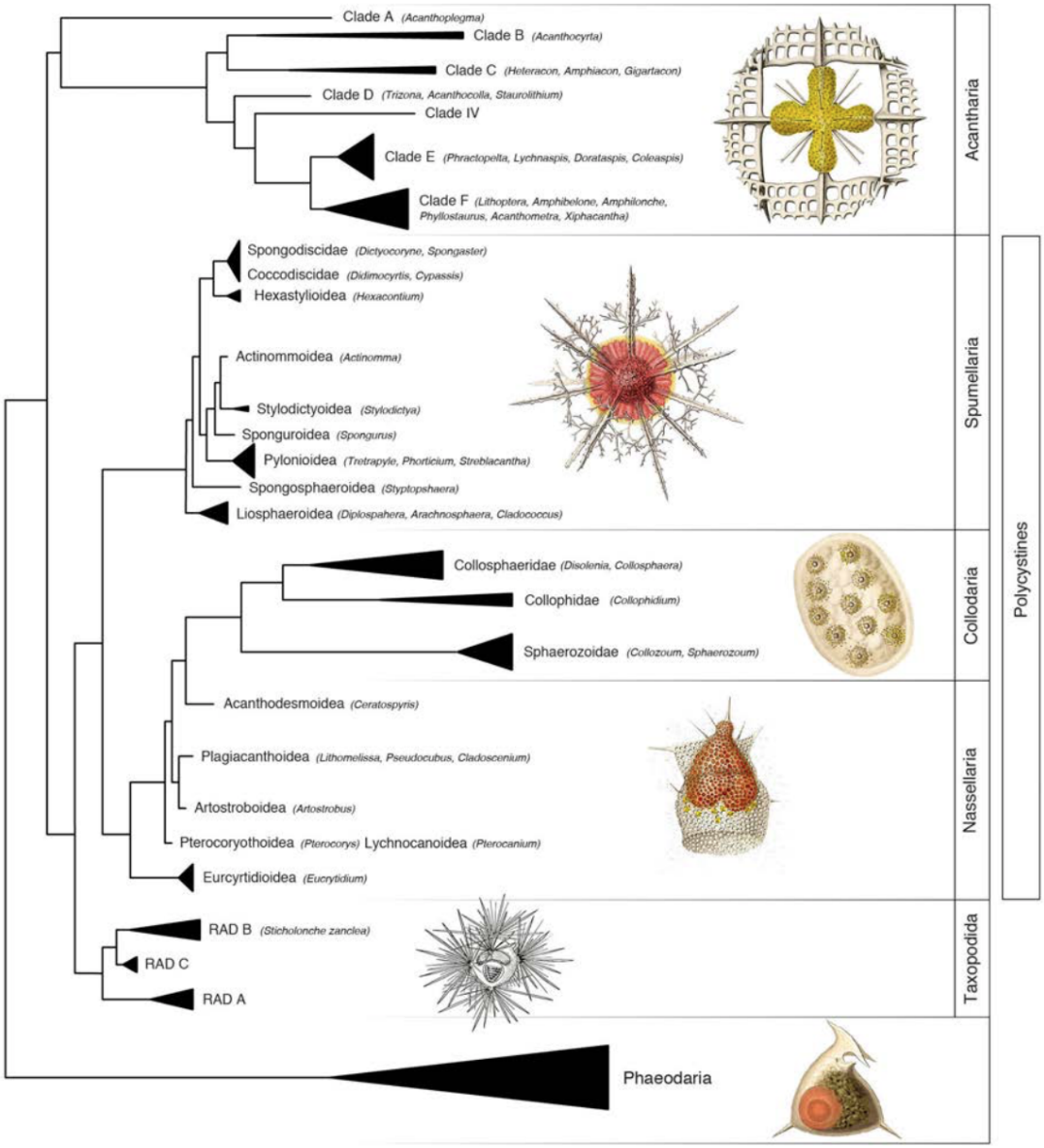


FIGURE 1.16 – Arbre phylogénétique schématique des radiolaires [Tristand BIARD 2016]. Les relations sont basées sur les connaissances morpho-moléculaires actuelles (critères morphologiques sur l'observation du squelette et phylogénie moléculaire sur la base des sous-unités d'ADNr). Le groupe externe (outgroup) des Phaeodaria (Cercozoa) permet d'enraciner l'arbre.

La taille des cellules de radiolaires est variable et peut s'étendre de quelques dizaines de micromètres jusqu'à plusieurs millimètres. Certaines espèces vivent isolées et d'autres forment des colonies pouvant atteindre plusieurs dizaines de centimètres (Figure 1.15). Les cellules de radiolaires se caractérisent par la présence d'une capsule centrale, au sein de laquelle est localisé le noyau, entouré de l'endoplasme qui

comporte la plupart des organites cellulaires. Une membrane, dénommée la membrane capsulaire, sépare l'endoplasme de l'ectoplasme dans lequel se retrouve les endosymbiontes s'ils sont présents dans l'espèce considérée [Suzuki et Aita 2011]. Toutes les lignées de radiolaires décrites à ce jour sont hétérotrophes* et nécessitent une source extérieure de matière organique pour survivre. Des études ont montré que les stratégies de nutrition employées par les radiolaires étaient étroitement liées à la morphologie de leur squelette [Matsuoka 2007 ; Suzuki et Sugiyama 2001]. Pour les trois taxons de polycystines (Nassellaria, Spumellaria et Collodaria), des observations *in vitro* ont mis en évidence l'utilisation des axopodes et pseudopodes des cellules dans le but de capturer leur proies (ciliés, flagellés et bactéries). Le mécanisme se divise en trois phases, (1) l'extension des axopodes/pseudopodes, (2) la capture de la proie piégée par les axopodes, (3) la rétractation des axopodes/pseudopodes pour ramener la proie vers la cellule du radiolaire. Pour les lignées symbiotiques, cette source peut provenir de leur endo-photo-symbiontes [Gast et Caron 2001]. Nos connaissances du cycle de vie des radiolaires restent encore très limitées car ces organismes sont actuellement non-cultivables *in vitro* et l'observation de plusieurs générations n'a pu être faite. On estime cependant que la durée de vie moyenne des radiolaires serait de deux à quatre semaines [W. S. Johnson et Allen 2012]. À ce jour, seule la reproduction asexuée, par fission binaire, a pu être observée, mais l'analyse de petites cellules appelées « *swarmers* » car bi-flagellés chez ces protistes, suggère l'existence d'une reproduction sexuée [Yuasa et Takahashi 2016]. Selon les espèces considérées, la distribution des radiolaires dans la colonne d'eau s'étend de la surface jusqu'à plusieurs centaines, voir milliers, de mètres de profondeur [Suzuki et Not 2015]. Cependant les radiolaires qui présentent des microalgues symbiotiques sont restreints aux couches superficielles des océans où la lumière pénètre. Du fait de la fragilité de ces organismes (en particulier de ceux qui forment des colonies) et de la difficulté à les échantillonner, nos connaissances sur ces lignées restent très limitées. Néanmoins, à l'aide de techniques d'imagerie *in situ*, de récents travaux démontrent l'abondance et l'importante biomasse que représentent les radiolaires de grande taille ($>$600 μm), en particulier en surface dans les zones intertropicale pour les larges colonies photosymbiotiques [Tristan Biard, Stemmann et al. 2016].

À ce jour, seulement trois projets de séquençage de génome de Rhizaria sont référencés sur la banque de données JGI (*Joint Genome Institute*) [Burki et Keeling 2014]. Éloignés des radiolaires d'un point de vu phylogénétique, deux génomes appartiennent au groupe des Cercozoa (*Bigelowiella natans* et *Paulinella chromatophora*) [Sibbald et Archibald 2017] et le troisième appartient à celui des Foraminifera (*Reticulomyxa filosa*) [Glöckner et al. 2014]. En parallèle, au delà des séquences des ADN ribosomaux 18S et 28S qui sont des gènes marqueurs pour la phylogénie et la phylogéographie [Krabberød, Orr et al. 2017 ; Takahashi et al. 2004 ; Zettler, Sogin et Caron 1997], des projets de transcriptomiques ont permis de générer des premières données génomiques pour les radiolaires autorisant un début d'exploration fonctionnelle sur la biologie de ces organismes [Balzano et al. 2015]. Cette étude a analysé les données de transcriptomique pour les Rhizaria suivants : un Acantharia (*Amphilonche elongata*), un Phaeodaria (*Aulacantha scolymantha*) et deux Polycystinea (*Collozoum* sp. et *Spongosphaera streptacantha*) [Balzano et al. 2015]. Les résultats obtenus révèlent que la présence de lectine de type C, protéine impliquée dans les processus de symbiose chez les eucaryotes, était spécifique aux lignées décrites comme symbiotiques (*i.e. Collozoum* sp., *Spongosphaera streptacantha*, *Amphilonche elongata*) et absente chez *Aulacantha scolymantha* (Phaeodaria non-symbiotique), caractérisant ainsi un peu plus les mécanismes symbiotiques en place chez ces organismes non-modèles*. L'étude indépendante des différents acteurs en place dans ces symbioses apparaît déterminante afin de comprendre les détails du fonctionnement des symbioses, de discerner les rôles des différents partenaires symbiotiques, mais également d'accéder à l'histoire évolutive de ces interactions.

1.3.4 Les dinoflagellés, des symbiontes communs

Les dinoflagellés sont des organismes unicellulaires eucaryotes appartenant à la lignée monophylétique* des Alveolata [Bachvaroff et al. 2014]. Les dinoflagellés furent décrits pour la première fois en 1753 par Henry Baker, naturaliste anglais du XVIII^{ème} siècle [Baker 1753]. Leur apparition est estimée à la période du Cambrien, il y a plus de 500 millions d'années [Raven 1998]. D'un point de vue morphologique, les dinoflagellés du grec δῖνος (« tournoiement ») et du latin flagellum (« fouet »)

possèdent deux flagelles : le flagelle transversal situé dans le cingulum (sillon équatorial), et le flagelle longitudinal situé dans le sulcus (sillon longitudinal). Certains possèdent une thèque et d'autres n'en développent pas et sont dits « nus ». La thèque correspond à un enchevêtrement de plaques cellulosiques rendant la surface de la cellule rigide (Figure 1.17). La cellule de dinoflagellées se décompose en deux parties : l'épithèque et l'hypothèque dont la délimitation est marquée par le cingulum. La forme du cingulum (ascendant, fermé, descendant ou croisé) est notamment un critère morphologique de différenciation des dinoflagellés. La taille d'une cellule de dinoflagellés peut varier de quelques micromètres à plus de 2 mm pour les espèces les plus grandes (*e.g. Noctiluca scintillans*). Ce groupe comprend une large variété d'espèces et des nouvelles espèces sont découvertes chaque année [Gómez 2014]. Huit classes de Dinophyta sont reconnues : les Gonyaulacales, les Prorocentrales, les Gymnodiniales, les Peridiniales, les Suessiales, les Noctilucales, les Syndiniales et les Blastodiniales [Lin 2011]. Certaines études distinguent le groupe des « *core* » dinoflagellés (regroupant les Peridiniales, Symbiodiniaceae, Gonyaulacales, Gymnodiniales, Noctilucales, Syndiniales) dont le caractère dérivé partagé est un noyau (appelé « dinocaryon ») qui se caractérise par des chromosomes maintenus condensés à tous les stades du cycle cellulaire [Janouškovec et al. 2016 ; Rizzo 2003] (Figure 1.18).

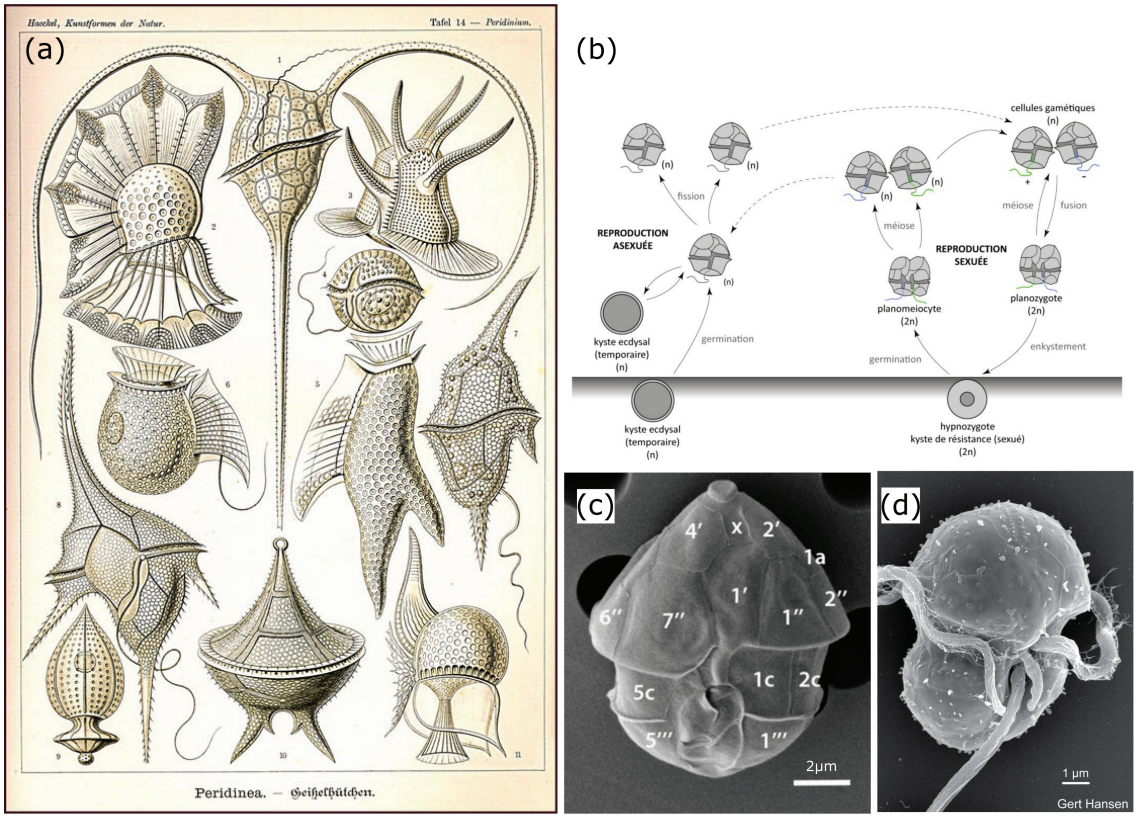


FIGURE 1.17 – Illustrations de Dinophyta. (a) Planche réalisée par Ernst Haeckel [HAECKEL 1904] dans laquelle sont représentées onze espèces de dinoflagellées. (b) Schéma du cycle de reproduction d'un dinoflagellé (http://archimer.ifremer.fr/doc/00191/30231/28647.pdf). (c) Observation au microscope à balayage électronique (MBE) d'une cellule de *Brandtodinium nutricula* (Peridiniales) dont les plaques de la thèque sont indicées [PROBERT et al. 2014]. (d) Une autre cellule de *Symbiodinium natans* observée au MBE [G. HANSEN et DAUGBJERG 2009]

Le cycle de vie des dinoflagellés alterne entre reproduction sexuée et asexuée, et pour certaines espèces entre phases pélagiques et benthiques [VARGAS, ZANINETTI et al. 1997]. Lorsque les conditions environnementales deviennent défavorables, les dinoflagellés forment des kystes qui correspondent à des stades benthiques au cours desquels la cellule s'entoure d'une paroie épaisse qui la préserve de l'environnement extérieur. Il existe deux formes de kystes : les kystes ecdysial (haploïdes, n chromosome) et les kystes de résistance (diploïdes, 2n chromosomes). Lorsque les conditions environnementales deviennent favorables, les kystes germent et la phase de vie pélagique des dinoflagellés commence (Figure 1.17). La grande variété de modes de vie des dinoflagellés permet d'expliquer leur succès écologique dans l'environnement [MURRAY et al. 2016]. Certaines espèces de dinoflagellés possèdent un chloroplaste*

fonctionnel et sont autotrophes*, d'autres n'ont pas de chloroplaste* ou seulement un chloroplaste* rémanent non-fonctionnel et sont hétérotrophes*. Finalement un certains nombre d'espèces sont mixotrophes* à la fois constitutifs ou non-constitutifs selon les espèces considérées [Mitra et al. 2016]. Indépendamment de leur mode trophique, certaines espèces de dinoflagellés peuvent être impliqués dans des relations de parasitisme ou de mutualisme (*e.g.* *Symbiodinium* spp.) avec des espèces de coraux (Cnidaria), d'éponges (Porifera) et même de protistes Rhizaria (*e.g.* foraminifères ou polycystines) [Carlos et al. 1999 ; T. LaJeunesse 2002 ; Probert et al. 2014] [Stal et Cretoiu 2016].

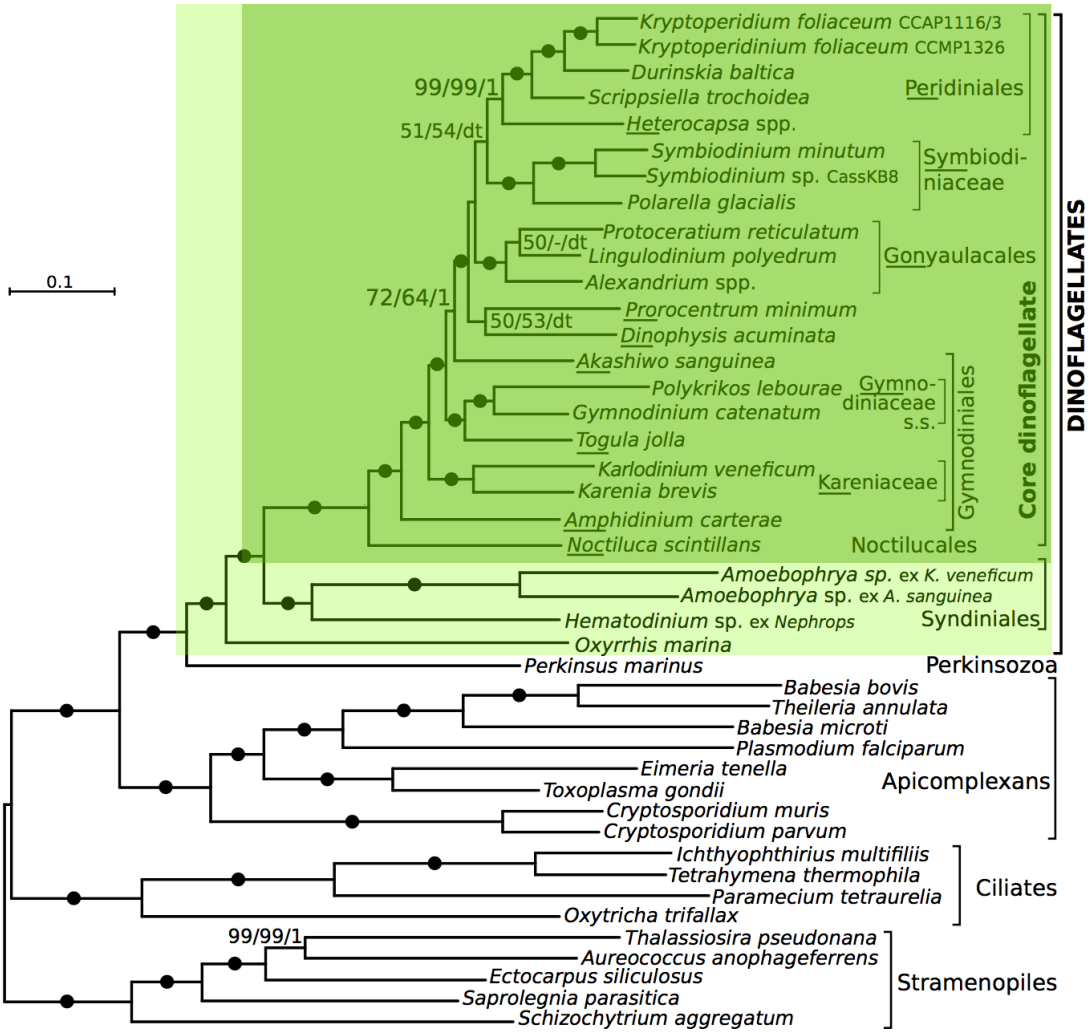


FIGURE 1.18 – Phylogénie moléculaire des dinoflagellés basée sur l'alignement multiple de 101 séquences protéiques pour 43 taxons distincts dont 25 espèces de dinoflagellés (21 en vert clair et 4 en vert foncé) [JANOUŠKOVEC et al. 2016]. Les « *cores* » dinoflagellés ont été encadrés en vert foncé. L'arbre a été obtenu par une méthode de maximum de vraisemblance (logiciel RAxML), et les points indiquent des noeuds avec des valeurs de boostrap de 100%.

Certaines espèces de dinoflagellés sont connues pour produire des composés chimiques dont certains peuvent être toxiques. Le DSP *diarrheic shellfish poisoning*, le CFP *ciguatera fish poisoning*, le PSP *paralytic shellfish poisoning*, le NSP *neurotoxic shellfish poisoning* ou encore l'AZP *azaspiracid shellfish poisoning*, ont des implications écologiques et économiques importantes notamment sur les activités de pisciculture et mytiliculture [D.-Z. WANG 2008]. L'impact des souches de dinoflagellés toxiques telles que *Alexandrium fundyense*, *Gymnodinium catenatum* ou encore

Pyrodinium bahamense sur l'environnement est particulièrement visible au cours des événements d'efflorescences (*harmful algal bloom* (HAB)) tels que les « marées rouges » (*red tides*) qui ont généralement lieu durant la saison estivale lorsque la concentration de nutriments à la surface des océans devient plus élevée. Les toxines sécrétées, comme la saxitoxine (une neurotoxine 100 000 fois plus puissante que la cocaïne), ont pour effet l'empoisonnement menant à la mort des organismes avoisinants leur environnement [Hackett et al. 2013 ; Stüken et al. 2011]. Cette capacité physiologique des dinoflagellés à secréter des toxines pourrait leur permettre de se protéger contre les prédateurs, alors que d'autres souches emploient la bioluminescence pour se protéger comme *Noctiluca scintillans* [HADHAZY 2009 ; Valiadi et Iglesias-Rodriguez 2013]. Le DMSP (diméthylsulfoniopropionate) est un autre composé chimique que certains dinoflagellés sont capables de produire. Cette molécule est connue pour ses propriétés osmolytiques agissant comme antioxydants contre les stress environnementaux, mais aussi pour son implication dans le cycle biogéochimique du soufre dans les océans [Gutierrez-Rodriguez et al. 2017]. Il a été montré que les concentrations en DMSP des dinoflagellés photosynthétiques impliqués dans des symbioses sont plus élevées que la moyenne (170–702 mmol L^{-1} contre 0.1^{-23} mmol L^{-1}) [Gutierrez-Rodriguez et al. 2017 ; Sunda et al. 2002]

Du fait de l'impact important des dinoflagellés sur l'environnement et de leur rôle dans diverses symbioses marines, la communauté scientifique s'est efforcée de comprendre les bases génomiques des mécanismes biologiques impliqués dans les propriétés fonctionnelles des dinoflagellés [Murray et al. 2016]. Dans un contexte où l'observation du phénomène de blanchissement des coraux est de plus en plus fréquente et où de nombreuses études s'intéressent à ce phénomène, les trois seuls génomes de dinoflagellés disponibles à ce jour sont : *Symbiodinium minutum* [Shoguchi et al. 2013] *Symbiodinium kawagutii* [Lin et al. 2015] et *Symbiodinium microadriaticum* [Aranda et al. 2016]. Ils appartiennent tous au même genre connu pour être le photo-symbionte indispensable à la survie des coraux. Ces génomes possèdent des tailles relativement petites (1 à 5 Gb pour les formes haploïdes) par rapport aux autres espèces de dinoflagellés qui varient de 1.5 Gb à 112 Gb [Aranda et al. 2016 ; Murray et al. 2016] (Figure 1.8). En outre, dans le cadre de l'étude des

impacts économiques et écologiques des dinoflagellés, des approches de transcriptomique ont été réalisées pour comprendre les processus fonctionnels impliqués et caractériser leur toxicité [Kohli et al. 2015], ou encore accroître nos connaissances sur les mécanismes de mise en place et de maintien de la symbiose chez les coraux [Sabourault et al. 2009]. Ainsi, que ce soit grâce aux études isolées ou à grande échelle (*i.e.* MMETSP [Keeling, Burki et al. 2014]), plus d'une centaine de transcriptomes de dinoflagellés, pour plus de 40 espèces différentes ont été générés à ce jour [Sibbald et Archibald 2017]. Ces jeux de données nous offrent une opportunité, rare pour un groupe de protistes, de mieux appréhender l'histoire évolutive grâce à des analyses phylogénomiques* [Janouškovec et al. 2016]. Compte tenu de l'importance des dinoflagellés dans les relations symbiotiques avec les Rhizaria du plancton marin, et en particulier les radiolaires [Decelle, Colin et Foster 2015 ; Probert et al. 2014], ces jeux de données représentent une possibilité inédite d'entamer la caractérisation, sur la base d'analyses de génomique fonctionnelles, du fonctionnement de ces relations, des interactions cellulaires et physiologiques qui ont lieux entre les partenaires impliqués.

1.4 Objectifs de la thèse

Cette thèse de bioinformatique a pour objectif l'étude de l'association symbiotique entre les radiolaires et les dinoflagellés à partir des transcriptomes de ces organismes en condition de symbiose. Ce travail de doctorat contribuera à une meilleure compréhension des mécanismes d'adaptation fonctionnelle et évolutive des organismes photosymbiotiques marins, et fournira des connaissances sur la physiologie des radiolaires ainsi que de leurs photosymbiontes dinoflagellés. Ces travaux ont été effectués en deux étapes : (1) le développement et mise en place opérationnelle d'une chaîne d'analyses pour effectuer l'assemblage *de novo* de jeux de données RNA-seq (assemblage de transcrits sans génome de référence), obtenus à partir du transcriptomes de dinoflagellés en phase libre d'une part et de radiolaires en symbiose avec des dinoflagellés d'autre part ; (2) l'identification des transcrits potentiellement impliqués dans les mécanismes qui régissent ces associations symbiotiques (*e.g.* com-

munication inter-espèces, transports membranaires, ou encore détection de transcrits non-annotés fonctionnellement).

Chapitre 2

Mise en place d'une chaîne d'analyses dédiée à l'étude *de novo* des transcriptomes d'organismes non-modèles

Afin d'étudier les mécanismes fonctionnels régissant la symbiose entre radiolaires et dinoflagellés, j'ai consacré la première partie de ma thèse à la mise en place d'une chaîne d'analyses pour l'assemblage *de novo* et le traitement de transcriptomes d'organismes non-modèles*. Cette chaîne a d'abord été élaborée à partir de l'analyse d'un des deux partenaires de cette symbiose, les dinoflagellés, pour lesquels, au moment de mon étude, davantage de données génomiques étaient disponibles comparativement à l'hôte. De plus, une littérature plus fournie, abordant des aspects de génomique fonctionnelle, est disponible pour les dinoflagellés. Dans le contexte global de mon projet de thèse, ces jeux de données et les informations relatives à la biologie des dinoflagellés m'ont permis de mettre en place un protocole d'analyses bioinformatiques à partir d'une sélection d'outils adaptés. Ce chapitre décrit l'élaboration ainsi que la mise en oeuvre de ce protocole.

La chaîne d'analyses développée (Figure 2.1) est composée de 4 étapes. La première étape correspond au prétraitement des séquences courtes (lectures) issues du séquençage des ARN. J'ai choisi Trimmomatic [Bolger, Lohse et Usadel 2014] pour effectuer le nettoyage des séquences (*i.e.* retrait des séquences de mauvaise qualité). Les paramètres ont été fixés comme suit : `MINLEN:32` (taille minimale des séquences en sortie de 32 pb (paires de bases)), `SLIDINGWINDOW:10:20` (retrait des séquences dont le score phred* d'une fenêtre de 20 pb est inférieure à 10). Cependant, l'utilisateur est libre de modifier ces paramètres grâce au fichier de configuration de la chaîne d'analyses. L'outil FastQC (`https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`) a été intégré afin d'évaluer la qualité de l'étape de prétraitement et la génération de rapports visuels. Pour l'élaboration de cette première étape, j'ai eu l'opportunité d'encadrer le stage de Master 1 de Anita Annamale (étudiante dans le master Biologie et Informatique de l'Université Paris 7 Diderot) en 2015 (de mars à juillet, c'est à dire pendant 5 mois). L'objectif du stage était de déterminer dans quelle mesure le prétraitement des lectures de façon drastique avant d'effectuer les assemblages avait un impact sur le temps d'assemblage ainsi que sur le taux de transcrits chimériques* obtenus. Au cours de son stage, je l'ai guidé et accompagné dans ses recherches bibliographiques sur les outils existants, dans les tests des outils pré-sélectionnés ainsi qu'à la création des scripts

Python constituant le module final qui a été déposé sur la plateforme GitHub (https://github.com/upmcgenomics/PREMSEQ).

La deuxième étape de la chaîne d'analyses est la phase d'assemblage *de novo* pour laquelle le programme Trinity [Grabherr et al. 2011] est utilisé avec ses paramètres par défaut. Trinity est par ailleurs largement employé pour la reconstruction de transcriptome sans référence de protistes marins (*e.g.* [Kodama et al. 2014]). Pour la suite du traitement, j'ai choisi d'intégrer une troisième étape afin d'évaluer la qualité des contigs* assemblés. Le programme Transrate [Smith-Unna et al. 2016] permet à l'utilisateur d'obtenir un résumé des caractéristiques des contigs*. Cette étape combine le calcul des métriques* caractérisant l'assemblage comme le nombre de contigs*, la valeur de N50* des contigs*, ou le taux de de bases GC avec l'alignement des lectures sur les contigs* assemblés (*mapping**) à l'aide du programme Bowtie2 (paramètres par défaut) [Langmead et Salzberg 2012]. La quatrième et dernière étape de la chaîne d'analyses consiste en la prédiction des peptides associés aux contigs* basée sur la détection de motifs protéiques connus. Cette étape est effectuée grâce au programme Transdecoder [Haas et al. 2013] dont les paramètres sont laissés par défaut. L'étape optionnelle permettant de filtrer les ORFs* (*open reading frames*) détectées dans les séquences de contigs* n'a pas été utilisée. Cette étape permet de valider les ORFs* sur la base d'homologies avec les banques de données SwissProt ou Pfam ce qui pourrait limiter la potentielle découverte de nouveaux domaines protéiques dont les ORFs* ne sont pas référencés dans ces banques de données. Les séquences de peptides sont ensuite annotées fonctionnellement via Interproscan 5 [Jones et al. 2014].

L'ensemble des étapes et outils que je viens de décrire ont été intégrés dans une chaîne d'analyses qui a été implémentée en langage de programmation Python (version 2.7.5+) sous la forme de modules indépendants dans le but de conserver une flexibilité dans l'enchaînement des quatre étapes (Figure 2.1). J'ai choisi d'utiliser le système de gestion de workflow Snakemake (https://bitbucket.org/johanneskoester/snakemake) afin de rendre possible l'utilisation d'un fichier de paramètres pour les utilisateurs. Le système Snakemake permet une gestion automatique des fichiers d'information (*log*) et la gestion de la parallélisation des différents

programmes qui composent la chaîne d'analyses. La structure modulaire ainsi que la syntaxe simplifiée du système de codage par Snakemake permet également l'intégration future de nouvelles étapes et/ou de nouveaux programmes de traitement en fonction des besoins et des questionnements biologiques à venir. Par exemple, on peut imaginer un module dédié à la détection des ARNr (programme SortMeRNA [Kopylova, Noé et Touzet 2012]) en amont de l'étape 1 afin de séparer les reads provenant d'ARNr et d'ARNm pour les traiter de manière indépendante. Enfin, il m'est apparu nécessaire de partager cette chaîne d'analyses afin d'assurer la reproductibilité des études effectuées. Les fichiers sources et scripts de la chaîne d'analyses sont ainsi disponibles sur GitHub : `https://github.com/arnaudmeng/dntap`. Elle est également désormais disponible et fonctionnelle sur le serveur de la plateforme ABIMS (*Analysis and Bioinformatics for Marine Science*) de la station biologique de Roscoff.

J'ai choisi le programme Trinity [Grabherr et al. 2011] parmi les différents outils dédiés à l'assemblage transcriptomique *de novo*. En effet, les recherches bibliographiques sur la comparaison de ces outils ainsi que les tests effectués au cours de mon stage de Master 2 et au début de ce travail de thèse, m'ont convaincu que Trinity était le choix approprié pour le traitement des données des organismes non-modèles* à étudier dans le cadre de ma thèse (Figure 2.2). Durant mes tests j'ai comparé les deux assembleurs transcriptomiques *de novo* les plus utilisés en 2014 (Trinity [Grabherr et al. 2011] et Velvet/Oases [Schulz et al. 2012 ; Zerbino et Birney 2008]) afin d'évaluer leur performance (temps de calcul et quantité de mémoire vive utilisée, Annexe A) et la qualité de l'assemblage résultant (métriques* d'assemblages, Annexe B) sur quatre transcriptomes de dinoflagellés. Mon choix s'est porté sur ces organismes non-modèles* car ils représentent des souches impliquées dans les associations symbiotiques avec des lignées de Rhizaria et auxquelles j'avais accès (Figure 2.2). Les résultats montrent que Trinity a tendance à assembler moins de contigs* que Velvet/Oases mais de plus grandes tailles (Figure 2.2 A et B). Par ailleurs, j'ai effectué une filtration des contigs* dont la valeur de FPKM* était inférieure à 1 (valeur empirique, comm. pers. Erwan Corre). Plus cette valeur est élevée plus l'expression estimée du contig* est importante. Les résultats montrent

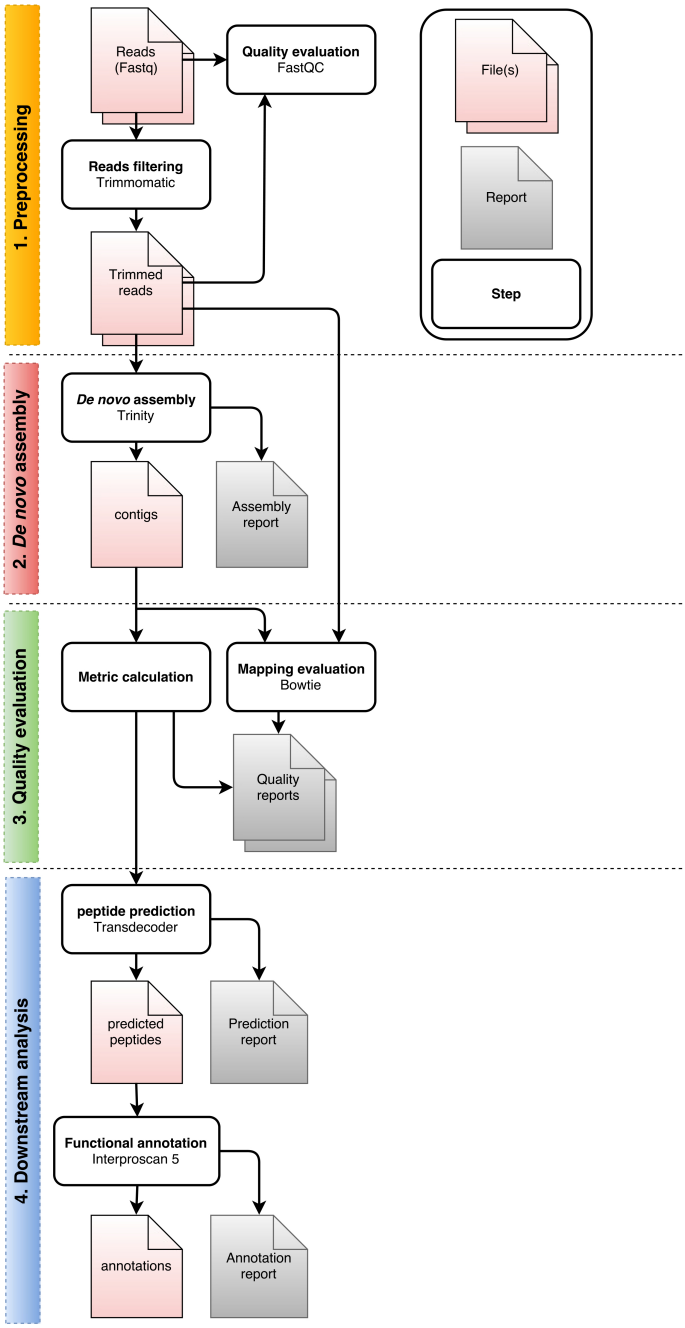


FIGURE 2.1 – Diagramme de la chaîne d'analyses dédiée à l'assemblage *de novo* de transcriptomes dans sa dernière version (juillet 2017). Les cinq étapes sont représentées par différentes couleurs : 1. La phase de pré-traitement des reads et d'évaluation de la qualité des séquences (en jaune), 2. La phase d'assemblage *de novo* (en rouge), 3. La phase d'évaluation des contigs* assemblés (en vert), 4. La phase de traitement des contigs* assemblés qui comprend la prédiction des domaines fonctionnels ainsi que l'annotation fonctionnelle des séquences des domaines prédits. Poster présenté lors du congrès international *European Council of Computational Biology* 2016, La Haye, Pays-Bas. Annexe D

que l'impact de cette filtration est plus important pour les transcriptomes assemblés avec Velvet/Oases et entraîne ainsi une plus grande perte par filtration des contigs* (Figure 2.2 C). Enfin, les taux de ré-alignements des lectures (*mapping**) sur les séquences de contigs* sont toujours plus élevés pour les assemblages issus de Trinity (Figure 2.2 D).

Outre la valorisation lors d'une communication affichée dans un congrès international (*European Council of Computational Biology* 2016, La Hayes, Pays-Bas), la méthodologie mise en place a été par ailleurs utilisée lors d'une collaboration avec Gaëlle Lelandais de l'institut Jacques Monod (Université de Paris 7 Diderot, Paris) et François-Yves Bouget de l'Observatoire océanologique de Banyuls-sur-Mer (Université Pierre et Marie Curie, Banyuls-sur-Mer). Le résultat de ce travail a donné lieu à la publication d'un article [Botebol et al. 2017] (Annexe C) qui porte sur la mise en évidence de l'adaptation fonctionnelle d'une souche de pico-algue verte (*Ostreococcus* sp. RCC802, Chlorophyta) aux conditions d'un environnement carencé en fer. Au cours de cette collaboration, j'ai participé au traitement des données, en particulier à l'étape d'assemblage *de novo* des transcriptomes d'*Ostreococcus* sp. RCC802. Dans le cadre de la continuité de cette collaboration, j'ai également eu l'occasion d'encadrer le stage de Master 1 de Quentin Letourneur (étudiant dans le master Biologie et Informatique de l'Université Paris 7 Diderot) de février à juillet 2016 (6 mois). Son stage a porté sur la mise en place d'un protocole permettant d'optimiser la sélection de transcrits reconstruits sur la base de leur qualité et en tenant compte de la dynamique d'expression des gènes. Je l'ai accompagné dans ses recherches bibliographiques et sur la mise en place du protocole d'analyses bioinformatiques.

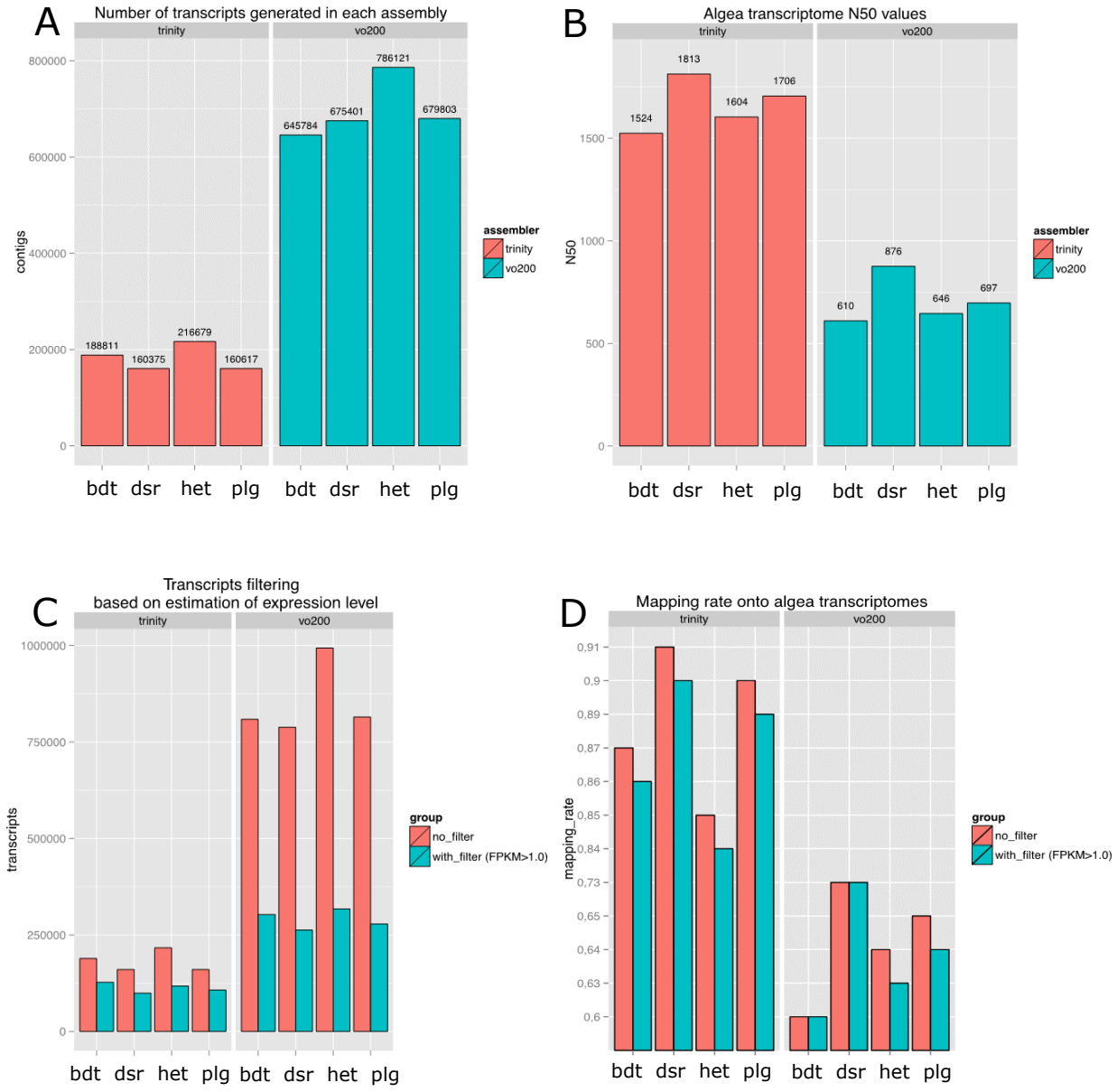


FIGURE 2.2 – Résultats des assemblages de quatres transcriptomes de dinoflagellés (*Brandtodinium nutricula* (bdt), *Gymnoxanthellae radiolarae* (dsr), *Heterocapsa* sp. (het) et *Pelagodinium beii* (plg)) assemblés avec Trinity (panneaux de gauche) et Velvet/Oases (panneaux de droite). (Extrait de mon rapport de stage de Master 2, 2014). (A) Le nombre de contigs* assemblés avec chaque assembleur ; (B) les valeurs de N50* ; (C) le nombre de contigs* avant (en rouge) et après (en bleu) le filtrage (FPKM*>1) ; (D) Les taux de ré-alignement (*mapping**) des reads sur les séquences des contigs* non filtrés (en rouge) et filtrés (en bleu).

Chapitre 3

Étude transcriptomique des dinoflagellés et recherche de séquences génomiques liées à la symbiose

Dans le cadre de ma thèse, j'ai utilisé d'une part des données issues du projet de séquençage RNA-seq MMETSP (*Marine Microbial Eukaryote Transcriptome Sequencing Project*) [Keeling, Burki et al. 2014] incluant 46 souches de dinoflagellés (provenant de 123 échantillons de dinoflagellés) correspondant à 56 transcriptomes, et d'autres part des données encore non publiées obtenues à partir de 4 souches de dinoflagellés maintenues dans la collection de culture de microalgues de Roscoff (RCC) à la Station Biologique de Roscoff et correspondant à 4 transcriptomes.

L'ensemble des jeux de données provenant de la MMETSP ainsi que de la RCC représente des transcriptomes pour 47 espèces distinctes de dinoflagellés (Article Meng et al *submitted*, p.55). Ces espèces correspondent à 35 genres, 19 familles et 11 ordres taxonomiques. Cette couverture taxonomique de 11 ordres est à mettre en regard des 21 ordres actuellement reconnus selon la banque de données WoRMS *World Register of Marine Species*, (`http://www.marinespecies.org/`), dont 19 comportent des espèces référencées dans la base de données AlgaeBase (`http://www.algaebase.org`). Ainsi, je dispose de représentants pour la moitié des différents ordres de dinoflagellés. Pour chaque espèce étudiée des recherches ont été effectuées pour 11 traits fonctionnels distincts, tels que la production de toxines, la possession de thèque ou encore la capacité à être symbiotique (Article Meng et al *submitted*, p.55). Cet ensemble de données constitue une base me permettant d'explorer la diversité fonctionnelle des dinoflagellés (Figure 3.1). Il est important de noter cependant, que ces jeux données sont issus d'échantillons de dinoflagellés en culture et ne peuvent donc pas être considérés comme représentatifs de l'environnement marin.

Au total 57 (des 60 transcriptomes disponibles) ont été assemblés à l'aide de la chaîne d'analyses développée dans ce travail de thèse (cf. chapitre 2). Sur les 60 transcriptomes possibles, 3 n'ont pu être assemblés par Trinity suite à une erreur de format provenant des données. Afin d'aller au-delà des études de phylogénomique* (Figure 3.1, Janouškovec et al. 2016), j'ai opté pour une comparaison globale de l'ensemble des transcrits produits, que ceux-ci soient annotés fonctionnellement ou non, grâce à des réseaux de similarité de séquences (SSN*). L'exploration de ces réseaux m'a permis d'extraire des sous-ensembles de transcrits clés (*i.e.* des composantes connexes, CCs) tels que ceux conservés chez l'ensemble des dinoflagellés

considérés dans cette étude et ceux potentiellement impliqués dans les traits fonctionnels identifiés tels que de la symbiose. Cette étude fait l'objet d'un article en premier auteur soumis le 29 juin 2017 à la revue *Molecular Ecology*.

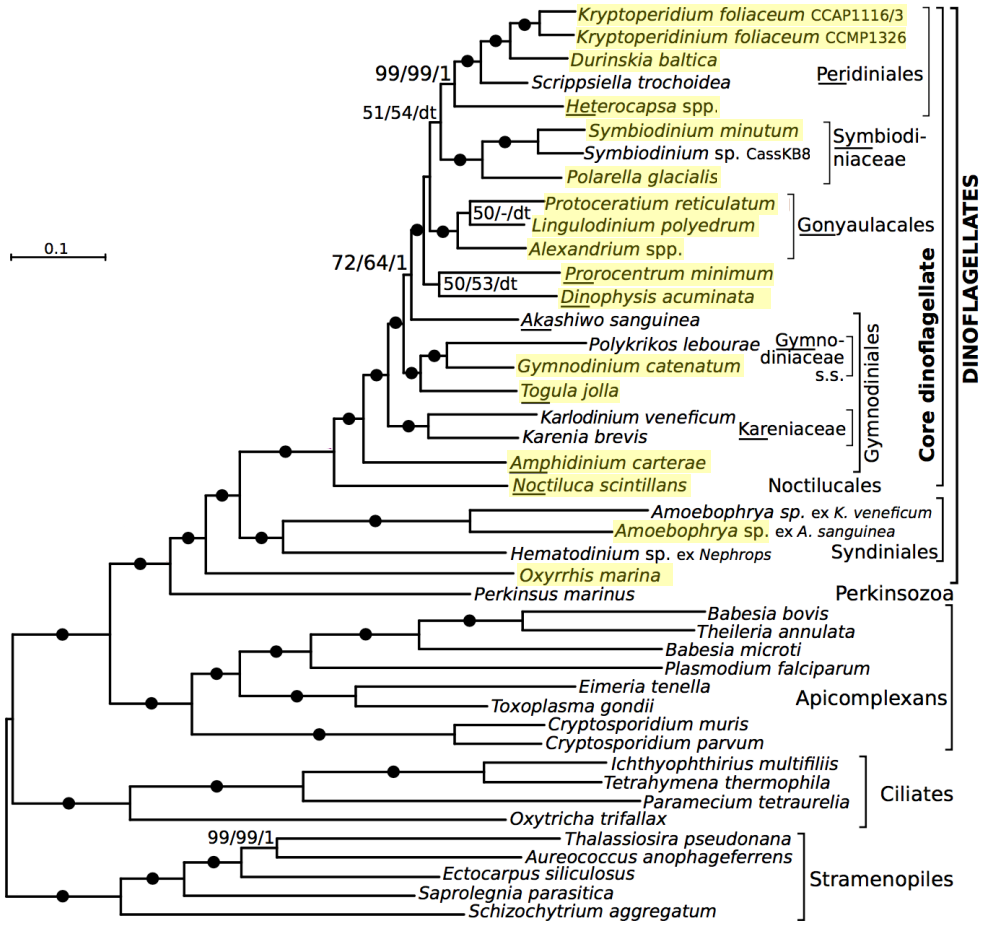


FIGURE 3.1 – Positionnement phylogénétique des Dinophyta utilisées dans cette étude (modifié de JANOUŠKOVEC et al. 2016). Sur l'arbre apparaissent 17 des 47 espèces que je considère dans mon étude. Les auteurs de cette phylogénie moléculaire n'ont tenu compte que d'un sous-échantillonnage des espèces (*i.e.* 26 espèces) présentent dans la base de données MMETSP.

ANALYSIS OF THE GENOMIC BASIS OF FUNCTIONAL DIVERSITY IN DINOFLAGELLATES USING A TRANSCRIPTOME-BASED SEQUENCE SIMILARITY NETWORK

List of authors

Arnaud Meng (AM)[a]*, Erwan Corre (EC)[b], Ian Probert (IP)[c,d], Andres Gutierrez-Rodriguez (AGR)[e], Raffaele Siano (RF)[f], Anita Annamale (AAN)[g,h,i], Adriana Alberti (AAL)[g,h,i], Corinne Da Silva (CDS)[g,h,i], Patrick Wincker (PW)[g,h,i], Stéphane Le Crom (SLC)[a], Fabrice Not (FN)[c,d]†, Lucie Bittner (LB)[a]†

Affiliations

[a] Sorbonne Universités, UPMC Univ Paris 06, Univ Antilles Guyane, Univ Nice Sophia Antipolis, CNRS, Evolution Paris Seine - Institut de Biologie Paris Seine (EPS - IBPS), 75005 Paris, France
[b] CNRS, UPMC, FR2424, ABiMS, Station Biologique, Roscoff 29680, France
[c] CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, Roscoff, France
[d] Sorbonne Universités, Université Pierre et Marie Curie (UPMC) Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, Roscoff, France
[e] National Institute for Water and Atmospheric Research (NIWA) Ltd, Private Bag 14-901, Kilbirnie, Wellington, New Zealand
[f] Ifremer – Centre de Brest, DYNECO PELAGOS, F-29280 Plouzané, France
[g] CEA - Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France.
[h] CNRS, UMR8030, CP5706, Evry, France
[i] Université d'Evry Val d'Essonne, Evry, France

*Corresponding author. E-Mail: arnaud.meng@etu.upmc.fr. † These authors contributed equally to this work.

keywords

Transcriptomics | Protists | Molecular Evolution | Microbial Biology |

Genomics/Proteomics

running title

Functional diversity exploration of Dinophyta

ABSTRACT

Dinoflagellates are one of the most abundant and functionally diverse groups of eukaryotes. Despite an overall scarcity of genomic information for dinoflagellates, constantly emerging high-throughput sequencing resources can be used to characterize and compare these organisms. We assembled *de novo* and processed 46 dinoflagellate transcriptomes and used a sequence similarity network (SSN) to compare the underlying genomic basis of functional features within the group. This approach constitutes the most comprehensive picture to date of the genomic potential of dinoflagellates. A core proteome composed of 252 connected components (CCs) of putative conserved protein domains (pCDs) was identified. Of these, 206 were novel and 16 lacked any functional annotation in public databases. Integration of functional information in our network analyses allowed investigation of pCDs specifically associated to functional traits. With respect to toxicity, sequences homologous to those of proteins involved in toxin biosynthesis pathways (e.g. *sxtA1-4* and *sxtG*) were not specific to known toxin-producing species. Although not fully specific to symbiosis, the most represented functions associated with proteins involved in the symbiotic trait were related to membrane processes and ion transport. Overall, our SSN approach led to identification of 45,207 and 90,794 specific and constitutive pCDs of respectively the toxic and symbiotic species represented in our analyses. Of these, 56% and 57% respectively (*i.e.* 25,393 and 52,193 pCDs) completely lacked annotation in public databases. This stresses the extent of our lack of knowledge, while emphasizing

the potential of SSNs to identify candidate pCDs for further functional genomic characterization.

INTRODUCTION

Dinoflagellates are unicellular eukaryotes belonging to the Alveolata lineage (Bachvaroff *et al.* 2014). This group encompasses a broad diversity of taxa that have a long and complex evolutionary history, play key ecological roles in aquatic ecosystems, and have significant economic impacts (Murray *et al.* 2016; Janouškovec *et al.* 2016). The ecological success of dinoflagellates in the marine planktonic environment is assumed to be due to their ability to exhibit various survival strategies associated with an extraordinary physiological diversity (Murray *et al.* 2016). Nearly half of dinoflagellates have chloroplasts, but most of these are likely mixotrophic, combining photosynthetic and heterotrophic modes of nutrition (Jeong *et al.* 2010; Stoecker *et al.* 2017). Many dinoflagellates produce toxins and form long-lasting harmful algal blooms with deleterious effects on fisheries or aquaculture (Flewelling *et al.* 2005). Some species of the genus *Alexandrium* can produce toxins that effect higher trophic levels in marine ecosystems (*i.e.* copepods, fish) and are harmful to humans (Orr *et al.* 2013; Kohli *et al.* 2016; Murray *et al.* 2016). Members of the genus *Symbiodinium* are known to establish mutualistic symbioses with a wide diversity of benthic hosts, sustaining reef ecosystems worldwide (Goodson *et al.* 2001; Lin *et al.* 2015). Interactions between dinoflagellates and other marine organisms are extremely diverse, including (photo)symbioses (Decelle *et al.* 2015), predation (Jeong *et al.* 2010), kleptoplasty (Gast *et al.* 2007), and parasitism (Siano *et al.* 2011). Dinoflagellates have been highlighted as important members of coastal and open-ocean protistan communities based on environmental molecular barcoding surveys (Massana *et al.* 2015; Le Bescot *et al.* 2016) and the parasitic syndiniales in particular

45 have been identified as key players that drive *in situ* planktonic interactions in the ocean
46 (Lima-Mendez *et al.* 2015).

47 Along with metabarcoding surveys based on taxonomic marker genes,
48 environmental investigations of protistan ecology and evolution involve genomic and
49 transcriptomic data. Interpretation of such large datasets is limited by the current lack of
50 reference data from unicellular eukaryotic planktonic organisms, resulting in a high
51 proportion of unknown sequences (Caron *et al.* 2016; Sibbald & Archibald 2017). This is
52 particularly significant for dinoflagellates as this taxon remains poorly explored at the
53 genome level, with only three full genome sequences published so far (Shoguchi *et al.*
54 2013; Lin *et al.* 2015; Aranda *et al.* 2016). Their genomes are notoriously big (0.5 to 40x
55 larger than the human haploid genome) and have a complex organization (Jaeckisch *et*
56 *al.* 2011; Shoguchi *et al.* 2013; Murray *et al.* 2016). Consequently, most recent studies
57 investigating functional diversity of dinoflagellates rely on transcriptomic data to probe
58 these non-model organisms.

59 The Moore Foundation Marine Microbial Eukaryotic Transcriptome Sequencing
60 Project (MMETSP, http://marinemicroeukaryotes.org/, (Keeling *et al.* 2014)) provided the
61 opportunity to produce a large quantity of reference transcriptomic data (Sibbald &
62 Archibald 2017). Among the 650 transcriptomes released, 56 were from 24 dinoflagellate
63 genera encompassing 46 distinct strains (Keeling *et al.* 2014). This dataset constitutes a
64 unique opportunity to investigate the genomic basis of the major evolutionary and
65 ecological traits of dinoflagellates (Janouškovec *et al.* 2016). Performing a global analysis
66 of such a large dataset (~3 million sequences) is challenging and requires innovative
67 approaches. Most studies published so far have targeted specific biological processes and
68 pathways, focusing on a small subset of the available data (Meyer *et al.* 2015; Dupont *et*
69 *al.* 2015; Kohli *et al.* 2016). In one recent study a 101-protein dataset was used to produce
70 a multiprotein phylogeny of dinoflagellates (Janouškovec *et al.* 2016). As a large fraction

of the sequences produced in the MMETSP project do not have any distant homologues in current reference databases, almost half (46%) of the data remains unannotated.

With the advent of high-throughput sequencing technologies and its inherent massive production of data, sequence similarity network (SSN) approaches (Atkinson *et al.* 2009; Cheng *et al.* 2014; Méheust *et al.* 2016) offer an alternative to classical methods, enabling inclusion of unknown sequences in the global analysis (Forster *et al.* 2015; Lopez *et al.* 2015). In a functional genomic context, SSNs facilitate large-scale comparison of sequences, including functionally unannotated sequences, and hypothesis design based on both model and non-model organisms. For instance, SSN has been used to define enolase protein superfamilies and assign function to nearly 50% of sequences composing the superfamilies that had unknown functions (Gerlt *et al.* 2012). Here we used a SSN approach involving 57 *de novo* assembled transcriptomes from the MMETSP project as well as new transcriptomes of four recently described dinoflagellates to unveil the core-, accessory-, and pan-proteome of dinoflagellates and to define gene sets characteristic of selected functional traits.

RESULTS

Dataset metrics overview

For the 57 assembled transcriptomes (53 from the MMETSP dataset and 4 from this study) the average number of transcript sequences was 93,685 with a mean N50 of 878 bp and remapping rates exceeded 60% on average (Tab. 1). Protein coding domains were predicted from the transcript sequences for each transcriptome to build what we consider here as proteomes. We found a mean of 49,281 protein-coding domains per proteome (Tab. 1). Globally, more than half of the protein-coding domains matched with functional annotations in InterPro (58%: 750,480 of 1,283,775) of which 552,846 had an identified Gene Ontology (GO) annotation. All individually assembled transcriptomes,

96 derived proteomes and their corresponding functional annotations are available at
97 http://application.sb-roscoff.fr/project/radiolaria/.

98 A first version of the SSN was created based on the 57 proteomes. The filtration
99 steps performed according to determined optimal settings (*i.e.* alignments with a minimum
100 60% sequence identity) (Fig. S1, see methods) resulted in the removal of 11 proteomes.
101 The final network was composed of 1,275,911 vertices (protein-coding domains) linked by
102 6,142,013 edges (corresponding to a pairwise sequence identity value ≥ 60%). The
103 network consisted of 350,267 connected components (CCs) with 11,568 of these having
104 a size from 10 to 100 vertices (Tab. S1). It encompassed 46 proteomes having a mean of
105 60,661 protein-coding domains with an average length of 307 bp. According to InterPro
106 functional annotations, 50.5% (*i.e.* 176,958) of the CCs were composed of unannotated
107 sequences only.

108 **Identification of core / accessory / pan connected components**

109 Analyses of CC composition revealed that 252 CCs included protein-coding
110 domains from all 43 proteomes considered in this analysis (core CCs), 160,431 CCs
111 exclusively included protein-coding domains from a single proteome (accessory CCs), and
112 347,551 CCs corresponded to the pan proteome of the dinoflagellates included in the
113 analysis (Fig. 1*A*). We extrapolated the trend of the core proteome CC number using a
114 non-linear regression model. The best-fit function was $y = a / x$, with y the predicted
115 number of core CCs, x the number of proteomes and a an estimated parameter. For 2 to
116 43 proteomes, this model had a correlation of 0.97 to our data (p-value of estimated
117 parameter a < 2e-16). The extrapolation of the number of core CCs for 50, 60 and 70
118 proteomes were 170, 144 and 123 CCs respectively, without displaying a saturation to a
119 fixed number of core CCs. The Pielou diversity indices calculated to explore CC
120 composition had a mean value of 0.96, indicating not only that core CCs were composed

of all proteomes by definition, but that they were also evenly structured, rarely being dominated by a single proteome.

Functional annotation using InterPro revealed that 91,4% of core protein-coding domains matched to the InterPro database. According to GOslim functional annotations, core CCs had an important contribution of protein-coding domains annotated as "ribosomal proteins" having a role in RNA translation (*i.e*. 7,968 of 37,842 core protein-coding domains) (Fig. S2). Other main functional annotations occurring in core CCs were protein phosphorylation (1,752 protein-coding domains), proteins involved in signal transduction (1,133 protein-coding domains) and cell redox homeostasis (562 protein-coding domains), the rest being composed of a variety of functions represented by few protein-coding domains (Fig. S2). The 37,842 protein-coding domains belonging to the 252 core CCs were further analyzed by comparison to other reference databases to identify protein-coding domains that are shared among the 43 dinoflagellate proteomes. The proportion of matching protein-coding domains reached up to 12.5% (involved in 51 CCs), 79.6% (involved in 190 CCs) and 93.7% (involved in 236 CCs) against BUSCO (Simão *et al.* 2015), UniProtKB/Swiss-Prot and nr, respectively (Fig. 1*B*). A total of 16 CCs (*i.e.* 946 protein-coding domains) from the core proteome did not have a match in any of the databases explored (Fig. 1*B*) (Tab. S2). The 101 orthologous alignments used for a recent phylogeny of dinoflagellates (Janouškovec *et al.* 2016) were compared to the protein-coding domains from the 252 core CCs. Results show that 1606 protein-coding domains from 46 CCs matched with at least one of the 101 orthologous alignments (Fig. S3), and that domains from our 16 unknown core CCs were not included in these 46 CCs.

Functional trait investigations

In the SSN based on the 46 proteomes, the number of CCs exclusively composed of protein-coding domains from species tagged for a functional trait (*i.e.* trait-CCs) has been reported for each trait investigated (Tab. S3-S12), as well as the percentage of trait-

147 CCs that are annotated (*e.g.* an annotated trait-CC includes at least one functionally
148 annotated protein-coding domain). As expected considering the taxonomic coverage of
149 our dataset, the analysis revealed a maximum number of trait-CCs for the “chloroplast”
150 trait (336,099 CCs) and a minimum for the “parasitism” trait (826 CCs). The “chloroplast”
151 trait had the highest percentage of annotated trait-CCs (93%) while the “parasitism” trait
152 had the lowest (23%) (Fig. S4). Among the trait-CCs, a total of 5 “harmful for human” trait-
153 CCs regrouping protein-coding domains of 7 of 14 possible proteomes were detected.
154 Likewise, we found 2 “symbiosis” trait-CCs including 8 of 12 possible proteomes (Tab. S4
155 & S6).

156 **Identification of toxin related sequences in “harmful for human” trait-CCs**

157 Well-described proteins involved in dinoflagellate toxicity, the polyketide synthases
158 (PKS) and saxitoxins (STX) were sought within the “harmful for human” trait-CCs. 36
159 protein-coding domains homologous to PKS were identified in 17 “harmful for human” trait-
160 CCs (composed of a total of 45 protein-coding domains) (Tab. S13). On the other hand,
161 646 protein-coding domains homologous to PKS were found in 165 CCs (composed of a
162 total of 1,144 protein-coding domains) belonging to the contrasting non-“harmful for
163 human” trait-CCs. All protein-coding domains from trait-CCs in which PKS homologues
164 were detected (*i.e.* 45 + 1,144 = 1189 protein coding domains) had either a Thiolase-like
165 functional annotation (1,159 protein-coding domains), which corresponds to the
166 superfamily of KS enzyme domains of PKS, or lacked annotation (30 protein-coding
167 domains) according to the InterPro database. The *sxtA* and *sxtG* genes have been
168 reported to be involved in the STX biosynthesis pathway. Based on 117 *sxtA1-4* and 20
169 *sxtG* reference gene sequences, no target protein-coding domain matched to “harmful for
170 human” trait-CCs (Tab. S14). In contrast, 99 and 3 unique protein-coding domains were
171 identified for *sxtA1-4* and *sxtG* respectively, in non-“harmful for human” trait-CCs. *sxtA1-*
172 *4* hits involved protein-coding domains from 20 CCs (composed of 166 protein-coding

173 domains), and *sxtG* hits belonged to 1 CC composed of 3 protein-coding domains. Of the
174 166 *sxtA1-4* homologues, 156 protein-coding domains had InterPro annotations related to
175 *sxtA* domains (*i.e.* pyridoxal phosphate-dependent transferase, PKS, GNAT domains,
176 Acyl-CoA N-acyltransferase) and the remaining 10 protein-coding domains were
177 unannotated. A single InterPro functional annotation was found for the CC involving *sxtG*
178 homologues (of the 3 protein-coding domains forming the CC) and corresponded to an
179 amidinotransferase domain that is known as a *sxtG* protein domain (Tab. S14).

180 We investigated the GO functional annotations of “harmful for human” trait-CCs.
181 At the cellular component functional level, “membrane” and “integral component of
182 membrane”, protein-coding domains represented 51% (3017 out of 5,998) and 27% (1,646
183 out of 5,998) of annotated protein-coding domains, respectively (Fig. 2*A*). At the biological
184 process annotation level, 14% (1,672 out of 11,187) of protein-coding domains were linked
185 to “ion transport” (Fig. 2*A*). At the molecular function annotation level, 24% (5,151 out of
186 21,337) corresponded to “protein binding” protein-coding domains (Fig. 2*A*). Differential
187 composition of functional annotations between proteomes of “harmful for human”
188 compared to non-“harmful for human” species was investigated to unveil functions that
189 are more likely to be observed in the proteome of toxic species. The pair differences of
190 each function occurring in “harmful for human” and non-“harmful for human” trait-CCs
191 showed that “ion transport” protein domains occured 7 times more often in “harmful for
192 human” trait-CCs. We also noticed that pentatricopeptide repeat, C2 domain, P-loop
193 containing nucleoside triphosphate hydrolase, Pyrrolo-quinoline quinone beta-propeller
194 repeat, Quinonprotein alcohol dehydrogenase-like and Thrombospondin type 1 repeat
195 domains occurred from 1 to 2 times more often in “harmful for human” trait-CCs (Fig. 2*B*).

196 The core “harmful for human” trait-CCs (*i.e*. CCs that are composed of protein-
197 coding domains from most toxic species representatives) was investigated to reveal
198 functions that are shared among toxic species only (Fig. 2*C*). We identified 5 of such core

199 “harmful for human” trait-CCs, corresponding to a total of 49 protein-coding domains.
200 These core trait-CCs encompassed 7 of 14 toxic dinoflagellate proteomes considered in
201 our analysis. Not a single of these 49 protein-coding domains had a GO annotation. Based
202 on InterPro functional annotations, 3 of the 5 CCs are respectively composed of 14
203 “nucleotide-binding alpha-beta plait” protein-coding domains, 7 “P-loop containing
204 nucleoside triphosphate hydrolase” protein-coding domains and 8 “nucleotide-diphospho-
205 sugar transferase” protein-coding domains. Two of these 5 CCs were entirely composed
206 of 7 and 15 unannotated protein-coding domains. The taxonomic distribution of the 49
207 protein-coding domains is represented by: 30 protein-coding domains from the
208 *Alexandrium* genus (61%), 2 from *Dinophysis acuminata*, 4 from *Protoceratium*
209 *reticulatum*, 6 from *Gambierdiscus australes*, 4 from *Pyrodinium bahamense* and 3 from
210 *Lingulodinium polyedra*.

211 31,496 “harmful for human” trait-CCs, composed in total of 70,359 protein-coding
212 domains, completely lacked functional annotations according to Interproscan. Alignments
213 (with an e-value of 1e-3) to nr database revealed 6,103 hits including 283 protein-coding
214 domains with sequence identity higher than 80%. These 283 protein-coding domains
215 originated from 157 CCs which were composed of 359 protein-coding domains in total.
216 Finally, we identified functions for 157 “harmful for human” trait-CCs while 25,393 “harmful
217 for human” trait-CCs remained without functional annotation.

218 **Exploration of “symbiosis” trait-CCs**

219 A large range of dinoflagellates, including symbiotic and non-symbiotic species,
220 express genes identified in the literature as potentially involved in symbiotic processes
221 (Tab. S15). 150 of these gene sequences were compared to protein-coding domains
222 included in “symbiosis” trait-CCs. Alignments revealed 8 protein-coding domains from 5
223 “symbiosis” trait-CCs. These 8 protein-coding domains were identified as belonging to
224 proteins involved in symbiosis establishment (nodulation protein noIO and

225 phosphoadenosine phosphosulfate reductase), cell recognition processes (merozoite
226 surface protein), and highlighted in cnidarian-algal symbiosis (peroxiredoxin, ferritin) (Tab.
227 S15). Similarly, 71 protein-coding domains (spread across 21 CCs) matching symbiosis
228 related query sequences were found in non-“symbiosis” trait-CCs. Functions of these 71
229 protein-coding domains are involved in symbiosis establishment (P-type H+-ATPase,
230 phosphoadenosine phosphosulfate reductase), cell recognition processes (merozoite
231 surface protein 1) and exposed in cnidarian-algal symbiosis (superoxide dismutase,
232 catalase, peroxiredoxin, glutathione peroxidase, g-glutamylcysteine synthetase).
233 We explored GO functional annotations from all “symbiosis” trait-CCs (Fig. 2*D*). At
234 the cellular component level, 83% of annotated protein-coding domains (11,103 out of
235 13,298) were “membrane” protein-coding domains. At the biological process level, 21%
236 (4,131 out of 19,380) were “ion transport” protein-coding domains and 18% were involved
237 in “protein phosphorylation”. At the molecular function level, 39% (15,585 out of 39,293)
238 were annotated as “protein-binding” protein-coding domains, 10% and 9.9% were involved
239 in “ion channel activity” and “calcium ion binding”, respectively. Pair differences of
240 normalized counts for each possible function that appeared in proteomes of both symbiotic
241 and non-symbiotic species revealed 4 annotations that occured 2 to 10 times more within
242 “symbiosis” trait-CCs. These were protein-coding domains containing an ion transport
243 domain, ankyrin repeat domains, EF-hand domain and zinc finger, and CCCH-type
244 annotations were 9.7, 4, 2.4, and 2.3 times more numerous, respectively, in “symbiosis”
245 trait-CCs (Fig. 2*E*).
246 Our attempt to identify core CCs of the “symbiosis” trait yielded 2 “symbiosis” trait-
247 CCs involving a maximum of 8 distinct proteomes of symbiotic species and 187
248 “symbiosis” trait-CCs involving 7 proteomes (of the 12 proteomes available for symbiotic
249 species) (Fig. 2*F*). GO annotations of these 189 core “symbiosis” trait-CCs revealed that
250 the major part of the protein-coding domains (*i.e.* 1400 out of 1896) could not be

functionally annotated. Among those that could be annotated, 73.8% of the protein-coding domains corresponded to “membrane proteins” (cellular component). The remainder corresponded to “proteins of photosystem I”, “extracellular region” and “spliceosomal complex”. With respect to biological process, 31.9% were involved in ion transport while 23.8% were involved in proteolytic processes (Tab. S16). Looking at taxonomic distribution, we found that the 189 CCs were dominated by protein-coding domains belonging to *Symbiodinium* spp. (1679 protein-coding domains from the 187 CCs and 13 protein-coding domains from the 2 CCs). *Pelagodinium beii* was represented by 211 protein-coding domains spread across the 187 CCs. Finally, few protein-coding domains (respectively 8 and 1) of *Gymnodinium radiolariae* and *Brandtodinium nutricula* were found within the 189 “symbiosis” trait-CCs.

The 52,491 “symbiosis” trait-CCs (composed of 130,673 protein-coding domains that lack InterPro functional annotation) were compared to nr database alignments (with an e-value of 1e-3) revealing 495 protein-coding domains with sequence identity higher than 80%. These 495 protein-coding domains belonged to 298 CCs, which were composed of 919 protein-coding domains in total. Finally, 52,193 “symbiosis” trait-CCs were completely unannotated.

DISCUSSION

Pipeline efficiency & sequence similarity network

Our *de novo* assembly and downstream pipeline analysis of multiple dinoflagellate transcriptomes overcame several biases inherent to *de novo* assembly processes (Fig. S5). For instance, the protein-coding domain prediction step we performed contributed to selection of transcripts in which ORFs and protein domains were detected but also allowed removal of truncated or chimeric transcripts (Yang & Smith 2013). Protein-coding domains derived from high quality transcriptomes enabled construction of sequence similarity

networks to focus on shared protein-coding domains among multiple proteomes. Considering our 46 proteomes, a mean value of 60,661 protein-coding domains was found, which is consistent with the previously estimated range of 34,156 to 75,461 protein-coding genes in dinoflagellates (Murray *et al.* 2016). The median length of the protein-coding domains was 307 bp, also consistent with the median protein length of 361 bp reported based on genomes of 5 model eukaryote species (*Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana*) (Brocchieri & Karlin 2005).

Sequence similarity networks represent an informative and pragmatic way to study massive datasets (Atkinson *et al.* 2009; Alvarez-Ponce *et al.* 2013; Cheng *et al.* 2014; Forster *et al.* 2015; Méheust *et al.* 2016). In (Cheng *et al.* 2014), 84 genome-derived proteomes of prokaryotes (*i.e.* 128,628 sequences) were used to study the impact of redox state changes on their gene content and evolution. The authors found that the core CCs revealed a correlation between their network structure and differences in respiratory phenotypes. Our SSN has allowed simultaneous exploration of 46 transcriptome-derived proteomes (1,275,911 sequences), including their overwhelming “dark matter” (*i.e.* here protein-coding domains totally lacking functional annotation). High identity and coverage threshold values used in our analyses to filter alignments ensured that only high quality alignments were included in the network (Bittner *et al.* 2010). The integration of 4 new dinoflagellate proteomes represented an increase of 14% of protein-coding domains in the SSN and overall the dataset represents the most comprehensive picture to date of the genomic potential of dinoflagellates. This new resource and comparative genomic

298 approach allow generation and testing of original hypotheses about the genomic basis for
299 evolutionary history and life style, functional traits, and specificities of dinoflagellates.

300 **Large-scale comparison of dinoflagellate proteomes**

301 The SSN analyses allowed characterization of the core and accessory proteomes
302 for this large dataset of non-model organisms. Because our analysis relied on a *de novo*
303 assembled, transcriptome-derived, proteome SSN rather than classical knowledge-based
304 genomics, it also promoted discovery of new CCs, each of which can be functionally
305 assimilated to a single putative conserved protein-domain (pCD) in such non-model
306 organisms (Lopez *et al.* 2015) (Fig. S6).

307 The core dinoflagellate proteome identified in our analysis was composed of 252
308 pCDs (Fig. 1*A*), a size that falls in the range of the latest estimates for bacteria (352 core
309 genes) (Yang *et al.* 2015) and eukaryotes (258 core genes in CEGMA, and more recently
310 429 single-copy orthologs in BUSCO) (Parra *et al.* 2007; Simão *et al.* 2015). The
311 extrapolation of the number of core CCs does not saturate, suggesting that the number of
312 core CCs for dinoflagellates could be less than 256. It also suggests that dinoflagellates
313 expressed only a fraction of core eukaryote genes referenced in databases. Our
314 comparative analysis with the most up-to-date eukaryotic orthologous gene database
315 BUSCO strongly stresses the need to generate more gene and protein data for non-model
316 marine organisms in order to populate reference databases (Armengaud *et al.* 2014). The
317 small overlap between core dinoflagellate pCDs identified here and the BUSCO database
318 suggests that essential functions expressed by dinoflagellates are distantly related to
319 those of current model eukaryotes.

320 Our SSN constitutes a strong basis for exploration and refinement of functional
321 annotations as our dataset encompassed a broad range of dinoflagellate taxa according
322 to recent phylogenetic analyses (Bachvaroff *et al.* 2014; Janouškovec *et al.* 2016).
323 However, the identified core proteome can only be considered partial as our dataset *i*- did

not include representatives of all described dinoflagellate lineages, and *ii*- relied on transcriptomic (*i.e.* gene expression) data that can vary according to eco-physiological conditions and/or life-cycle stage. The content of our SSN can be updated permanently to refine these estimates as new dinoflagellate genomic data are accumulated (Shoguchi *et al.* 2013; Lin *et al.* 2015; Aranda *et al.* 2016). 236 (93%) core CCs involving one or more functionally annotated protein-coding domains (Fig. 2*B*) can be exploited to extend annotation to other aligned protein-coding domains within each CC, therefore leading to a more comprehensive pCD description. For instance, looking for the HSP70 conserved protein domain, which is ubiquitous in all eukaryotic organisms (Germot & Philippe 1999), we found 320 protein-coding domain sequences annotated as HSP70, all belonging to a single CC composed of 328 protein-coding domain sequences. The 8 remaining protein-coding domain sequences were either imprecisely annotated as chaperone DnaK (1 sequence), cyclic nucleotide-binding domain (2 sequences), heat shock protein 70 family (3 sequences) or annotation was simply missing (2 sequences) (Tab. S17). As this CC was 97% represented by HSP70 annotation, it is reasonable to extend this annotation to all protein-coding domain sequences forming the connected component. Considering only CCs that were at least half composed of annotated protein-coding domain sequences, this approach could be applied to complement the functional characterization of 49 CCs (583 unannotated protein-coding domains) through extension of functional annotations.

Relatively few previous studies have relied on functional aspects such as protein alignments to investigate dinoflagellate phylogeny. A recent study used for the first time a multi-protein dataset providing a robust phylogeny for dinoflagellates (Janouškovec *et al.* 2016). The comparison of the 101 orthologous alignments used in (Janouškovec *et al.* 2016) with our 252 pCDs revealed 206 additional core pCDs that are good candidates for

refining dinoflagellate phylogeny, increasing by nearly 200% the quantity of information available for such studies.

Among the 176,958 distinct CCs entirely composed of unannotated protein-coding domains in the total network, 16 CCs or pCDs (composed of 946 protein-coding domain sequences) belonged to our core dinoflagellate proteome (Fig. 1*B*). This highlights the fact that many fundamental genomic features remain to be characterized in this group. These unknown groups of homologous domains are excellent potential candidate markers to further investigate dinoflagellate genomics at a broad scale and might also be useful for identification of dinoflagellates within complex environmental genomic datasets.

Investigating harmful dinoflagellates

Toxic dinoflagellates represent about 80% of toxic eukaryotic phytoplankton species (Janouškovec *et al.* 2016). Production of toxins by dinoflagellates is well known and can cause major health and economic problems. *Karenia brevis*, for example, is known to produce brevetoxins which cause fish mortality and can affect human health through the consumption of contaminated seafood or direct exposure to harmful algal blooms (HABs) (Flewelling *et al.* 2005). To date, several dinoflagellate toxins have been chemically and genetically characterized (Wang 2008; Kellmann *et al.* 2010; Stüken *et al.* 2011; Cusick & Sayler 2013). In our SSN analyses, protein-coding domains homologous to PKS were identified in CCs composed of domains from both “harmful for human” and non-“harmful for human” species. This result validates a previous report that PKS proteins are not exclusive to toxic species (Kohli *et al.* 2016). PKS are in fact involved in the production of a variety of natural products such as small acids, acetyl-CoA or propionyl-Co (Khosla *et al.* 2014). Spreading information among unannotated protein-coding domains in both “harmful for human” and non-“harmful for human” trait-CCs in which PKS were identified allowed extension of the potential PKS-like annotation to 9 protein-coding domains from “harmful for human” trait-CCs and 498 protein-coding domains from non-

374 “harmful for human” trait-CCs. PKS protein-coding domains for 4 extra species
375 (*Alexandrium catenella*, *Kryptoperidinium foliaceum*, *Protoceratium reticulatum* and
376 *Crypthecodinium cohnii*) were also detected compared to the database from (Kohli *et al.*
377 2016) (Tab. S13) confirming that synthesis of PKS proteins is not exclusive to toxic species
378 (Kohli *et al.* 2016).

379 With respect to saxitoxin production, we did not detect either *sxtA* or *sxtG* related
380 protein-coding domains in “harmful for human” trait-CCs which suggests that such proteins
381 are not exclusively expressed by toxic species. Robust alignments for both *sxtA* and *sxtG*
382 protein-coding domains in non-“harmful for human” trait-CCs were found. Our results
383 differed somewhat from those of a previous study (Murray *et al.* 2015) based on the same
384 initial dataset from the MMETSP. Specifically, we were not able to detect *sxtA* in 7 species
385 in which this gene was detected in (Murray *et al.* 2015) and in contrast *sxtA* domains were
386 detected in 9 species in which the protein was not detected previously (Murray *et al.* 2015)
387 (Tab. S15). These differences may be due to the use of distinct *de novo* assembly tools
388 and pCD prediction processes, illustrating the requirement to ultimately combine *in vitro*
389 and *in silico* methods in order to unambiguously characterize toxic species. Nevertheless,
390 our results seem consistent with biological knowledge since the expression of both *sxtA*
391 and *sxtG* proteins, which are involved in toxin biosynthesis process in toxic species
392 (Hackett *et al.* 2013), has already been revealed for *P. bahamense* and *G. catenatum* that
393 are known to be STX-producing dinoflagellates and reported as toxic species (Tab. S14).
394 We also confidently detected 2 protein-coding domains including *sxtA* domains and 1
395 protein-coding domain including *sxtG* domains in *P. beii*, an *a priori* non-toxic symbiotic
396 species that has never been reported as a STX-producer. This is consistent with the fact
397 that *sxtG* has previously been identified in non-toxic species (Orr *et al.* 2013). The detailed
398 investigation of the 2 first domains sharing similarity to *sxtA* showed that one shares an
399 aminotransferase domain with *sxtA* (Hackett *et al.* 2013) and the second shares GNAT

400 domains with *sxtA*. These results alone are not sufficient to prove toxin production and
401 toxicity tests must be performed *in vitro* to confirm the synthesis of saxitoxins by *P. beii*.
402 Previous studies showed that two forms of *sxtA* (long and short) are present in some
403 *Alexandrium* genera (Stüken *et al.* 2011) and that the shorter form is not related to toxin
404 production (Murray *et al.* 2015). We were not able to associate the domains identified in
405 our study to either the long or short form of *sxtA*, and we cannot exclude the possibility
406 that the *sxtA* domain identified could belong to a molecule that is synthesized through the
407 saxitoxin biosynthesis pathway but that is not a functional saxitoxin. From an evolutionary
408 point of view, as PKS and STX genes are also found in species currently described as
409 non-toxic, it seems that like for snake venoms, dinoflagellate toxins evolved by recruitment
410 of genes encoding regular proteins followed by gene duplication and neo-functionalization
411 of the domains (Vonk *et al.* 2013).

412 Based on exploration of the SSN for putative coding domains specific for “harmful
413 for human” trait-CCs, we found that membrane located protein domains and more
414 specifically ion transport protein domains were important components characterizing toxic
415 species. This is in agreement with reports that ion channel proteins and proteins involved
416 in neurotransmission are mediators of dinoflagellate toxicity (Wang 2008; Cusick & Sayler
417 2013). For our SSN analysis, we made the assumption that the more species are
418 represented in a CC, the more the corresponding gene set is likely to be specific for this
419 particular functional trait. In the case of “harmful for human” trait-CCs we noticed that 2 of
420 the 5 CCs with the most toxic representatives (*i.e.* 7 species) were exclusively composed
421 of unannotated domains, representing essential functions constitutively expressed by toxic

422 species only and for which further investigations are required to better characterize toxic
423 dinoflagellates.

424 **Focus on symbiosis**

425 The “symbiotic” gene set compiled from the literature based on their involvement
426 in the establishment and maintenance of symbiosis (Lehnert *et al.* 2014; Lin *et al.* 2015)
427 was found here in both “symbiosis” trait-CCs and in non-“symbiosis” trait-CCs (Tab. S15),
428 suggesting that these proteins are constitutively expressed by all dinoflagellate species.
429 This result may reflect the fact that the transcriptomes of dinoflagellate strains were not
430 directly isolated from symbiotic conditions, but rather from their free-living stages
431 maintained in culture. Symbiotic genes identified from the literature were originally inferred
432 from studies on holobionts (*i.e.* host and symbionts), but proved here not to be exclusive
433 to symbiotic dinoflagellates when performing global comparison of multiple datasets.

434 Functional annotations of “symbiosis” trait-CCs revealed an overall clear
435 domination of proteins involved in phosphorylation and ion transport domains (*e.g.*
436 sodium, potassium and calcium ion channel proteins) located within membrane
437 compartments (Fig. 2*D*). The 4 most prominent functions that occured 2 to 10 times more
438 often in “symbiosis” trait-CCs than in non-“symbiosis” trait-CCs (Fig. 2*E*) were related to
439 ion transport domains and regulation processes. The results indicate that the two functions
440 are constitutively more expressed in symbiotic compared to non-symbiotic species.
441 Protein phosphorylation is known to take part in cellular mechanisms in response to the
442 environment (Day *et al.* 2016) and play a key role in signal transduction to other cells in
443 plant parasitism and symbiosis models (Lionetti & Metraux 2015). The specific dominant
444 presence of ion transport domains (also involved in cell signaling and cell adaptation to
445 the environment) in symbiotic dinoflagellates could represent a constitutive characteristic
446 of symbiotic species facilitating establishment and maintenance of the symbiosis. Notably,
447 the role of ion channel proteins has been highlighted as essential in plant root

endosymbiosis (Charpentier *et al.* 2008; Matzke *et al.* 2009). This suggests that symbiotic species are likely to be constitutively better equipped for environmental adaptations.

Exploring the SSN, we found that 45% of the overall protein-coding domains associated to symbiotic species were functionally annotated (Tab. S19) which implies that a large suite of uncharacterized functions are specifically associated to symbiotic dinoflagellates. 129,754 protein coding domains from 52,193 “symbiosis” trait-CCs remained unannotated according to the InterPro and nr databases. We found 187 “symbiosis” trait-CCs composed of 7 of 12 possible symbiotic species and 2 CCs composed of 8 of the 12 (Fig. 2*F*). Protein-coding domains from 2 of the 3 newly added symbiotic proteomes were found in both the 187 and 2 pCDs, contributing to revealing pCDs specific of symbiotic species. The 2 “symbiosis” trait-CCs encompassing the 8 distinct species were exclusively composed of unannotated domains, suggesting that these 2 CCs represent pCDs with fundamental, yet unknown, functions constitutively expressed by symbiotic species. Overall, our analyses demonstrate that SSN has significant potential to reveal the variety of annotated and unknown pCDs that constitute good candidates for further study to characterize and understand the genomic basis of symbioses involving dinoflagellates.

M&M

Dataset building

The dataset used in our study included all dinoflagellate transcriptomes available in the MMETSP project repository (http://marinemicroeukaryotes.org/resources) as well as 4 transcriptomes generated for this study (more details in the following section) (Fig. S7). This dataset represented transcriptomes of 47 distinct species from 35 genera, 19 families, and 11 of the 21 current dinoflagellate taxonomic orders according to the taxonomic framework of the WoRMS database (http://www.marinespecies.org/index.php)

(Tab. 1). Taxonomy and functional trait information (i.e. chloroplast occurrence and origin, trophic mode, harmfulness for humans, ability to live in symbiosis, to perform kleptoplasty, to be a parasite or to be toxic for fauna) were indicated for each organism considered (Tab. 1).

Culturing and RNA sequencing for four dinoflagellate strains

Free-living clonal strains of the dinoflagellate species *Brandtodinium nutricula* (RCC3468) (Probert *et al.* 2014) and *Gymnoxanthella radiolariae* (RCC3507) (Yuasa *et al.* 2016) isolated from symbiotic Radiolaria, *Pelagodinium beii* (RCC1491) (Siano *et al.* 2010) isolated from a foraminiferan host, and the non-symbiotic *Heterocapsa* sp. (RCC1516) were obtained from the Roscoff Culture Collection (www.roscoff-culture-collection.org). Triplicate 2-L acid-washed, autoclaved polycarbonate Nalgene bottles were filled with 0.2 micron filter-sterilized (Stericup-GP, Millipore) seawater with K/2 (-Tris,-Si) medium supplements (Keller *et al.* 1987) and inoculated with an exponentially growing culture of each strain. All cultures were maintained at 18°C, ~80 µmol photon m^{-2} s^{-1} light intensity and 14:10 light:dark cycle. Cell abundance was monitored daily by flow cytometry with a FACSAria flow cytometer (Becton Dickinson, San José, CA, USA) and derived cell division rates were used to monitor the growth phase of the culture. Light and dark phase samples for transcriptome analyses were taken from exponential and stationary phase cultures. 100 mL aliquots from each culture were filtered onto 3 micron pore-size polycarbonate filters with an autoclaved 47 mm glass vacuum filter system (Millipore) and a hand-operated PVC vacuum pump with gauge to maintain the vacuum pressure below 5 mm Hg during filtration. The filter was then placed in a sterile 15 mL falcon tube filled with ca. 5 ml TriZol and stored at -80°C.

Total RNA was purified directly from the filters stored in TriZol using the Direct-zol RNA Miniprep kit (ZymoResearch, Irvine, CA). First, the tube containing the filter immersed in TriZol was incubated for 10 min at 65°C. Then, after addition of an equal

volume of 100% EtOH and vortexing, the mixture was loaded into a Zymo-SpinIIC column and centrifuged for 1 min at 12,000 *g*. The loading and centrifugation steps were repeated until exhaustion of the mixture. RNA purification was completed by prewash and wash steps following the manufacturer's instructions and RNA was directly eluted in 45 µL nuclease-free water. The in-column DNAse step was replaced by a more efficient post-extraction DNAse treatment using the Turbo DNA-free kit (Thermo Fisher Scientific, Waltham, MA) according to the manufacturer's rigorous DNase treatment procedure. After two rounds of 30 minutes incubation at 37°C, the reaction mixture was purified with the RNA Clean and Concentrator-5 kit (ZymoResearch) following the procedure described for retention of >17nt RNA fragments. Total RNA, eluted in 20 µl nuclease-free water, was quantified with RNA-specific fluorimetric quantification on a Qubit 2.0 Fluorometer using Qubit RNA HS Assay (ThermoFisher Scientific). RNA quality was assessed by capillary electrophoresis on an Agilent Bioanalyzer using the RNA 6000 Pico LabChip kit (Agilent Technologies, Santa Clara, CA).

RNA-Seq library preparations were carried out from 1 µg total RNA using the TruSeq Stranded mRNA kit (Illumina, San Diego, CA), which allows mRNA strand orientation. Briefly, poly(A)$^+$ RNA was selected with oligo(dT) beads, chemically fragmented and converted into single-stranded cDNA using random hexamer priming. Then, the second strand was generated to create double-stranded cDNA. Strand specificity was achieved by quenching the second strand during final amplification thanks to incorporation of dUTP instead of dTTP during second strand synthesis. Then, ready-to-sequence Illumina libraries were quantified by qPCR using the KAPA Library Quantification Kit for Illumina libraries (KapaBiosystems, Wilmington, MA), and library profiles evaluated with an Agilent 2100 Bioanalyzer (Agilent Technologies). Each library

was sequenced using 101 bp paired-end read chemistry on a HiSeq2000 Illumina sequencer.

Data filtering and *de novo* assembly

For each strain considered (Tab. S7 and our 4 strains), sequenced reads were pooled resulting in 60 datasets, then filtered using Trimmomatic (Tab. 1) (34). Reads with quality below 30 Q on a sliding window size of 10 were excluded. Remaining reads were assembled with the *de novo* assembler Trinity version 2.1.1 (Grabherr *et al.* 2011) using default parameters for the paired reads method. Of the initial 60 transcriptome datasets (56 from the MMETSP repository and 4 produced in this study), 57 were successfully assembled. The assembly process could not be completed properly for 3 datasets due to incompatibility between the version of the assembly software and the datasets (*Karenia brevis* strain CCMP 2229, Wilson SP1 and SP3 as a combined assembly, *Oxyrrhis marina* strain CCMP1795 and *Symbiodinium kawaguti* strain CCMP2468) (Tab. 1).

Assembled transcripts were evaluated based on: (i) sequence metrics, and (ii) read remapping rates calculated respectively with homemade scripts and Bowtie 2 in local mode (Langmead *et al.* 2009) (Tab. 1). Based on these assembly analyses, two classes of assembly quality were defined: those with >30,000 transcripts with a N50 > 400 bp and read remapping rate >50% were tagged as “high quality” transcriptomes whereas the rest were tagged as “low quality” transcriptomes. An exception was made for one poor quality transcriptome corresponding to the species *Oxyrrhis marina* (LB1974 and NA strain) composed of 18,275 assembled transcripts that was intentionally tagged as a “high quality” transcriptome because this basal species holds a key evolutionary and ecological

545 position among dinoflagellates (Montagnes *et al.* 2011; Lee *et al.* 2014; Bachvaroff *et al.*
546 2014).

547 **Coding domain prediction and functional annotation**

548 For each transcriptome, coding domain prediction of assembled transcripts was
549 conducted with Transdecoder version 2.0.1 (Haas *et al.* 2013) to obtain peptide
550 sequences of corresponding domains. We defined each set of predicted protein domains
551 as a proteome. The optional step of Transdecoder consisting in the identification of ORFs
552 in the protein domain database Pfam was not executed in order to avoid a comparative
553 approach that would result in a limited discovery of new sequences. The predicted coding
554 domains were then processed with the Interproscan 5 functional annotation program
555 version 5.11-51.0 (Jones *et al.* 2014) to scan for protein signatures. Default parameters
556 were used to obtain each proteome. Finally, to get a broad overview of the ontology
557 content of our datasets, GO slims were retrieved from the Gene Ontology Consortium to
558 build a summary of the GO annotations without the detail of the specific fine-grained terms
559 (http://geneontology.org/page/go-slim-and-subset-guide).

560 **Sequence similarity network**

561 A sequence similarity network (SSN) is a graph in which vertices are genomic
562 sequences and the edges represent similarity between sequences. A SSN is composed
563 of connected components (CC) (subgraphs or subnetworks, including at least two vertices
564 disconnected from other subgraphs in the total network). As information can be linked to
565 sequences (e.g. in our study: taxonomy, functional annotation, functional traits), the SSN
566 and its structure can be explored accordingly. Using predicted protein domain sequences,
567 a SSN was constructed with the BLASTp alignment method (Altschul *et al.* 1990) with an

568 e-value of 1e-25 using the DIAMOND software (Buchfink *et al.* 2015). Similarities
569 satisfying query and subject sequence coverages higher than 80% were kept.
570 Whenever predicted coding domains aligned together forming a CC it can be
571 assumed that they potentially share a similar molecular function (Marchler-Bauer *et al.*
572 2005) and form putative conserved domains (pCDs). SSN exploration and analyses were
573 performed using personal scripts and functions implemented in the igraph R package
574 (Csárdi & Nepusz 2006). Diverse biological information related to the species considered
575 were mapped on each vertex, and missing information were marked as <NA>.
576 In our approach, CC number, structure and composition were impacted when edge
577 sequence identity cut off was shifted. We thus tested different similarity thresholds and
578 chose an optimal threshold according to the two following criteria: maximizing the number
579 of large CCs (i.e. minimum of 30 vertices) and the number of CCs involving a single
580 homogeneous functional annotation (i.e. a unique GOslim term at the Biological Process
581 level). An optimal sequence identity threshold at 60% similarity with our dataset was
582 inferred (Fig. S1). As a last filtering step, we chose to consider only vertices of proteomes
583 derived from “high quality” tagged transcriptomes.
584 43 proteomes composed of comparable numbers of protein domains (*i.e.* a
585 minimum of 9,000 domains) (Fig. S8) were used to define the core-, accessory- and pan-
586 proteomes. The core-proteome corresponds to the CCs composed of sequences from
587 every single proteome considered, whereas the accessory-proteome corresponds to the
588 CCs composed of sequences from a single proteome. The pan-proteome corresponds to
589 the total number of CCs identified in the network. In addition to the Interproscan annotation
590 process, sequences belonging to core CCs were compared to 3 databases: (i) BUSCO

591 core eukaryotic gene set (Simão *et al.* 2015), (ii) the UniProtKB/Swiss-Prot database, and
592 (iii) the nr database, using BLASTp and an e-value of 1e-25.

593 To further explore the composition and structure of the CCs, we computed the
594 Pielou equitability index (Mulder *et al.* 2004), classically used in ecology in order to
595 estimate the richness and/or evenness of species in a sample. Here the Pielou index was
596 used to estimate the contribution of each proteome in any given CC, for instance for
597 assessing whether a CC is mainly composed of domains from a limited number of
598 proteomes. The index ranges from 0 to 1, the more homogeneous the composition of a
599 CC, the higher the index.

600 **Investigation of the functional annotation**

601 Analyses of functional traits were based on the SSN encompassing the 46
602 proteomes derived from “high quality” transcriptomes. The information about 10 selected
603 functional traits was retrieved from the literature (Tab. 1). The details about plastid origin
604 and presence were retrieved from (Caruana & Malin 2014). Dinoflagellates that are
605 capable of mixotrophy were listed in (Jeong *et al.* 2010). The information on species
606 harmful to humans (AZP, DSP, NSP, PSP, CFP syndromes) or to marine fauna
607 (ichyotoxicity) was obtained from the Taxonomic Reference List of Harmful MicroAlgae of
608 the IOC-UNESCO (http://www.marinespecies.org/hab/index.php). Dinoflagellate plastidy
609 is reviewed in (Gagat *et al.* 2014). Dinoflagellates which have the capacity to produce
610 DMSP in high cellular concentration were described in (Caruana *et al.* 2012). Presence of
611 the theca, characteristic of thecate dinoflagellates, has been studied in (Lin 2011; Orr *et*
612 *al.* 2012). In (Rengefors *et al.* 1998) authors studied dinoflagellates species that go
613 through a cyst stage during their life cycle. Symbiotic taxa are characterized in (Trench &
614 Blank 1987; Siano *et al.* 2010; Decelle *et al.* 2012; Probert *et al.* 2014; Yuasa *et al.* 2016).

We later focused on CCs that are specific to a given trait, called “trait-CCs”, defined by CCs exclusively composed of vertices tagged with this trait (and excluding <NA> tags).

Following an exploratory approach, among trait-CCs, CCs including a maximum of distinct proteomes were sought (except for the “parasite” trait, as only one of the two proteomes are represented in the network). In this study, we examined more specifically the functional composition for the “harmful for human” and “symbiosis” trait-CCs. To validate the SSN capacity to detect trait-CCs characteristic for a given function, we followed a knowledge-based approach searching for sequence similarities through BLASTp (e-value 1e-3) to well-known genes from the literature.

Research of toxin sequences in “harmful for human” trait-CCs

Specific studies on toxic dinoflagellate species have led to the establishment of defined gene sets related to toxin production that can be used as a reference for knowledge-based approaches (Snyder *et al.*; Monroe & Van Dolah 2008; Wang 2008; Sheng *et al.* 2010; Kellmann *et al.* 2010; Stüken *et al.* 2011; Salcedo *et al.* 2012; Hackett *et al.* 2013; Cusick & Sayler 2013; Lehnert *et al.* 2014; Perini *et al.* 2014; Zhang *et al.* 2014; Kohli *et al.* 2015, 2016; Meyer *et al.* 2015; Murray *et al.* 2015; Beedessee *et al.* 2015). Many of the toxic metabolites produced by some dinoflagellate species are of polyketide origin (Kellmann *et al.* 2010). 2,632 polyketide synthase (PKS) peptide sequences from (Kohli *et al.* 2016) (supplementary data 3) were compared to sequences from “harmful for human” trait-CCs to unveil PKS presence as well as non-“harmful for human” trait-CCs as a control (retained alignments show 80% sequence identity and 80% sequence coverage). Previous studies have also identified *sxt* genes involved in saxitoxin (STX) biosynthesis (Orr *et al.* 2013) that we compared to “harmful for human” and non-“harmful for human” trait-CCs using a specific threshold (retained alignments with 80% sequence identity and 90% sequence coverage). Of these, two genes have been highlighted to be related with the STX biosynthesis pathway: *sxtA* and *sxtG*. We based

641 our investigations on 117 *sxtA1-4* and 20 *sxtG* sequences from (Murray *et al.* 2015) (Tab.
642 S20). The differential composition of functional annotations between “harmful for human”
643 and non-“harmful for human” trait-CCs was investigated to detect functions that are likely
644 more represented in toxic species. The counts of each annotation found in each functional
645 category were respectively normalized by the total number of sequences that composed
646 both trait-CCs. Finally, the difference of pair normalized counts for the same annotation in
647 “harmful for human” and non-“ harmful for human” trait-CCs was calculated

648 **Probing “symbiosis” trait-CCs**

649 In this study, three additional transcriptomes of symbiotic species were added to
650 the MMETSP data to increase the number of transcriptomes of symbiotic species from 10
651 to 13. Following a similar strategy as for the “harmful for human” CC-trait, investigation of
652 the “symbiosis” trait in our network was based on reported sets of genes potentially
653 involved in the symbiotic lifestyle for *Symbiodinium kawaguti* (Lin *et al.* 2015) and coral
654 symbiotic relationships (Tab. S21). We combined this set with other putative proteins
655 highly up-regulated in anemone-dinoflagellate symbiosis (Lehnert *et al.* 2014). The
656 distribution of 150 “symbiotic” marker sequences was studied across “symbiosis” trait-CCs
657 (Tab. S15). The differential composition of functional annotations between “symbiosis” and
658 non-“symbiosis” trait-CCs was investigated as previously described for “harmful for
659 human” trait-CCs.

660

661 # DATA ACCESSIBILITY

662 Link to data: http://application.sb-roscoff.fr/project/radiolaria/

663 # Acknowledgments

664 We thank Gaëlle Lelandais and Éric Pelletier for their support and critical discussions. We
665 are also grateful to RCC staff for providing dinoflagellate cultures as well as ABIMs staff

666 for the help on computational facilities. This work was supported by a 3-year Ph.D. grant
667 from “Interface Pour le Vivant” (IPV) program at the Université Pierre et Marie Curie
668 (UPMC), Paris. This project was supported by grants from Région Ile-de-France.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO (2013) Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, E1594-1603.

Aranda M, Li Y, Liew YJ *et al.* (2016) Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle. *Scientific Reports*, **6**.

Armengaud J, Trapp J, Pible O *et al.* (2014) Non-model organisms, a species endangered by proteogenomics. *Journal of Proteomics*, **105**, 5–18.

Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC (2009) Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. *PLoS ONE*, **4**.

Bachvaroff TR, Gornik SG, Concepcion GT *et al.* (2014) Dinoflagellate phylogeny revisited: Using ribosomal proteins to resolve deep branching dinoflagellate clades. *Molecular Phylogenetics and Evolution*, **70**, 314–322.

Beedessee G, Hisata K, Roy MC, Satoh N, Shoguchi E (2015) Multifunctional polyketide synthase genes identified by genomic survey of the symbiotic dinoflagellate, Symbiodinium minutum. *BMC Genomics*, **16**.

Bittner L, Halary S, Payri C *et al.* (2010) Some considerations for analyzing biodiversity using integrative metagenomics and gene networks. *Biology Direct*, **5**, 47.

Brocchieri L, Karlin S (2005) Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research*, **33**, 3390–3400.

Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, **12**, 59–60.

Caron DA, Alexander H, Allen AE *et al.* (2016) Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nature Reviews Microbiology*, **advance online publication**.

Caruana AMN, Malin G (2014) The variability in DMSP content and DMSP lyase activity in marine dinoflagellates. *Progress in Oceanography*, **120**, 410–424.

Caruana AMN, Steinke M, Turner SM, Malin G (2012) Concentrations of dimethylsulphoniopropionate and activities of dimethylsulphide-producing enzymes in batch cultures of nine dinoflagellate species. *Biogeochemistry*, **110**, 87–107.

Charpentier M, Bredemeier R, Wanner G *et al.* (2008) Lotus japonicus CASTOR and POLLUX Are Ion Channels Essential for Perinuclear Calcium Spiking in Legume Root Endosymbiosis. *The Plant Cell*, **20**, 3467–3479.

Cheng S, Karkar S, Bapteste E *et al.* (2014) Sequence similarity network reveals the imprints of major diversification events in the evolution of microbial life. *Frontiers in Ecology and Evolution*, **2**.

Csárdi G, Nepusz T (2006) The igraph software package for complex network research. *InterJournal Complex Systems*.

Cusick KD, Sayler GS (2013) An Overview on the Marine Neurotoxin, Saxitoxin: Genetics, Molecular Targets, Methods of Detection and Ecological Functions. *Marine Drugs*, **11**, 991–1018.

Day EK, Sosale NG, Lazzara MJ (2016) Cell signaling regulation by protein phosphorylation: a multivariate, heterogeneous, and context-dependent process. *Current Opinion in Biotechnology*, **40**, 185–192.

Decelle J, Colin S, Foster RA (2015) Photosymbiosis in Marine Planktonic Protists. In: *Marine Protists* (eds Ohtsuka S, Suzaki T, Horiguchi T, Suzuki N, Not F), pp. 465–500. Springer Japan.

Decelle J, Probert I, Bittner L *et al.* (2012) An original mode of symbiosis in open ocean plankton. *Proceedings of the National Academy of Sciences*, **109**, 18000–18005.

Dupont CL, McCrow JP, Valas R *et al.* (2015) Genomes and gene expression across light and productivity gradients in eastern subtropical Pacific microbial communities. *The ISME Journal*, **9**, 1076–1092.

Flewelling LJ, Naar JP, Abbott JP *et al.* (2005) Brevetoxicosis: Red tides and marine mammal mortalities. *Nature*, **435**, 755–756.

Forster D, Bittner L, Karkar S *et al.* (2015) Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. *BMC Biology*, **13**, 16.

Gagat P, Bodył A, Mackiewicz P, Stiller JW (2014) Tertiary Plastid Endosymbioses in Dinoflagellates. In: *Endosymbiosis* (ed Löffelhardt W), pp. 233–290. Springer Vienna.

Gast RJ, Moran DM, Dennett MR, Caron DA (2007) Kleptoplasty in an Antarctic dinoflagellate: caught in evolutionary transition? *Environmental Microbiology*, **9**, 39–45.

Gerlt JA, Babbitt PC, Jacobson MP, Almo SC (2012) Divergent Evolution in Enolase Superfamily: Strategies for Assigning Functions. *The Journal of Biological Chemistry*, **287**, 29–34.

Germot A, Philippe H (1999) Critical Analysis of Eukaryotic Phylogeny: A Case Study Based on the HSP70 Family. *Journal of Eukaryotic Microbiology*, **46**, 116–124.

Goodson MS, Whitehead LF, Douglas AE (2001) Symbiotic dinoflagellates in marine Cnidaria: diversity and function. *Hydrobiologia*, **461**, 79–82.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome (Trinity). *Nature Biotechnology*, **29**, 644–652.

Haas BJ, Papanicolaou A, Yassour M *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, **8**, 1494–1512.

Hackett JD, Wisecaver JH, Brosnahan ML *et al.* (2013) Evolution of Saxitoxin Synthesis in Cyanobacteria and Dinoflagellates. *Molecular Biology and Evolution*, **30**, 70–78.

Jaeckisch N, Yang I, Wohlrab S *et al.* (2011) Comparative Genomic and Transcriptomic Characterization of the Toxigenic Marine Dinoflagellate Alexandrium ostenfeldii. *PLOS ONE*, **6**, e28012.

Janouškovec J, Gavelis GS, Burki F *et al.* (2016) Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics. *Proceedings of the National Academy of Sciences*, 201614842.

Jeong HJ, Yoo YD, Kim JS *et al.* (2010) Growth, feeding and ecological roles of the mixotrophic and heterotrophic dinoflagellates in marine planktonic food webs. *Ocean Science Journal*, **45**, 65–91.

Jones P, Binns D, Chang H-Y *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)*, **30**, 1236–1240.

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

Keller MB, Lavori PW, Friedman B *et al.* (1987) The Longitudinal Interval Follow-up Evaluation. A comprehensive method for assessing outcome in prospective longitudinal studies. *Archives of General Psychiatry*, **44**, 540–548.

Kellmann R, Stüken A, Orr RJS, Svendsen HM, Jakobsen KS (2010) Biosynthesis and Molecular Genetics of Polyketides in Marine Dinoflagellates. *Marine Drugs*, **8**, 1011–1048.

Khosla C, Herschlag D, Cane DE, Walsh CT (2014) Assembly Line Polyketide Synthases: Mechanistic Insights and Unsolved Problems. *Biochemistry*, **53**, 2875–2883.

Kohli GS, John U, Figueroa RI *et al.* (2015) Polyketide synthesis genes associated with toxin production in two species of Gambierdiscus (Dinophyceae). *BMC Genomics*, **16**, 410.

Kohli GS, John U, Van Dolah FM, Murray SA (2016) Evolutionary distinctiveness of fatty acid and polyketide synthesis in eukaryotes. *The ISME Journal*.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Le Bescot N, Mahé F, Audic S *et al.* (2016) Global patterns of pelagic dinoflagellate diversity across protist size classes unveiled by metabarcoding. *Environmental Microbiology*, **18**, 609–626.

Lee R, Lai H, Malik SB *et al.* (2014) Analysis of EST data of the marine protist Oxyrrhis marina, an emerging model for alveolate biology and evolution. *BMC Genomics*, **15**, 122.

Lehnert EM, Mouchka ME, Burriesci MS *et al.* (2014) Extensive Differences in Gene Expression Between Symbiotic and Aposymbiotic Cnidarians. *G3: Genes|Genomes|Genetics*, **4**, 277–295.

Lima-Mendez G, Faust K, Henry N *et al.* (2015) Determinants of community structure in the global plankton interactome. *Science*, **348**, 1262073.

Lin S (2011) Genomic understanding of dinoflagellates. *Research in Microbiology*, **162**, 551–569.

Lin S, Cheng S, Song B *et al.* (2015) The Symbiodinium kawagutii genome illuminates dinoflagellate gene expression and coral symbiosis. *Science*, **350**, 691–694.

Lionetti V, Metraux J-P (2015) *Plant cell wall in pathogenesis, parasitism and symbiosis*. Frontiers Media SA.

Lopez P, Halary S, Bapteste E (2015) Highly divergent ancient gene families in metagenomic samples are compatible with additional divisions of life. *Biology Direct*, **10**, 64.

Marchler-Bauer A, Anderson JB, Cherukuri PF *et al.* (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Research*, **33**, D192–D196.

Massana R, Gobet A, Audic S *et al.* (2015) Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environmental Microbiology*, **17**, 4035–4049.

Matzke M, Weiger TM, Papp I, Matzke AJM (2009) Nuclear membrane ion channels mediate root nodule development. *Trends in Plant Science*, **14**, 295–298.

Méheust R, Zelzion E, Bhattacharya D, Lopez P, Bapteste E (2016) Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. *Proceedings of the National Academy of Sciences*, **113**, 3579–3584.

Meyer JM, Rödelsperger C, Eichholz K *et al.* (2015) Transcriptomic characterisation and genomic glimps into the toxigenic dinoflagellate Azadinium spinosum, with emphasis on polykeitde synthase genes. *BMC Genomics*, **16**.

Monroe EA, Van Dolah FM (2008) The Toxic Dinoflagellate Karenia brevis Encodes Novel Type I-like Polyketide Synthases Containing Discrete Catalytic Domains. *Protist*, **159**, 471–482.

Montagnes DJS, Lowe CD, Roberts EC *et al.* (2011) An introduction to the special issue: Oxyrrhis marina, a model organism? *Journal of Plankton Research*, **33**, 549–554.

Mulder CPH, Bazeley-White E, Dimitrakopoulos PG *et al.* (2004) Species evenness and productivity in experimental plant communities. *Oikos*, **107**, 50–63.

Murray SA, Diwan R, Orr RJS, Kohli GS, John U (2015) Gene duplication, loss and selection in the evolution of saxitoxin biosynthesis in alveolates. *Molecular Phylogenetics and Evolution*, **92**, 165–180.

Murray SA, Suggett DJ, Doblin MA *et al.* (2016) Unravelling the functional genetics of dinoflagellates: a review of approaches and opportunities. *Perspectives in Phycology*, 37–52.

Orr RJS, Murray SA, Stüken A, Rhodes L, Jakobsen KS (2012) When Naked Became Armored: An Eight-Gene Phylogeny Reveals Monophyletic Origin of Theca in Dinoflagellates (S Lin, Ed,). *PLoS ONE*, **7**, e50004.

Orr RJS, Stüken A, Murray SA, Jakobsen KS (2013) Evolutionary Acquisition and Loss of Saxitoxin Biosynthesis in Dinoflagellates: the Second "Core" Gene, sxtG. *Applied and Environmental Microbiology*, **79**, 2128–2136.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics (Oxford, England)*, **23**, 1061–1067.

Perini F, Galluzzi L, Dell'Aversano C *et al.* (2014) SxtA and sxtG Gene Expression and Toxin Production in the Mediterranean Alexandrium minutum (Dinophyceae). *Marine Drugs*, **12**, 5258–5276.

Probert I, Siano R, Poirier C *et al.* (2014) Brandtodinium gen. nov. and B. nutricula comb. Nov. (Dinophyceae), a dinoflagellate commonly found in symbiosis with polycystine radiolarians. *Journal of Phycology*, **50**, 388–399.

Rengefors K, Karlsson I, Hansson L-A (1998) Algal cyst dormancy: a temporal escape from herbivory. *Proceedings of the Royal Society B: Biological Sciences*, **265**, 1353–1358.

Salcedo T, Upadhyay RJ, Nagasaki K, Bhattacharya D (2012) Dozens of Toxin-Related Genes Are Expressed in a Nontoxic Strain of the Dinoflagellate Heterocapsa circularisquama. *Molecular Biology and Evolution*, **29**, 1503–1506.

Sheng J, Malkiel E, Katz J, Adolf JE, Place AR (2010) A dinoflagellate exploits toxins to immobilize prey prior to ingestion. *Proceedings of the National Academy of Sciences*, **107**, 2082–2087.

Shoguchi E, Shinzato C, Kawashima T *et al.* (2013) Draft Assembly of the Symbiodinium minutum Nuclear Genome Reveals Dinoflagellate Gene Structure. *Current Biology*, **23**, 1399–1408.

Siano R, Alves-de-Souza C, Foulon E *et al.* (2011) Distribution and host diversity of Amoebophryidae parasites across oligotrophic waters of the Mediterranean Sea. *Biogeosciences*, **8**, 267–278.

Siano R, Montresor M, Probert I, Not F, de Vargas C (2010) Pelagodinium gen. nov. and P. béii comb. nov., a dinoflagellate symbiont of planktonic foraminifera. *Protist*, **161**, 385–399.

Sibbald SJ, Archibald JM (2017) More protist genomes needed. *Nature Ecology & Evolution*, **1**, 145.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, btv351.

Snyder RV, Gibbs PDL, Palacios A *et al.* Polyketide Synthase Genes from Marine Dinoflagellates. *Marine Biotechnology*, **5**, 1–12.

Stoecker DK, Hansen PJ, Caron DA, Mitra A (2017) Mixotrophy in the Marine Plankton. *Annual Review of Marine Science*, **9**, 311–335.

Stüken A, Orr RJS, Kellmann R *et al.* (2011) Discovery of Nuclear-Encoded Genes for the Neurotoxin Saxitoxin in Dinoflagellates. *PLOS ONE*, **6**, e20096.

Trench RK, Blank RJ (1987) Symbiodinium Microadriaticum Freudenthal, S. Goreauii Sp. Nov., S. Kawagutii Sp. Nov. and S. Pilosum Sp. Nov.: Gymnodinioid Dinoflagellate Symbionts of Marine Invertebrates 1. *Journal of Phycology*, **23**, 469–481.

Vonk FJ, Casewell NR, Henkel CV *et al.* (2013) The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proceedings of the National Academy of Sciences*, **110**, 20651–20656.

Wang D-Z (2008) Neurotoxins from Marine Dinoflagellates: A Brief Review. *Marine Drugs*, **6**, 349–371.

Yang Y, Smith SA (2013) Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics*, **14**, 328.

Yang L, Tan J, O'Brien EJ *et al.* (2015) Systems biology definition of the core proteome of metabolism and expression is consistent with high-throughput data. *Proceedings of the National Academy of Sciences*, **112**, 10810–10815.

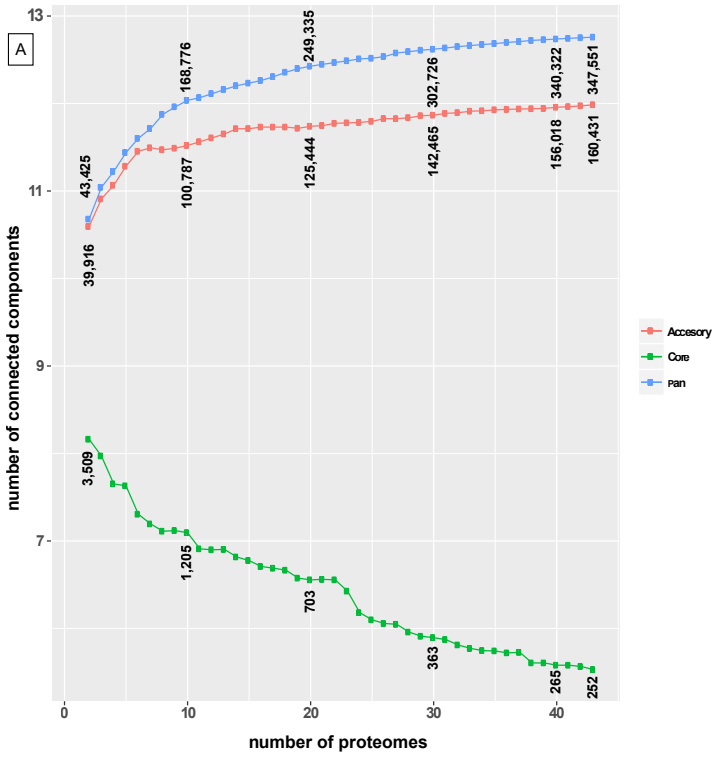
Yuasa T, Horiguchi T, Mayama S, Takahashi O (2016) Gymnoxanthella radiolariae gen. et sp. nov. (Dinophyceae), a dinoflagellate symbiont from solitary polycystine radiolarians. *Journal of Phycology*, **52**, 89–104.

Zhang Y, Zhang S-F, Lin L, Wang D-Z (2014) Comparative Transcriptome Analysis of a Toxin-Producing Dinoflagellate Alexandrium catenella and Its Non-Toxic Mutant. *Marine Drugs*, **12**, 5698–5718.

	TAXONOMY					ASSEMBLY METRICS					NETWORK	FUNCTIONAL TRAITS									
ID	order	family	genus	specie	strain	# contigs	N50	remapping rates	# protein coding domains	# annotations	presence	chloroplast type	mixotrophy	harmful for humans	symbiont	kleptoplasty	ichyotoxicity	parasitism	DMSP	thecate	cyst forming
1	Dinophysiales	Dinophysiaceae	*Dinophysis*	*acuminata*	DAEP01	123 473	747	73,07	57 612	21 401	y	C	y	DSP	n	y	n	n	y	y	n
2	Gonyaulacales	Ceratiaceae	*Ceratium*	*fusus*	PA161109	147 425	1 234	81,94	77 757	28 582	y	P	y	n	n	n	n	n	n	n	n
3	Gonyaulacales	Crypthecodiniaceae	*Crypthecodinium*	*cohnii*	Seligo	102 139	1 396	84,64	37 992	15 703	y	R	n	n	n	n	n	n	y	y	n
4	Gonyaulacales	Goniodomataceae	*Gambierdiscus*	*australes*	CAWD149	95 306	812	83,2	47 902	17 321	y	P	n	CFP	n	n	n	n	n	n	n
5	Gonyaulacales	Goniodomataceae	*Pyrodinium*	*bahamense*	pbaha01	142 061	772	75,18	73 648	26 001	y	P	n	PSP	n	n	n	n	n	n	y
6	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*andersoni*	CCMP2222	97 010	438	56,64	41 556	13 166	y	P	y	PSP	n	n	n	n	n	y	y
7	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*catenella*	OF101	95 316	570	69,07	51 078	17 364	y	P	y	PSP	n	n	n	n	n	y	y
8	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*fundyense*	CCMP1719	6 491	413	3,14	2 477	1 092	n	P	n	PSP	n	n	n	y	y	y	y
9	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*margalefi*	AMGDE01CS-322	145 973	825	80,58	87 070	29 537	y	P	n	n	n	n	n	n	n	y	y
10	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*minutum*	CCMP113	21 364	550	41	10 572	3 817	y	P	y	PSP	n	n	n	n	y	y	y
11	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*monilatum*	CCMP3105	114 652	1 404	84,63	75 921	26 620	y	P	y	PSP	n	n	n	n	n	y	y
12	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*tamarense*	CCMP1771	176 197	1 065	79,45	91 414	36 592	y	P	y	PSP	n	n	n	n	y	y	y
13	Gonyaulacales	Gonyaulacaceae	*Gonyaulax*	*spinifera*	CCMP409	70 621	634	73,93	33 151	12 928	y	P	y	n	n	n	n	n	y	y	y
14	Gonyaulacales	Gonyaulacaceae	*Lingulodinium*	*polyedra*	CCMP1738	131 324	1 278	86,04	80 900	28 396	y	P	y	DSP	n	n	n	n	y	y	y
15	Gonyaulacales	Gonyaulacaceae	*Protoceratium*	*reticulatum*	CCCM535=CCMP1889	96 484	650	70,13	50 156	18 700	y	P	n	DSP	n	n	n	n	n	n	y
16	Gymnodiniales	Gymnodiniaceae	*Akashiwo*	*sanguinea*	CCCM885	1 016	287	5,1	103	19	n	P	y	n	n	n	n	n	y	n	y
17	Gymnodiniales	Gymnodiniaceae	*Amphidinium*	*carterae*	CCMP1314	60 662	1 590	74	37 749	15 747	y	P	y	n	n	n	y	n	y	n	n
18	Gymnodiniales	Gymnodiniaceae	*Amphidinium*	*massartii*	CS-259	76 973	1 280	85,32	40 678	16 207	y	P	y	n	n	n	n	n	n	n	n
19	Gymnodiniales	Gymnodiniaceae	*Gymnodinium*	*catenatum*	GC744	124 421	836	74,72	54 459	22 417	y	P	y	PSP	n	n	n	n	n	n	y
20	Gymnodiniales	Gymnodiniaceae	*Gymnoxanthella*	*radiolariae*	RCC3507	160 971	1 683	93,11	102 709	39 515	y	P	n	n	y	n	n	n	y	n	n
21	Gymnodiniales	Gymnodiniaceae	*Gyrodinium*	*dominans*	SPMC103	513	277	2,58	67	5	n	R	n	n	n	n	n	n	n	n	n
22	Gymnodiniales	Gymnodiniaceae	*Togula*	*jolla*	CCCM725	73 075	1 054	81,34	35 840	15 309	y	P	NA	n	n	n	n	n	n	n	n
23	Gymnodiniales	Gymnodiniaceae	*Karlodinium*	*micrum*	CCMP2283	142 286	1 330	84,41	65 800	28 395	y	H	y	n	n	n	y	n	n	n	n
24	Gymnodiniales	Kareniaceae	*Karenia*	*brevis*	CCMP2229;Wilson;SP3;SP1						n	H	y	NSP	n	n	y	n	y	n	n
25	noctilucales	noctilucaceae	*Noctiluca*	*scintillans*	NA	66 050	1 230	84,07	33 017	14 223	y	R	n	n	n	n	n	n	n	n	n
26	Oxyrrhinales	Oxyrrhinaceae	*Oxyrrhis*	*marina*	NA;LB1974	18 275	569	42,15	5 189	2 402	y	R	n	n	n	n	n	n	n	n	y
27	Oxyrrhinales	Oxyrrhinaceae	*Oxyrrhis*	*marina*	CCMP1788	340	266	2,03	45	25	n	R	n	n	n	n	n	n	n	n	y
28	Oxyrrhinales	Oxyrrhinaceae	*Oxyrrhis*	*marina*	CCMP1795						n	R	n	n	n	n	n	n	n	n	y
29	Peridiniales	Heterocapsaceae	*Heterocapsa*	*sp.*	RCC1516	225 203	1 289	88,62	107 673	36 966	y	P	y	n	n	n	n	n	y	y	n
30	Peridiniales	Heterocapsaceae	*Heterocapsa*	*arctica*	CCMP445	62 237	628	66,3	33 122	12 078	y	P	NA	n	n	n	n	n	n	y	n
31	Peridiniales	Heterocapsaceae	*Heterocapsa*	*rotundata*	SCCAPK-0483	69 955	774	72,65	39 543	14 077	y	P	y	n	n	n	n	n	y	y	n
32	Peridiniales	Heterocapsaceae	*Heterocapsa*	*triquetra*	CCMP448	89 751	698	68,45	44 370	16 265	y	P	y	n	n	n	n	n	y	y	n
33	Peridiniales	*incertae sedis*	*Azadinium*	*spinsum*	3D9	152 890	1 269	83,7	76 500	30 065	y	P	NA	AZP	n	n	n	n	n	y	n
34	Peridiniales	Lessardiaceae	*Lessardia*	*elongata*	SPMC104	1 418	285	5,18	118	33	n	R	n	n	n	n	n	n	n	y	n
35	Peridiniales	Peridiniaceae	*Brandtodinium*	*nutricula*	RCC3387	92 032	672	66,47	59 250	19 378	y	P	n	n	y	n	n	n	y	y	n
36	Peridiniales	Peridiniaceae	*Brandtodinium*	*nutricula*	RCC3468	187 598	1 199	89,84	115 229	36 197	y	P	n	n	y	n	n	n	y	y	n
37	Peridiniales	Peridiniaceae	*Durinskia*	*baltica*	CSIRO_CS-38	158 433	836	77,96	71 415	29 330	y	D	NA	n	n	n	n	n	n	y	NA
38	Peridiniales	Peridiniaceae	*Glenodinium*	*foliaceum*	CCAP1116/3	154 714	746	76,33	82 653	29 409	y	P	NA	n	n	n	n	n	n	n	y
39	Peridiniales	Peridiniaceae	*Kryptoperidinium*	*foliaceum*	CCMP1326	254 192	792	70,28	135 557	48 835	y	D	NA	n	n	y	n	n	y	y	n
40	Peridiniales	Peridiniaceae	*Peridinium*	*aciculiferum*	PAER-2	38 083	352	30,02	14 079	4 952	n	P	NA	n	n	n	n	n	n	y	y
41	Peridiniales	Peridiniaceae	*Scrippsiella*	*hangoei*	SHTV-5	194 233	1 526	86,93	114 374	37 917	y	P	n	n	n	n	n	n	n	y	y
42	Peridiniales	Peridiniaceae	*Scrippsiella*	*hangoei-like*	SHHI-4	406	311	4,42	130	55	n	P	n	n	n	n	n	n	n	y	y
43	Peridiniales	Peridiniaceae	*Scrippsiella*	*trochoidea*	CCMP3099	160 890	1 386	82,78	90 198	34 206	y	P	y	n	n	n	n	n	n	y	y
44	Prorocentrales	Prorocentraceae	*Prorocentrum*	*lima*	CCMP684	3 911	311	10,19	524	123	n	P	y	DSP	n	n	n	n	n	n	y
45	Prorocentrales	Prorocentraceae	*Prorocentrum*	*micans*	CCCM845	7 504	311	4,45	909	259	n	P	y	n	n	n	n	n	n	y	y
46	Prorocentrales	Prorocentraceae	*Prorocentrum*	*minimum*	CCMP1329;CCMP2233	110 115	710	66,53	45 564	17 693	y	P	y	DSP	n	n	n	n	n	n	y
47	Pyrocystales	Pyrocystaceae	*Pyrocystis*	*lunula*	CCCM517	1 085	281	2,44	106	11	n	P	y	n	n	n	n	n	n	y	n
48	Suessiales	Suessiaceae	*Pelagodinium*	*beii*	RCC1491	154 473	1 513	92,2	111 658	36 604	y	P	n	n	y	n	n	n	y	n	n
49	Suessiales	Suessiaceae	*Pelagodinium*	*beii*	RCC1491	99 728	946	75,84	44 901	18 705	y	P	n	n	y	n	n	n	n	y	n
50	Suessiales	Suessiaceae	*Polarella*	*glacialis*	CCMP1383;CCMP2088	108 029	794	68,56	46 056	17 766	y	P	n	n	n	n	n	n	y	n	y
51	Suessiales	Symbiodiniaceae	*Symbiodinium*	*kawagutii*	CCMP2468						n	P	n	n	y	n	n	n	n	y	n
52	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	D1a	142 720	493	53,9	52 578	20 131	y	P	n	n	y	n	n	n	y	y	n
53	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	CCMP421	136 116	965	75,98	81 880	29 878	y	P	n	n	y	n	n	n	y	y	n
54	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	C15	101 453	1 024	80,66	48 884	18 687	y	P	n	n	y	n	n	n	y	y	n
55	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	C1	89 177	1 181	84,07	52 745	21 085	y	P	n	n	y	n	n	n	y	y	n
56	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	CCMP2430	79 016	1 082	87,26	50 160	19 992	y	P	n	n	y	n	n	n	y	y	n
57	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	Mp	74 565	1 416	89,35	46 513	18 550	y	P	n	n	y	n	n	n	y	y	n
58	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	cladeA	72 446	947	80,77	41 846	14 844	y	P	n	n	y	n	n	n	y	y	n
59	Syndiniales	Amoebophryaceae	*Amoebophrya*	*sp.*	Ameob2	26 721	1 856	87,46	5 548	2 075	y	R	n	n	n	n	n	y	n	n	n
60	Thoracosphaerales	Thoracosphaeraceae	*Thoracosphaera*	*heimii*	CCCM670=CCMP1069	517	277	2,32	81	3	n	P	n	n	n	n	n	n	n	y	n	n

Tab. 1: Summary table of the 60 transcriptomes (and the corresponding strains) analyzed in this study, ranked based on their taxonomy. Assembly metrics are reported for each transcriptome encompassing: the number of assembled contigs, N50, the remapping rate of initial reads, the number of predicted protein domains found in transcript sequences and the number of functional annotations identified through Interproscan 5. The network presence column indicates the “high quality” transcriptomes for which derived proteomes were included in the final network. Based on a literature survey, information about functional traits for each species included in the dataset is provided: chloroplast type (P: peridinin, H: haptophyte-like, C: cryptomonad-like, D: diatom-like and R: remnant or absent plastid), mixotrophy, ability to produce toxins harmful for humans (DSP:Diarrhetic shellfish poisoning, CFP:Ciguatera Fish Poisoning, PSP:Paralytic shellfish poisoning, AZP:Azaspiracid Shellfish Poisoning, NSP:Neurologic Shellfish Poisoning), ability to be symbionts, kleptoplasty, ichyotoxicity, parasitism, ability to produce DSMP, presence of a theca, ability to form cysts during life-cycle. <NA> corresponds to a lack of information.

A



B

		Number of sequence	Percentage of sequences (*)	Number of involved components	Percentage involved components (**)
sequence homology	BUSCO version: 06/2016	4,737	12.5%	51	20.2%
	UniProtKB/SwissProt version: 06/2016	30,138	79.6%	190	75.4%
	nr version: 12/12/2015	35,455	93.7%	236	93.7%
protein domain	InterPro version: 06/2016	334,573	91.4%	226	89.7%
	Remaining unnanotated domains	946	2.5%	16	6.3%

* Total number of core sequences = 37,842
** Total number of core components = 252

Fig. 1: (A) Number of connected components (CCs) in the core (green), accessory (red) and pan (blue) dinoflagellate proteomes, considering 2 to 43 proteomes. (B) Comparison of the 37,842 protein domains included in the 252 core dinoflagellates CCs to BUSCO, UniProtKB/Swiss-Prot and nr databases. The number and percentage of core sequences with at least one match in each database, and the number and percentage of their corresponding CCs.

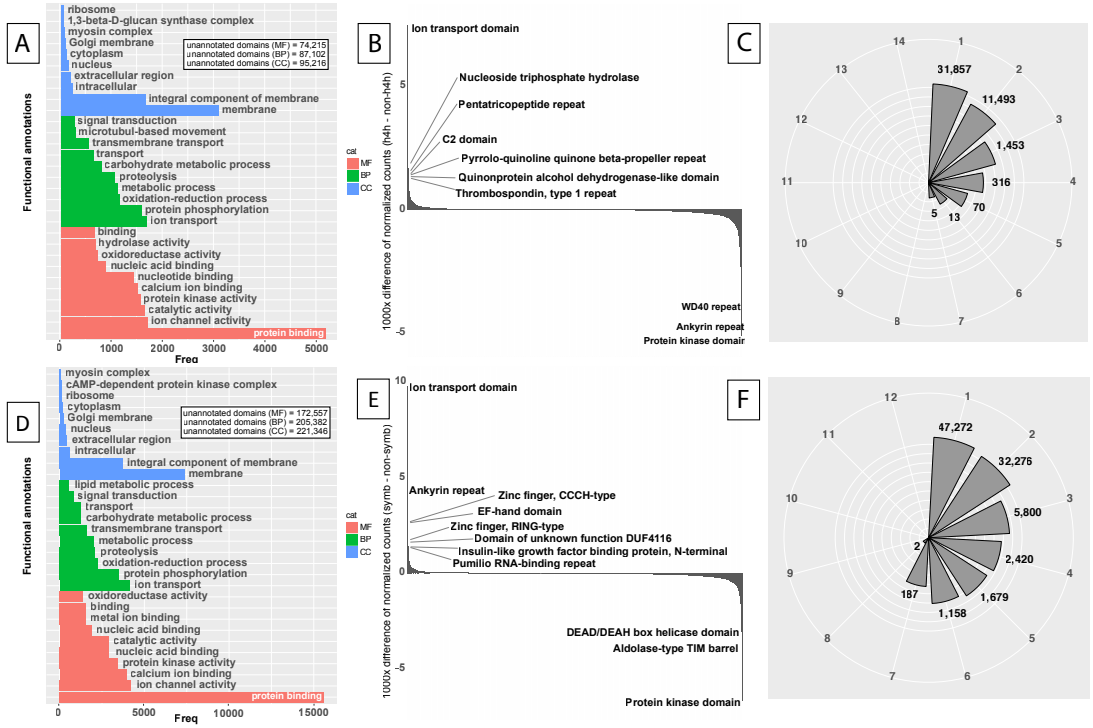


Fig. 2: (A and D) Top 10 functional annotations (GOslim levels) of sequences belonging to the 45,207 “harmful for human” trait-CCs (A) and to the 90,794 “symbiosis” trait-CCs (D). (B and E) Differential composition of functional annotations between “harmful for human” and non-“harmful for human” trait-CCs (B) and “symbiosis” and non-“symbiosis” trait-CCs (E). (C and F) The circular barplot shows the number of connected components that include 1 to 14 proteome(s) of the transcriptomes assigned to toxic species (C) and the number of connected components that include 1 to 12 proteome(s) of the transcriptomes assigned to symbiotic species (F).

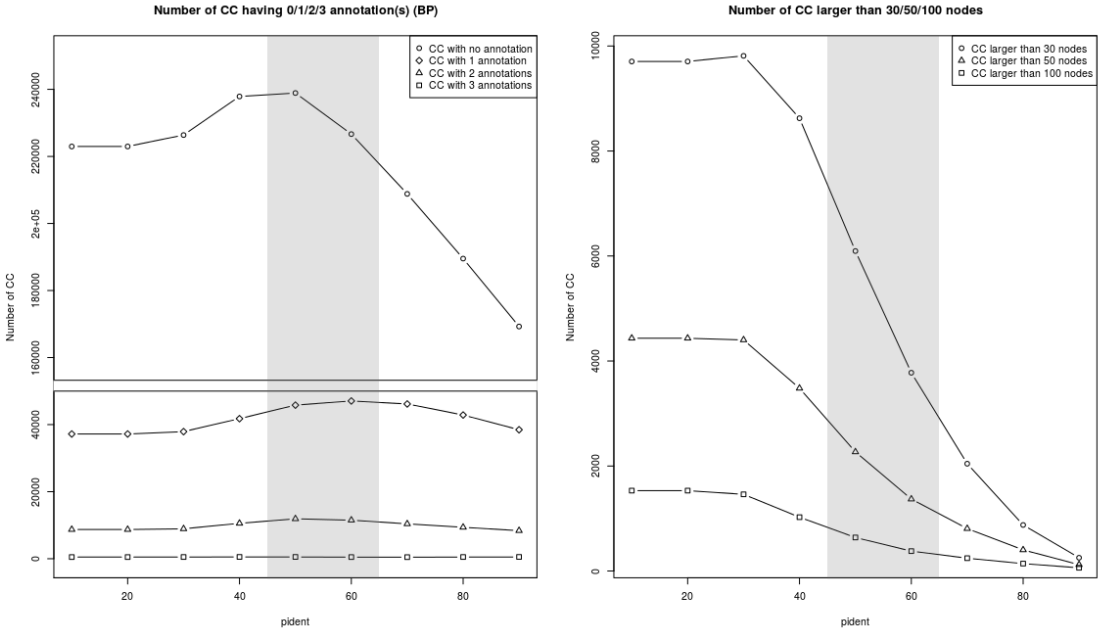


Fig. S 1: Optimal sequence identity threshold selection. The cutoff was chosen such that: (A) the network contains a maximum of connected components with homogeneous functional annotation (i.e. a unique GO Slim term for all annotated protein coding domains in each CC) and (B) the network conserved a maximum of « large » connected components.

Fig. S 2 : Top 10 functional annotations (GOslim levels) in all core components. The three levels of annotation are represented: Molecular Function level (MF, red), Biological Process (BP, green) and Cellular Component (CC, blue).

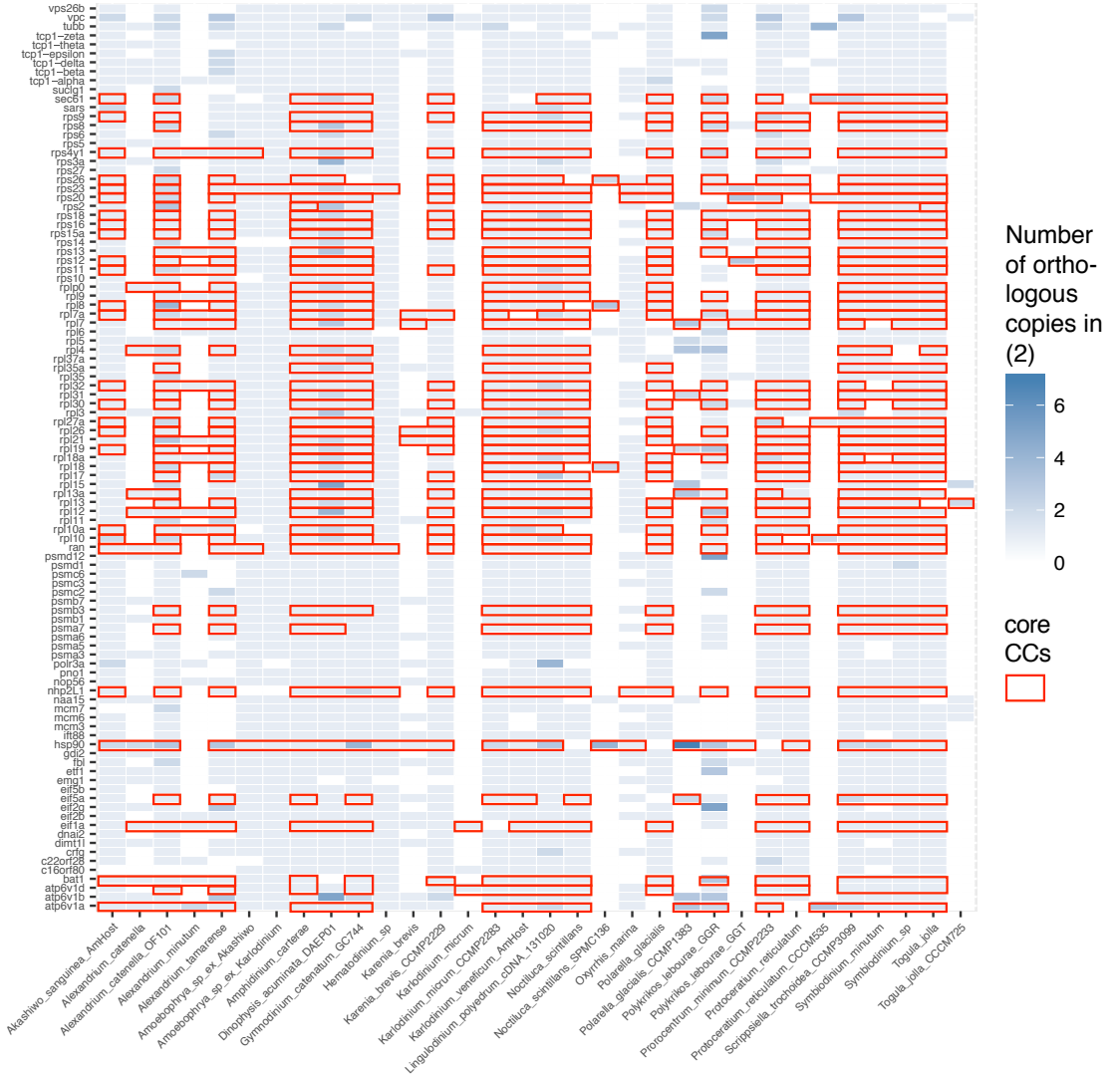


Fig. S 3 : Overlap between the 101 multiple protein alignments from (2) (used for phylogenomics) and our core dinoflagellate proteome. The number of copies of the 101 proteins is represented here in a heat map: the blue color gradient represents the number of orthologous sequences available of a particular protein (y-axis) for each studied dinoflagellate species in (2) (x-axis). The overlap between the data set from (2) and our core proteome is represented by the red boxes: i.e. each protein sequence from (2) that aligned to at least a core protein coding domain is here delimited in red.

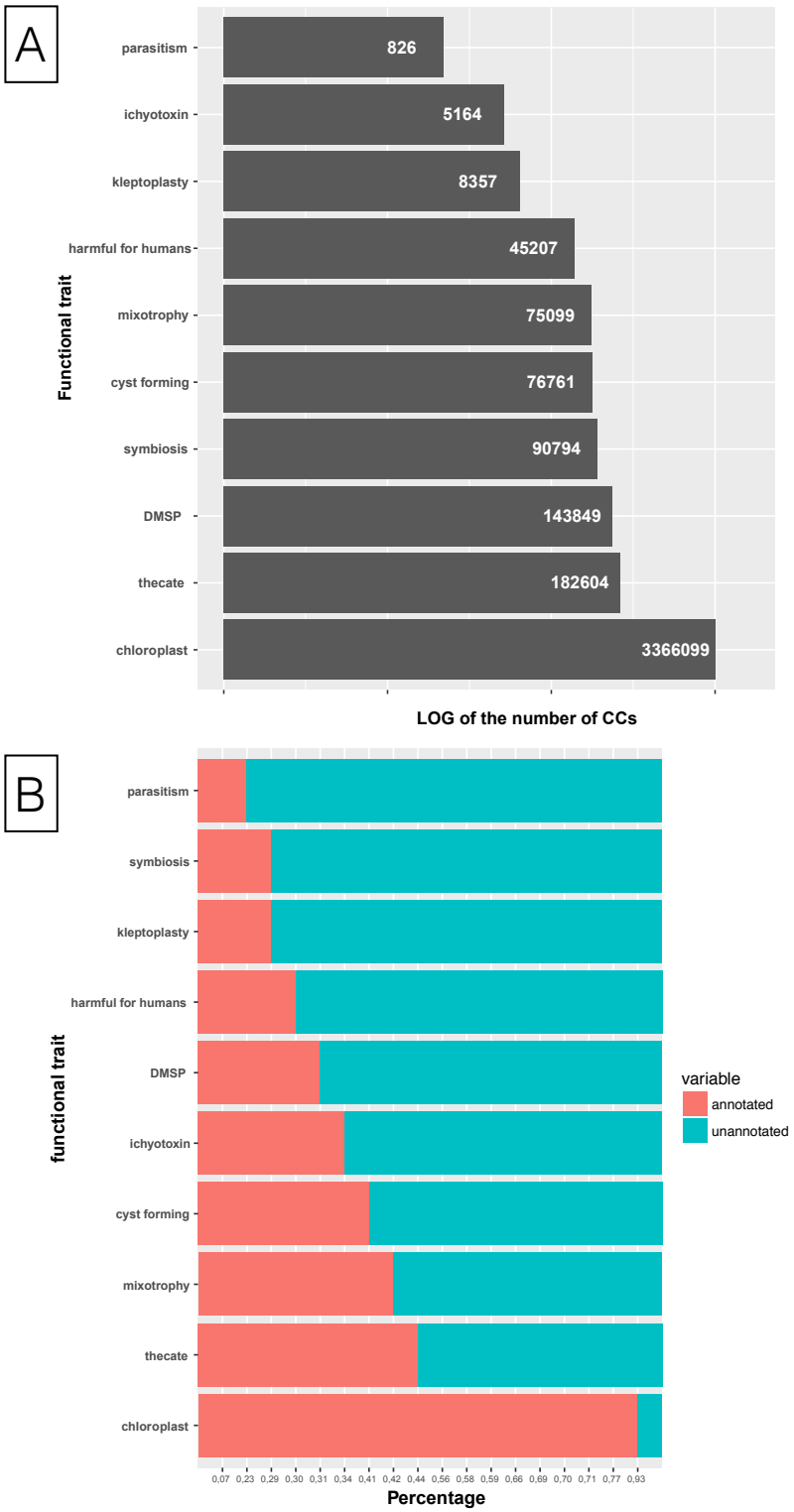


Fig. S 4: (A) Number of connected components for each functional trait. (B) Proportion of annotated sequences of connected components for each functional trait.

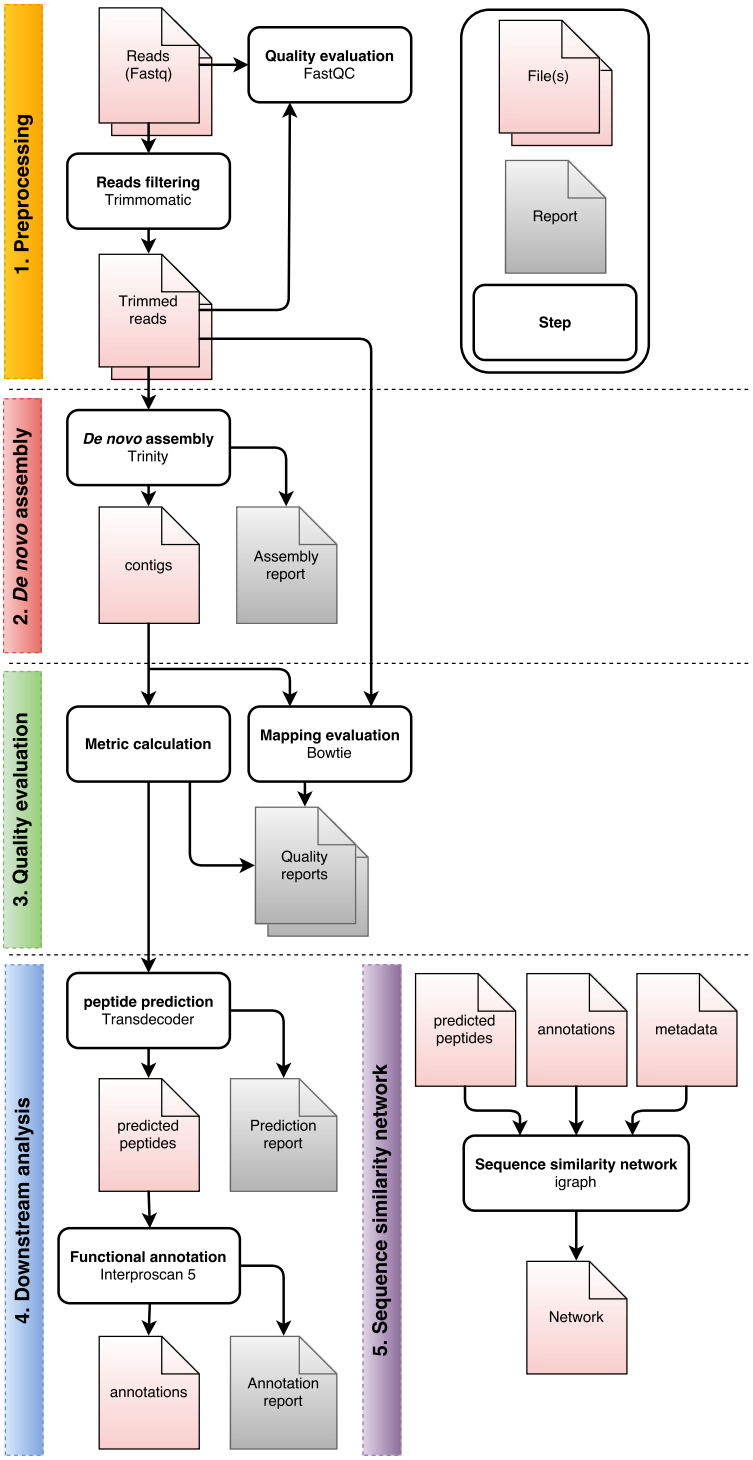


Fig. S 5: Pipeline diagram of our analysis composed of 5 distinct steps (for more details see Material & Methods): (1) Preprocessing step including read quality evaluation and filtering; (2) De novo assembly step in which assembled contigs were generated from cleaned reads with Trinity (ref. 57). (3) Quality evaluation of the previously assembled contigs. (4) Downstream analysis divided into two parts, with first detection of likely coding domains within contig sequences and then functional annotation of previously detected domains. (5) Construction of a sequence similarity network based on de novo assembly and downstream analysis results.

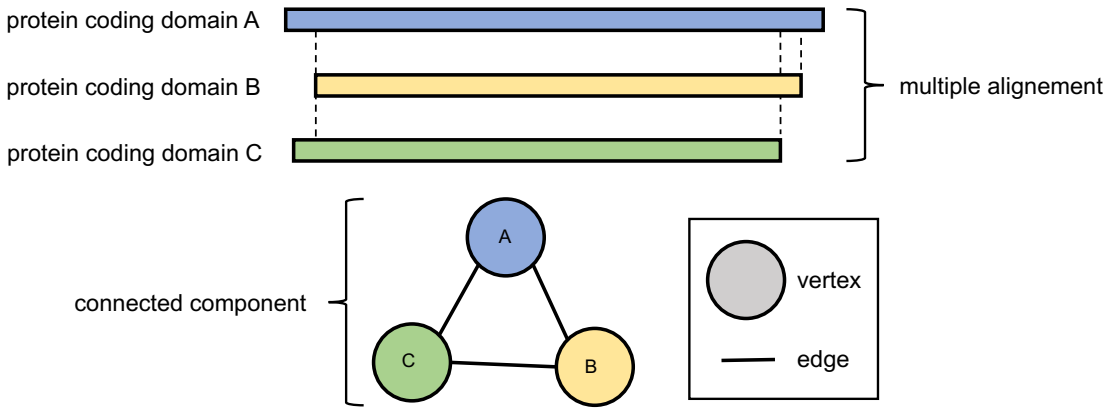


Fig. S 6: A connected component outline. At the top, a multiple alignment of 3 protein coding domain sequences A, B and C. The 3 alignments respect sequence identity threshold (>60%) and sequence coverage threshold (>80%). At the bottom, a sketch of the corresponding connected component where protein coding domain sequences are represented by vertices and each alignment between two sequences is represented by and edge.

MMETSP RNA-seq datasets

ID	TAXONOMY				strain	MMETSP ID
	order	family	genus	specie		
1	Dinophysiales	Dinophysiaceae	*Dinophysis*	*acuminata*	DAEP01	MMETSP0797
2	Gonyaulacales	Ceratiaceae	*Ceratium*	*fusus*	PA161109	MMETSP1074 , MMETSP1075
3	Gonyaulacales	Crypthecodiniaceae	*Crypthecodinium*	*cohnii*	Seligo	MMETSP0323_2, MMETSP0324_2, MMETSP0325_2, MMETSP0326_2
4	Gonyaulacales	Goniodomataceae	*Gambierdiscus*	*australes*	CAWD149	MMETSP0766_2
5	Gonyaulacales	Goniodomataceae	*Pyrodinium*	*bahamense*	pbaha01	MMETSP0796
6	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*andersonii*	CCMP2222	MMETSP1436
7	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*catenella*	OF101	MMETSP0790
8	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*fundyense*	CCMP1719	MMETSP0196, MMETSP0197, MMETSP0347
9	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*margalefi*	AMGDE01CS-322	MMETSP0661
10	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*minutum*	CCMP113	MMETSP0328
11	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*monilatum*	CCMP3105	MMETSP0093, MMETSP0095, MMETSP0096, MMETSP0097
12	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*tamarense*	CCMP1771	MMETSP0378, MMETSP0380, MMETSP0382, MMETSP0384
13	Gonyaulacales	Gonyaulacaceae	*Gonyaulax*	*spinifera*	CCMP409	MMETSP1439
14	Gonyaulacales	Gonyaulacaceae	*Lingulodinium*	*polyedra*	CCMP1738	MMETSP1032, MMETSP1033, MMETSP1034, MMETSP1035
15	Gonyaulacales	Gonyaulacaceae	*Protoceratium*	*reticulatum*	CCCM535=CCMP1889	MMETSP0228
16	Gymnodiniales	Gymnodiniaceae	*Akashiwo*	*sanguinea*	CCCM885	MMETSP0223_2
17	Gymnodiniales	Gymnodiniaceae	*Amphidinium*	*carterae*	CCMP1314	MMETSP0258, MMETSP0259, MMETSP0398, MMETSP0399
18	Gymnodiniales	Gymnodiniaceae	*Amphidinium*	*massartii*	CS-259	MMETSP0689_2
19	Gymnodiniales	Gymnodiniaceae	*Gymnodinium*	*catenatum*	GC744	MMETSP0784
20	Gymnodiniales	Gymnodiniaceae	*Gyrodinium*	*dominans*	SPMC103	MMETSP1148
21	Gymnodiniales	Gymnodiniaceae	*Togula*	*jolla*	CCCM725	MMETSP0224
22	Gymnodiniales	Kareniaceae	*Karenia*	*brevis*	CCMP2229 Wilson SP3 SP1 Wilson	MMETSP0027, MMETSP0029, MMETSP0030, MMETSP0031 MMETSP0201, MMETSP0202 MMETSP0527_2, MMETSP0528_2 MMETSP0573, MMETSP0574 MMETSP0648_2, MMETSP0649_2
23	Gymnodiniales	Gymnodiniaceae	*Karlodinium*	*micrum*	CCMP2283	MMETSP1015, MMETSP1016, MMETSP1017
24	Noctilucales	Noctilucaceae	*Noctiluca*	*scintillans*	NA	MMETSP0253
25	Oxyrrhinales	Oxyrrhinaceae	*Oxyrrhis*	*marina*	NA LB1974	MMETSP0468, MMETSP0469, MMETSP0470, MMETSP0471 MMETSP1424, MMETSP1425, MMETSP1426
26	Oxyrrhinales	Oxyrrhinaceae	*Oxyrrhis*	*marina*	CCMP1788	MMETSP0044
27	Oxyrrhinales	Oxyrrhinaceae	*Oxyrrhis*	*marina*	CCMP1795	MMETSP0451_2, MMETSP0452_2
28	Peridiniales	Heterocapsaceae	*Heterocapsa*	*arctica*	CCMP445	MMETSP1441
29	Peridiniales	Heterocapsaceae	*Heterocapsa*	*rotundata*	SCCAPK-0483	MMETSP0503
30	Peridiniales	Heterocapsaceae	*Heterocapsa*	*triquetra*	CCMP448	MMETSP0448
31	Peridiniales	*incertae sedis*	*Azadinium*	*spinosum*	3D9	MMETSP1036_2, MMETSP1037_2, MMETSP1038_2
32	Peridiniales	Lessardiaceae	*Lessardia*	*elongata*	SPMC104	MMETSP1147
33	Peridiniales	Peridiniaceae	*Brandtodinium*	*nutricula*	RCC3387	MMETSP1462
34	Peridiniales	Peridiniaceae	*Durinskia*	*baltica*	CSIRO_CS-38	MMETSP0116_2, MMETSP0117_2
35	Peridiniales	Peridiniaceae	*Glenodinium*	*foliaceum*	CCAP1116/3	MMETSP0118_2, MMETSP0119_2
36	Peridiniales	Peridiniaceae	*Kryptoperidinium*	*foliaceum*	CCMP1326	MMETSP0120_2, MMETSP0121_2
37	Peridiniales	Peridiniaceae	*Peridinium*	*aciculiferum*	PAER-2	MMETSP0370_2, MMETSP0371_2
38	Peridiniales	Peridiniaceae	*Scrippsiella*	*hangoei*	SHTV-5	MMETSP0359, MMETSP0360, MMETSP0361
39	Peridiniales	Peridiniaceae	*Scrippsiella*	*hangoei-like*	SHHI-4	MMETSP0367, MMETSP0368, MMETSP0369
40	Peridiniales	Peridiniaceae	*Scrippsiella*	*trochoidea*	CCMP3099	MMETSP0270, MMETSP0271, MMETSP0272
41	Prorocentrales	Prorocentraceae	*Prorocentrum*	*lima*	CCMP684	MMETSP0252
42	Prorocentrales	Prorocentraceae	*Prorocentrum*	*micans*	CCCM845	MMETSP0251_2
43	Prorocentrales	Prorocentraceae	*Prorocentrum*	*minimum*	CCMP1329 CCMP2233	MMETSP0053, MMETSP0055, MMETSP0056, MMETSP0057 MMETSP0267, MMETSP0268, MMETSP0269
44	Pyrocystales	Pyrocystaceae	*Pyrocystis*	*lunula*	CCCM517	MMETSP0229_2
45	Suessiales	Suessiaceae	*Pelagodinium*	*beii*	RCC1491	MMETSP1338
46	Suessiales	Suessiaceae	*Polarella*	*glacialis*	CCMP1383	MMETSP0227
47	Suessiales	Symbiodiniaceae	*Symbiodinium*	*kawagutii*	CCMP2468	MMETSP0132_2, MMETSP0133_2, MMETSP0134_2, MMETSP0135_2
48	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	D1a	MMETSP1377
49	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	CCMP421	MMETSP1110
50	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	C15	MMETSP1370, MMETSP1371
51	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	C1	MMETSP1367, MMETSP1369
52	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	CCMP2430	MMETSP1115, MMETSP1116, MMETSP1117
53	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	Mp	MMETSP1122, MMETSP1123, MMETSP1124, MMETSP1125
54	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	cladeA	MMETSP1374
55	Syndiniales	Amoebophryaceae	*Amoebophrya*	*sp.*	Ameob2	MMETSP0795
56	Thoracosphaerales	Thoracosphaeraceae	*Thoracosphaera*	*heimii*	CCCM670=CCMP1069	MMETSP0225

Fig. S 7: Table of the MMETSP subsets with their ID that has been pooled into 56 single sets of reads. Each of the 56 sets was assembled de novo to create 56 transcriptomes.

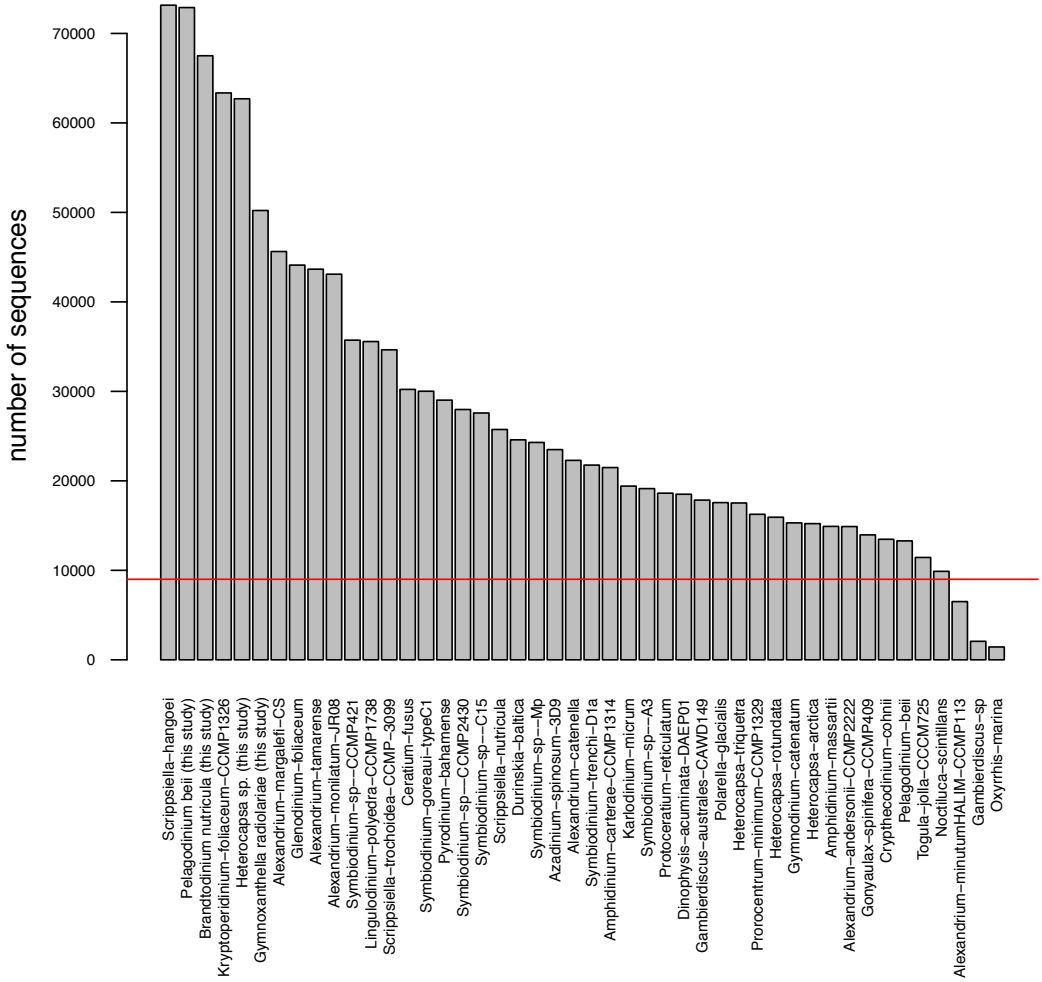


Fig. S 8: Number of peptide sequences per proteome derived from "high quality" transcriptomes. Red line represents the minimum number of sequences threshold (9,000 peptide sequences) required to perform core/accessory/pan proteome analysis.

SI Appendix Tab. S1 : Similarity network metrics

	Number of transcriptomes
MMETSP	56
Unpublished (our contribution)	4
Total	60
(A) : Total successully assembled	57
(A) of « low quality »	11
(B) : (A) of « high quality »	46

	Number of sequences	Average length (bp)
likely coding domains from (A)	1 283 775	307
likely coding domains from (A) having Interpro annotations	750 480	
likely coding domains from (A) having GO annotations	552 846	
likely coding domains from (B)	1 275 911	
likely coding domains from (B) having Interpro annotations	746 074	
likely coding domains from (B) having GO annotations	549 459	

	SNN metrics		
	Number of vertices	Number of edges	Number of CC
SSN-1 : build from (A)	1 283 755	6 213 111	352 153
SSN-2 : build from (B)	1 275 911	6 142 013	350 267

	Number of CC
SSN-2 : CC <= 10 vertices	338 327
SSN-2 : 10 <= CC <= 100 vertices	11 568
SSN-2 : 100 <= CC <= 1000 vertices	369
SSN-2 : CC >= 1000 vertices	3

	Number of vertices
SSN-2 : minimum of CC sizes	2
SSN-2 : ^{1}st Qu. Of CC sizes	2
SSN-2 : Median of CC sizes	2
SSN-2 : Mean of CC sizes	4
SSN-2 : ^{3}rd Qu. Of CC sizes	3
SSN-2 : maximum of CC sizes	1600

	Number of CC
SSN-2 : CC having a single homogeneous annotation (BP)	47 061
SSN-2 : CC having 2 annotations (BP)	11 508
SSN-2 : CC having 3 annotations (BP)	487
SSN-2 : CC having more than 3 annotations (BP)	139

SI Appendix Tab. S2 : core / accessory / pan CCs

	Number of CC	Number of sequences
core CC	252	37 842
accessory CC	160 431	NA
pan CC	347 551	NA

	Number of CC
core : minimum of CC sizes	44,00
core : 1st Qu. Of CC sizes	62,25
core : Median of CC sizes	116,00
core : Mean of CC sizes	150,20
core : 3rd Qu. Of CC sizes	215,00
core : maximum of CC sizes	564,00

	Number of hits	Number of involved core sequences	Number of involved core CC
blastp : core vs. nr (09/2016) (evalue=1e-25)	871 901	35 455	236
blastp : core vs. sp (09/2016) (evalue=1e-25)	637 746	30 138	190
core vs. BUSCO (09/2016)	5 263	4 737	51
non annotated CC (vs. Interpro)	NA	1 315	27
non annotated core sequences (vs. All databases)	NA	946	16

	Number of sequences
core vs. BUSCO : complete	1
core vs. BUSCO : duplicated	4 009
core vs. BUSCO : fragmented	1 253
core vs. BUSCO : missing	387

Top 10 occuring Goslim functional annotations of core sequences		
annotation	Freq	level
structural constituent of ribosome	7 968	MF
protein kinase activity	1 752	MF
catalytic activity	1 739	MF
nucleic acid binding	1 712	MF
RNA binding	1 630	MF
GTPase activity	1 383	MF
calcium ion binding	1 054	MF
GTP binding	1 043	MF
protein binding	1 028	MF
hydrolase activity	701	MF
translation	7 981	BP
protein phosphorylation	1 752	BP
signal transduction	1 133	BP
cell redox homeostasis	562	BP
protein peptidyl-prolyl isomerization	518	BP
metabolic process	472	BP
protein folding	421	BP
glycolytic process	397	BP
carbohydrate metabolic process	390	BP
protein retention in ER lumen	382	BP
intracellular	5 609	CC
ribosome	2 264	CC
integral component of membrane	636	CC
membrane	575	CC
phosphopyruvate hydratase complex	397	CC
large ribosomal subunit	342	CC
proton-transporting two-sector ATPase complex, proton-transporting domain	278	CC
nucleus	235	CC
cytoplasm	213	CC
ribonucleoprotein complex	188	CC

SI Appendix Tab. S3 : Function trait : chloroplast

Number of transcriptomes corresponding to chloroplatic species
42

Number of CC composed of chloroplastic species sequences	Unannotated
336 099	220 081

Number of CC composed of N transcriptomes	
N	# CC
1	156 073
2	117 449
3	29 256
4	11 841
5	6 819
6	4 482
7	2 565
8	1 570
9	1 107
10	898
11	656
12	499
13	375
14	350
15	277
16	242
17	201
18	164
19	163
20	129
21	142
22	114
23	97
24	75
25	71
26	63
27	63
28	53
29	47
30	43
31	38
32	25
33	37
34	23
35	29
36	21
37	16
38	12
39	4
40	6
41	1
42	3
43	0

Most occuring BP sequenceannotations for chloroplatic CC	
annotations	occurencies
NA	833713
oxidation-reduction process	18445
protein phosphorylation	18042
metabolic process	15740
ion transport	11810
proteolysis	11550
carbohydrate metabolic process	7854
transmembrane transport	7144
transport	6715
microtubule-based movement	4216
photosynthesis, light harvesting	4145
translation	3684
signal transduction	2848
photosynthesis	2739
lipid metabolic process	2692
protein folding	2326
cell redox homeostasis	1862
biosynthetic process	1754
regulation of transcription, DNA-templated	1753
pseudouridine synthesis	1582
protein dephosphorylation	1274
protein glycosylation	1255
glycolytic process	1241
carbohydrate transport	1235
ATP synthesis coupled proton transport	1231
cation transport	1204
DNA repair	1198
ubiquitin-dependent protein catabolic process	1196
protein peptidyl-prolyl isomerization	1106
tRNA aminoacylation for protein translation	926

Most occuring MF sequence annotations for chloroplatic CC	
annotations	occurencies
NA	680507
protein binding	66778
catalytic activity	24140
protein kinase activity	17794
calcium ion binding	15259
nucleotide binding	13829
oxidoreductase activity	12732
ion channel activity	12259
nucleic acid binding	10346
ATP binding	7663
binding	7044
RNA binding	5891
metal ion binding	5661
microtubule motor activity	5256
hydrolase activity, hydrolyzing O-glycosyl compounds	5254
hydrolase activity	5093
DNA binding	5077
zinc ion binding	4683
iron ion binding	4503
transporter activity	4391
methyltransferase activity	3525
structural constituent of ribosome	3472
NAD+ ADP-ribosyltransferase activity	2647
transferase activity, transferring phosphorus-containing groups	2501
aspartic-type endopeptidase activity	2477
GTP binding	2457
serine-type endopeptidase activity	2455
ATPase activity, coupled to transmembrane movement of substances	1945
cysteine-type peptidase activity	1825
ubiquitin-protein transferase activity	1786

Most occuring CC sequenceannotations for chloroplatic CC	
annotations	occurencies
NA	942249
membrane	31877
integral component of membrane	19919
intracellular	4253
nucleus	2448
cytoplasm	2146
extracellular region	1410
Golgi membrane	1271
ribosome	930
axonemal dynein complex	798
photosystem I	770
myosin complex	601
proton-transporting ATP synthase complex, coupling factor F(o)	555
cAMP-dependent protein kinase complex	535
mitochondrion	535
extrinsic component of membrane	500
1,3-beta-D-glucan synthase complex	371
spliceosomal complex	310
proton-transporting ATP synthase complex, catalytic core F(1)	303
nucleosome	297
nuclear pore	292
cytochrome b6f complex	289
phosphopyruvate hydratase complex	284
kinesin complex	280
photosystem II	235
endoplasmic reticulum	234
integral component of plasma membrane	230
Golgi apparatus	212
integral component of thylakoid membrane	210
viral capsid	206

SI Appendix Tab. S4: Function trait : symbiosis

Number of transcriptomes corresponding to symbiont species
12

Number of CC composed of symbiont species sequences	Unannotated (Interpro)
90 794	52 491

Number of CC composed of N transcriptomes	
N	# CC
1	47 272
2	32 276
3	5 800
4	2 420
5	1 679
6	1 158
7	187
8	2
9	0
10	0
11	0
12	0

Most occuring MF sequence annotations for symbiotic CC	
annotations	occurencies
NA	172557
protein binding	15585
ion channel activity	4221
calcium ion binding	3935
protein kinase activity	3450
nucleotide binding	2921
catalytic activity	2887
nucleic acid binding	1888
metal ion binding	1569
binding	1565
oxidoreductase activity	1372
RNA binding	1195
zinc ion binding	947
ATP binding	932
DNA binding	928
hydrolase activity	917
transporter activity	820
hydrolase activity, hydrolyzing O-glycosyl compounds	751
iron ion binding	694
microtubule motor activity	569
NAD+ ADP-ribosyltransferase activity	533
structural constituent of ribosome	446
GTP binding	427
cysteine-type peptidase activity	421
methyltransferase activity	414
aspartic-type endopeptidase activity	396
ubiquitin-protein transferase activity	376
ATPase activity, coupled to transmembrane movement of substances	371
transferase activity, transferring phosphorus-containing groups	354
serine-type endopeptidase activity	331

Most occuring BP sequenceannotations for symbiotic CC	
annotations	occurencies
NA	205382
ion transport	4131
protein phosphorylation	3498
oxidation-reduction process	2250
proteolysis	2103
metabolic process	1987
transmembrane transport	1604
carbohydrate metabolic process	1242
transport	1242
signal transduction	838
lipid metabolic process	552
microtubule-based movement	477
translation	459
photosynthesis, light harvesting	356
cell redox homeostasis	325
protein folding	298
regulation of transcription, DNA-templated	290
protein ubiquitination	262
protein dephosphorylation	226
pseudouridine synthesis	222
carbohydrate transport	214
intracellular signal transduction	204
cation transport	200
photosynthesis	177
ubiquitin-dependent protein catabolic process	174
oxygen transport	171
potassium ion transport	170
dephosphorylation	161
ammonium transport	159
DNA replication	157

Most occuring CC sequenceannotations for symbiotic CC	
annotations	occurencies
NA	221346
membrane	7359
integral component of membrane	3744
intracellular	621
extracellular region	433
nucleus	375
Golgi membrane	222
cytoplasm	214
ribosome	132
cAMP-dependent protein kinase complex	128
myosin complex	70
viral capsid	67
nucleosome	65
1,3-beta-D-glucan synthase complex	64
nuclear pore	63
voltage-gated potassium channel complex	59
photosystem I	56
kinesin complex	54
cytochrome b6f complex	48
spliceosomal complex	48
extrinsic component of membrane	45
axonemal dynein complex	39
protein complex	32
endoplasmic reticulum	30
integral component of plasma membrane	29
origin recognition complex	29
plasma membrane	29
light-harvesting complex	26
dynactin complex	23
extracellular space	21

SI Appendix Tab. S5 : Function trait : mixotrophy

Number of transcriptomes corresponding to mixotrophic species
18

Number of CC composed of mixotrophic species sequences	Unannotated
75 099	43 888

Number of CC composed of N transcriptomes	
N	# CC
1	59 753
2	13 800
3	1 335
4	173
5	29
6	4
7	5
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0
17	0
18	0

Most occuring MF sequence annotations for mixotrophic CC	
annotations	occurencies
NA	128854
protein binding	10002
ion channel activity	3071
protein kinase activity	2913
calcium ion binding	2614
catalytic activity	2495
nucleotide binding	2184
nucleic acid binding	1277
oxidoreductase activity	1119
metal ion binding	1027
binding	978
RNA binding	907
ATP binding	796
hydrolase activity, hydrolyzing O-glycosyl compounds	780
DNA binding	634
microtubule motor activity	607
zinc ion binding	576
transporter activity	486
hydrolase activity	481
iron ion binding	435
methyltransferase activity	394
aspartic-type endopeptidase activity	374
NAD+ ADP-ribosyltransferase activity	361
serine-type endopeptidase activity	360
structural constituent of ribosome	307
ATPase activity, coupled to transmembrane movement of substances	272
cysteine-type peptidase activity	268
transferase activity, transferring phosphorus-containing groups	263
GTP binding	249
magnesium ion binding	207

Most occuring BP sequenceannotations for mixotrophic CC	
annotations	occurencies
NA	151493
ion transport	3026
protein phosphorylation	2935
oxidation-reduction process	1663
proteolysis	1650
metabolic process	1527
carbohydrate metabolic process	1014
transport	805
transmembrane transport	770
signal transduction	663
microtubule-based movement	479
translation	323
lipid metabolic process	319
protein glycosylation	269
photosynthesis, light harvesting	253
pseudouridine synthesis	206
photosynthesis	195
cell redox homeostasis	189
cation transport	179
protein folding	177
biosynthetic process	168
glycolytic process	163
ammonium transport	155
intracellular signal transduction	151
regulation of transcription, DNA-templated	148
ubiquitin-dependent protein catabolic process	145
DNA repair	133
protein dephosphorylation	121
transcription, DNA-templated	115
carbohydrate transport	111

Most occuring CC sequenceannotations for mixotrophic CC	
annotations	occurencies
NA	164331
membrane	5303
integral component of membrane	2437
intracellular	394
extracellular region	339
nucleus	271
cytoplasm	195
Golgi membrane	119
myosin complex	95
axonemal dynein complex	89
ribosome	87
cAMP-dependent protein kinase complex	78
1,3-beta-D-glucan synthase complex	50
extrinsic component of membrane	45
cytochrome b6f complex	43
photosystem I	40
kinesin complex	39
phosphopyruvate hydratase complex	33
voltage-gated potassium channel complex	29
clathrin coat of trans-Golgi network vesicle	28
integral component of plasma membrane	28
nucleosome	28
viral capsid	28
COPII vesicle coat	21
spliceosomal complex	21
condensin complex	19
dynactin complex	19
proton-transporting V-type ATPase, V0 domain	19
nuclear pore	18
chromosome	16

SI Appendix Tab. S6 : Function trait : harmful for human

Number of transcriptomes corresponding to harmful for human species
14

Number of CC composed of harmful for human species sequences	Unannotated
45 207	31 496

Number of CC composed of N transcriptomes	
N	# CC
1	31 857
2	11 493
3	1 453
4	316
5	70
6	13
7	5
8	0
9	0
10	0
11	0
12	0
13	0
14	0

Most occuring MF sequence annotations for harmful for human CC	
annotations	occurencies
NA	74215
protein binding	5151
ion channel activity	1690
catalytic activity	1638
protein kinase activity	1554
calcium ion binding	1491
nucleotide binding	1414
nucleic acid binding	877
oxidoreductase activity	719
hydrolase activity, hydrolyzing O-glycosyl compounds	667
binding	649
metal ion binding	533
RNA binding	521
ATP binding	492
DNA binding	389
transporter activity	344
microtubule motor activity	321
zinc ion binding	304
hydrolase activity	300
aspartic-type endopeptidase activity	273
iron ion binding	260
methyltransferase activity	233
serine-type endopeptidase activity	226
NAD+ ADP-ribosyltransferase activity	223
ATPase activity, coupled to transmembrane movement of substances	193
transferase activity, transferring hexosyl groups	187
sulfotransferase activity	178
cysteine-type peptidase activity	176
GTP binding	176
structural constituent of ribosome	158

Most occuring BP sequenceannotations for harmful for human CC	
annotations	occurencies
NA	87102
ion transport	1672
protein phosphorylation	1567
oxidation-reduction process	1133
metabolic process	1108
proteolysis	1038
carbohydrate metabolic process	786
transport	629
transmembrane transport	528
microtubule-based movement	279
signal transduction	264
protein glycosylation	212
lipid metabolic process	203
translation	168
pseudouridine synthesis	161
cation transport	120
photosynthesis, light harvesting	119
regulation of transcription, DNA-templated	113
biosynthetic process	112
cell redox homeostasis	107
ubiquitin-dependent protein catabolic process	103
photosynthesis	101
protein folding	91
glycolytic process	86
protein ubiquitination	85
intracellular signal transduction	83
protein dephosphorylation	82
transcription, DNA-templated	81
sulfate transport	80
carbohydrate transport	76

Most occuring CC sequenceannotations for harmful for human CC	
annotations	occurencies
NA	95216
membrane	3071
integral component of membrane	1646
intracellular	223
extracellular region	186
nucleus	145
cytoplasm	106
Golgi membrane	80
myosin complex	62
1,3-beta-D-glucan synthase complex	53
ribosome	50
cAMP-dependent protein kinase complex	41
axonemal dynein complex	34
cytochrome b6f complex	28
phosphopyruvate hydratase complex	26
extrinsic component of membrane	24
voltage-gated potassium channel complex	24
clathrin coat of trans-Golgi network vesicle	23
photosystem I	22
proton-transporting V-type ATPase, V0 domain	20
integral component of plasma membrane	17
integral component of thylakoid membrane	17
nucleosome	15
proton-transporting ATP synthase complex, coupling factor F(o)	14
protein phosphatase type 2A complex	13
light-harvesting complex	12
outer dynein arm	12
proton-transporting ATP synthase complex, catalytic core F(1)	12
chromosome	11
kinesin complex	11

SI Appendix Tab. S7 : Function trait : DMSP

Number of transcriptomes corresponding to DMSP species
25

Number of CC composed of DMSP species sequences	Unannotated
143 849	99 906

Number of CC composed of N transcriptomes	
N	# CC
1	79 493
2	50 652
3	7 397
4	2 779
5	1 824
6	1 271
7	320
8	102
9	6
10	1
11	3
12	1
13	0
14	0
15	0
16	0
17	0
18	0
19	0
20	0
21	0
22	0
23	0
24	0
25	0

Most occuring MF sequence annotations for DMSP CC	
annotations	occurencies
NA	258670
protein binding	22553
ion channel activity	5891
calcium ion binding	5744
protein kinase activity	5692
catalytic activity	5271
nucleotide binding	4578
nucleic acid binding	3094
oxidoreductase activity	2468
metal ion binding	2375
binding	2352
RNA binding	1898
ATP binding	1769
DNA binding	1535
zinc ion binding	1414
hydrolase activity	1393
hydrolase activity, hydrolyzing O-glycosyl compounds	1359
transporter activity	1352
microtubule motor activity	1330
iron ion binding	1140
NAD+ ADP-ribosyltransferase activity	818
structural constituent of ribosome	810
methyltransferase activity	775
GTP binding	702
ATPase activity, coupled to transmembrane movement of substances	636
aspartic-type endopeptidase activity	624
cysteine-type peptidase activity	615
transferase activity, transferring phosphorus-containing groups	579
serine-type endopeptidase activity	564
ubiquitin-protein transferase activity	558

Most occuring BP sequenceannotations for DMSP CC	
annotations	occurencies
NA	308710
protein phosphorylation	5756
ion transport	5716
oxidation-reduction process	3971
metabolic process	3470
proteolysis	3186
transmembrane transport	2391
carbohydrate metabolic process	2183
transport	2016
signal transduction	1150
microtubule-based movement	1084
lipid metabolic process	849
translation	839
photosynthesis, light harvesting	542
protein folding	532
cell redox homeostasis	457
regulation of transcription, DNA-templated	441
protein ubiquitination	375
protein glycosylation	374
pseudouridine synthesis	359
protein dephosphorylation	339
cation transport	332
photosynthesis	329
carbohydrate transport	313
intracellular signal transduction	303
ubiquitin-dependent protein catabolic process	301
biosynthetic process	300
DNA replication	270
oxygen transport	251
glycolytic process	241

Most occuring CC sequenceannotations for DMSP CC	
annotations	occurencies
NA	336097
membrane	10764
integral component of membrane	5823
intracellular	1048
nucleus	642
extracellular region	618
cytoplasm	434
Golgi membrane	328
ribosome	223
axonemal dynein complex	188
cAMP-dependent protein kinase complex	177
myosin complex	132
1,3-beta-D-glucan synthase complex	118
photosystem I	88
nucleosome	80
nuclear pore	78
viral capsid	78
extrinsic component of membrane	76
spliceosomal complex	73
voltage-gated potassium channel complex	70
cytochrome b6f complex	66
kinesin complex	61
phosphopyruvate hydratase complex	57
integral component of plasma membrane	54
endoplasmic reticulum	48
plasma membrane	38
protein complex	35
endoplasmic reticulum membrane	33
integral component of peroxisomal membrane	33
light-harvesting complex	33

SI Appendix Tab. S8 : Function trait : kleptoplasty

Number of transcriptomes corresponding to kleptoplastic species
2

Number of CC composed of kleptoplastic species sequences	Unannotated
8 357	5 948

Number of CC composed of N transcriptomes	
N	# CC
1	8 295
2	62

Most occuring MF sequence annotations for kleptoplastic CC	
annotations	occurencies
NA	13433
protein binding	763
protein kinase activity	360
calcium ion binding	318
catalytic activity	313
ion channel activity	287
nucleotide binding	245
nucleic acid binding	140
metal ion binding	121
hydrolase activity, hydrolyzing O-glycosyl compounds	111
oxidoreductase activity	104
ATP binding	93
binding	78
RNA binding	70
zinc ion binding	69
transporter activity	61
iron ion binding	60
aspartic-type endopeptidase activity	56
DNA binding	56
structural constituent of ribosome	45
cysteine-type peptidase activity	41
NAD+ ADP-ribosyltransferase activity	40
methyltransferase activity	37
microtubule motor activity	36
serine-type endopeptidase activity	33
hydrolase activity	30
magnesium ion binding	30
GTP binding	29
transferase activity, transferring glycosyl groups	27
sugar:proton symporter activity	24

Most occuring BP sequenceannotations for kleptoplastic CC	
annotations	occurencies
NA	15521
protein phosphorylation	362
ion transport	278
oxidation-reduction process	204
metabolic process	198
proteolysis	171
carbohydrate metabolic process	150
transmembrane transport	109
transport	76
signal transduction	56
translation	49
DNA replication	42
microtubule-based movement	34
glycolytic process	26
tricarboxylic acid cycle	26
carbohydrate transport	24
lipid metabolic process	22
potassium ion transport	20
protein folding	20
protein glycosylation	19
ATP hydrolysis coupled proton transport	18
cation transport	16
intracellular signal transduction	16
oxygen transport	16
DNA repair	15
immune response	15
sulfate transport	15
protein dephosphorylation	14
regulation of transcription, DNA-templated	14
tRNA aminoacylation for protein translation	14

Most occuring CC sequenceannotations for kleptoplastic CC	
annotations	occurencies
NA	17006
membrane	493
integral component of membrane	238
intracellular	51
extracellular region	43
nucleus	43
cytoplasm	30
Golgi membrane	24
ribosome	10
proton-transporting V-type ATPase, V0 domain	8
phosphopyruvate hydratase complex	7
condensin complex	6
myosin complex	6
proton-transporting two-sector ATPase complex, catalytic domain	6
cAMP-dependent protein kinase complex	5
voltage-gated potassium channel complex	5
1,3-beta-D-glucan synthase complex	4
clathrin coat of trans-Golgi network vesicle	4
COPII vesicle coat	4
eukaryotic translation initiation factor 3 complex	4
extrinsic component of membrane	4
Golgi apparatus	4
intracellular membrane-bounded organelle	4
kinesin complex	4
mitochondrial outer membrane	4
nucleosome	4
outer dynein arm	4
retromer complex	4
integral component of peroxisomal membrane	3
checkpoint clamp complex	2

SI Appendix Tab. S9 : Function trait : theca

Number of transcriptomes corresponding to thecated species
28

Number of CC composed of thecated species sequences	Unannotated
182 604	102 372

Number of CC composed of N transcriptomes	
N	# CC
1	104 081
2	59 346
3	11 398
4	4 246
5	2 067
6	1 218
7	127
8	66
9	32
10	11
11	2
12	7
13	2
14	0
15	1
16	0
17	0
18	0
19	0
20	0
21	0
22	0
23	0
24	0
25	0
26	0
27	0
28	0

Most occuring MF sequence annotations for thecated CC	
annotations	occurencies
NA	331947
protein binding	27766
catalytic activity	7672
protein kinase activity	7638
ion channel activity	6950
calcium ion binding	6765
nucleotide binding	5637
nucleic acid binding	3889
oxidoreductase activity	3485
binding	3092
metal ion binding	2833
ATP binding	2592
RNA binding	2570
hydrolase activity, hydrolyzing O-glycosyl compounds	2117
DNA binding	1988
microtubule motor activity	1908
transporter activity	1838
hydrolase activity	1737
zinc ion binding	1632
iron ion binding	1566
NAD+ ADP-ribosyltransferase activity	1014
methyltransferase activity	1009
structural constituent of ribosome	1005
GTP binding	844
ATPase activity, coupled to transmembrane movement of substances	834
serine-type endopeptidase activity	823
transferase activity, transferring phosphorus-containing groups	803
cysteine-type peptidase activity	801
aspartic-type endopeptidase activity	752
ubiquitin-protein transferase activity	717

Most occuring BP sequenceannotations for thecated CC	
annotations	occurencies
NA	395308
protein phosphorylation	7771
ion transport	6765
oxidation-reduction process	5652
metabolic process	4750
proteolysis	4301
carbohydrate metabolic process	3008
transmembrane transport	2919
transport	2661
microtubule-based movement	1549
signal transduction	1422
translation	1042
lipid metabolic process	1001
regulation of transcription, DNA-templated	754
protein folding	690
photosynthesis, light harvesting	659
pseudouridine synthesis	607
cell redox homeostasis	546
protein glycosylation	494
biosynthetic process	485
protein ubiquitination	477
cation transport	464
ubiquitin-dependent protein catabolic process	443
protein dephosphorylation	428
intracellular signal transduction	423
glycolytic process	417
carbohydrate transport	392
photosynthesis	391
ammonium transport	379
DNA replication	378

Most occuring CC sequenceannotations for thecated CC	
annotations	occurencies
NA	432969
membrane	13420
integral component of membrane	7629
intracellular	1304
nucleus	898
extracellular region	777
cytoplasm	571
Golgi membrane	404
ribosome	299
mitochondrion	279
axonemal dynein complex	266
1,3-beta-D-glucan synthase complex	239
myosin complex	221
cAMP-dependent protein kinase complex	207
cytochrome b6f complex	106
viral capsid	103
photosystem I	101
phosphopyruvate hydratase complex	100
kinesin complex	99
nucleosome	90
extrinsic component of membrane	85
voltage-gated potassium channel complex	84
integral component of plasma membrane	75
spliceosomal complex	73
nuclear pore	70
endoplasmic reticulum	62
proton-transporting V-type ATPase, V0 domain	47
membrane coat	46
Golgi apparatus	41
clathrin coat of trans-Golgi network vesicle	40

SI Appendix Tab. S10 : Function trait : ichyotoxin

Number of transcriptomes corresponding to ichyotoxic species
3

Number of CC composed of Ichyotoxic species sequences	Unannotated
5 164	3 404

Number of CC composed of N transcriptomes	
N	# CC
1	5 161
2	3
3	0

Most occuring MF sequence annotations for ichyotoxic CC	
annotations	occurencies
NA	8298
protein binding	731
ion channel activity	256
protein kinase activity	249
calcium ion binding	217
nucleotide binding	217
catalytic activity	144
nucleic acid binding	118
metal ion binding	77
binding	70
oxidoreductase activity	62
zinc ion binding	57
RNA binding	48
structural constituent of ribosome	46
ATP binding	45
DNA binding	39
aspartic-type endopeptidase activity	34
iron ion binding	32
microtubule motor activity	31
cysteine-type peptidase activity	28
transporter activity	26
3',5'-cyclic-nucleotide phosphodiesterase activity	25
hydrolase activity	24
methyltransferase activity	24
NAD+ ADP-ribosyltransferase activity	24
transferase activity, transferring phosphorus-containing grou	23
metalloendopeptidase activity	20
phosphatidylinositol binding	18
sialyltransferase activity	18
motor activity	17

Most occuring BP sequenceannotations for ichyotoxic CC	
annotations	occurencies
NA	9924
ion transport	258
protein phosphorylation	248
proteolysis	121
oxidation-reduction process	120
metabolic process	92
photosynthesis, light harvesting	72
signal transduction	72
translation	48
transport	45
transmembrane transport	41
photosynthesis	27
protein glycosylation	27
carbohydrate metabolic process	25
cell redox homeostasis	24
lipid metabolic process	23
protein folding	23
microtubule-based movement	19
intracellular signal transduction	15
regulation of transcription, DNA-templated	14
calcium ion transport	13
potassium ion transport	13
protein ADP-ribosylation	12
phosphatidylinositol metabolic process	11
protein dephosphorylation	11
protein peptidyl-prolyl isomerization	11
regulation of protein phosphorylation	11
cilium movement	10
oxygen transport	10
biosynthetic process	9

Most occuring CC sequenceannotations for ichyotoxic CC	
annotations	occurencies
NA	10809
membrane	478
integral component of membrane	123
intracellular	50
cytoplasm	20
nucleus	20
myosin complex	17
photosystem I	13
axonemal dynein complex	12
ribosome	12
cAMP-dependent protein kinase complex	11
cytochrome b6f complex	9
Golgi membrane	8
voltage-gated potassium channel complex	8
nucleosome	7
nuclear pore	6
plasma membrane	6
extracellular region	5
1,3-beta-D-glucan synthase complex	4
extrinsic component of membrane	4
integral component of plasma membrane	4
large ribosomal subunit	4
phosphopyruvate hydratase complex	4
U1 snRNP	4
condensin complex	3
endoplasmic reticulum membrane	3
small ribosomal subunit	3
COPI vesicle coat	2
endoplasmic reticulum	2
eukaryotic translation initiation factor 3 complex	2

SI Appendix Tab. S11 : Function trait : parasitism

Number of transcriptomes corresponding to parasitic species
1

Number of CC composed of parasitic species sequences	Unannotated
826	640

Number of CC composed of N transcriptomes	
N	# CC
	826

Most occuring MF sequence annotations for parasitic CC	
annotations	occurencies
NA	1417
protein binding	69
catalytic activity	39
protein kinase activity	21
calcium ion binding	20
microtubule motor activity	18
DNA binding	14
ion channel activity	14
zinc ion binding	11
ATP binding	9
GTPase activity	8
nucleic acid binding	8
oxidoreductase activity	8
RNA binding	8
metal ion binding	7
DNA photolyase activity	6
binding	4
hydrolase activity	4
iron ion binding	4
nucleotide binding	4
transferase activity, transferring hexosyl groups	4
cAMP-dependent protein kinase regulator activity	3
3-hydroxyanthranilate 3,4-dioxygenase activity	2
acyl-CoA dehydrogenase activity	2
AMP deaminase activity	2
ATPase activity, coupled to transmembrane movement	2
calcium-activated potassium channel activity	2
carbonate dehydratase activity	2
carbon-nitrogen ligase activity, with glutamine as amid	2
chromatin binding	2

Most occuring BP sequenceannotations for parasitic CC	
annotations	occurencies
NA	1587
protein phosphorylation	21
microtubule-based movement	18
ion transport	16
metabolic process	16
oxidation-reduction process	15
DNA repair	8
carbohydrate metabolic process	7
cellular protein modification process	6
DNA replication	6
proteolysis	6
transport	6
peroxisome fission	4
regulation of protein phosphorylation	3
response to stress	3
autophagy	2
biosynthetic process	2
dephosphorylation	2
DNA replication, synthesis of RNA primer	2
glutamate catabolic process to 2-oxoglutarate	2
guanine catabolic process	2
IMP salvage	2
lipid metabolic process	2
magnesium ion transport	2
molybdate ion transport	2
nucleoside metabolic process	2
peptidyl-amino acid modification	2
potassium ion transport	2
protein dephosphorylation	2
protein ubiquitination	2

Most occuring CC sequenceannotations for parasitic CC	
annotations	occurencies
NA	1702
membrane	24
integral component of membrane	20
integral component of peroxisomal membrane	4
cAMP-dependent protein kinase complex	3
cohesin core heterodimer	2
cytoplasm	2
nucleosome	2
nucleus	2
intracellular	1

SI Appendix Tab. S12 : Function trait : cyst formation

Number of transcriptomes corresponding to cyst foming species
16

Number of CC composed of cyst foming species sequences	Unannotated
76 761	45 160

Number of CC composed of N transcriptomes	
N	# CC
1	57 350
2	16 222
3	2 414
4	559
5	166
6	33
7	11
8	5
9	0
10	1
11	0
12	0
13	0
14	0
15	0
16	0

Most occuring MF sequence annotations for cyst foming CC	
annotations	occurencies
NA	136072
protein binding	9406
protein kinase activity	2664
calcium ion binding	2576
ion channel activity	2475
catalytic activity	2447
nucleotide binding	2220
nucleic acid binding	1599
oxidoreductase activity	1195
metal ion binding	1184
binding	1093
RNA binding	1045
hydrolase activity, hydrolyzing O-glycosyl compounds	871
ATP binding	829
DNA binding	728
zinc ion binding	599
hydrolase activity	583
transporter activity	511
microtubule motor activity	467
iron ion binding	460
NAD+ ADP-ribosyltransferase activity	411
aspartic-type endopeptidase activity	408
structural constituent of ribosome	378
methyltransferase activity	358
serine-type endopeptidase activity	358
transferase activity, transferring glycosyl groups	312
magnesium ion binding	296
ATPase activity, coupled to transmembrane movement of substances	290
cysteine-type peptidase activity	286
transferase activity, transferring phosphorus-containing groups	286

Most occuring BP sequenceannotations for cyst foming CC	
annotations	occurencies
NA	158919
protein phosphorylation	2720
ion transport	2440
oxidation-reduction process	1924
metabolic process	1651
proteolysis	1635
carbohydrate metabolic process	1157
transport	890
transmembrane transport	850
signal transduction	463
regulation of transcription, DNA-templated	445
translation	403
microtubule-based movement	402
lipid metabolic process	374
protein glycosylation	296
pseudouridine synthesis	292
photosynthesis, light harvesting	271
protein folding	255
photosynthesis	207
glycolytic process	189
protein ubiquitination	188
cation transport	174
cell redox homeostasis	171
intracellular signal transduction	165
ubiquitin-dependent protein catabolic process	164
protein dephosphorylation	154
transcription, DNA-templated	145
biosynthetic process	139
DNA repair	132
carbohydrate transport	116

Most occuring CC sequenceannotations for cyst foming CC	
annotations	occurencies
NA	172348
membrane	4678
integral component of membrane	2558
intracellular	451
nucleus	320
extracellular region	298
mitochondrion	264
cytoplasm	177
Golgi membrane	124
ribosome	108
myosin complex	87
1,3-beta-D-glucan synthase complex	80
phosphopyruvate hydratase complex	59
photosystem I	56
cAMP-dependent protein kinase complex	52
axonemal dynein complex	48
extrinsic component of membrane	46
cytochrome b6f complex	42
integral component of plasma membrane	33
nucleosome	31
spliceosomal complex	28
proton-transporting ATP synthase complex, coupling factor F(o)	22
kinesin complex	21
proton-transporting V-type ATPase, V0 domain	21
voltage-gated potassium channel complex	21
plasma membrane	20
chromosome	18
clathrin coat of trans-Golgi network vesicle	18
Golgi apparatus	18
viral capsid	18

SI Appendix Tab. S13 : comparison of PKS (literature) vs. harmful for human trait-CCs

	Number of sequences
From harmful for human CC	101 513
PKS from Kohli et al. 2015	2 632
From non-harmful for human CC	1 174 398

	Number of alignments of PKS (Kohli) vs. H4h CC sequences	Number of unique sequences	Number of CC/sequences
A : evalue 1e-25	298 751	NA	NA
B : (A) + pident >= 80	1 024	505	233/523
C : (B) + qcov >= 80 / scov >= 80	48	36	17/45

		Alignments from (C)											
	qseqid (source: Kohli et al. 2015)	sseqid	pident	length	mismatch	gapopen	qstart	qend	sstart	send	evalue	bitscore	qlen	slen
1	a_andernossi_CAMNT_0043413087_frame1	Alexandrium-andersonii-CCMP2222_TRINITY_DN57913_c15_g1_i1|m.102832	100.00	390	0	0	1	390	1	390	0.0	807	395	397
2	a_andernossi_CAMNT_0043419317_frame1	Alexandrium-andersonii-CCMP2222_TRINITY_DN51432_c1_g1_i1|m.69300	98.59	569	0	1	1	569	2	562	0.0	1138	667	563
3	alex_catgp4_ccmp1598_GAJB01000547_frame2r_dino_ks_full	Alexandrium-tamarense_TRINITY_DN47904_c0_g1_i1|m.76925	95.62	1142	50	0	168	1309	1	1142	0.0	2290	1349	1148
4	alex_catgp4_ccmp1598_GAJB01000695_frame3_dino_ks_full	Alexandrium-tamarense_TRINITY_DN100837_c0_g1_i1|m.270703	93.71	1208	76	0	36	1243	1	1208	0.0	2358	1282	1209
5	alex_fundy_38-3_gtm253_GAIV01007556_frame1_ks_full	Alexandrium-tamarense_TRINITY_DN100837_c0_g1_i1|m.270703	94.45	1208	67	0	37	1244	1	1208	0.0	2376	1275	1209
6	alex_fundy_38-3_gtm253_GAIV01065921_frame3_ks_full	Alexandrium-tamarense_TRINITY_DN47904_c0_g1_i1|m.76925	95.82	1147	48	0	162	1308	1	1147	0.0	2295	1335	1148
7	Alexandrium_monilatum_CAMNT_0046713273_frame2r	Alexandrium-monilatum-JR08_TRINITY_DN40765_c0_g1_i1|m.253782	100.00	1142	0	0	1	1142	113	1254	0.0	2351	1166	1255
8	atsp1b_gi|509914460|gb|GAIT01091826.1|_frame1_dino_ks_full	Alexandrium-tamarense_TRINITY_DN47904_c0_g1_i1|m.76925	99.74	1147	3	0	162	1308	1	1147	0.0	2383	1330	1148
9	atsp1b_gi|509914460|gb|GAIT01091826.1|_frame1_dino_ks_full	Alexandrium-monilatum-JR08_TRINITY_DN11248_c0_g1_i1|m.29052	80.37	1177	230	1	127	1303	104	1279	0.0	1992	1330	1281
10	atsp1b_gi|509962767|gb|GAIT01052927.1|_frame3_dino_ks_full	Alexandrium-tamarense_TRINITY_DN100837_c0_g1_i1|m.270703	99.54	1096	5	0	1	1096	113	1208	0.0	2272	1123	1209
11	Alexandrium_tema_CAMNT_0028940359_frame1R_dino_ks_full	Alexandrium-tamarense_TRINITY_DN47904_c0_g1_i1|m.76925	100.00	1147	0	0	102	1248	1	1147	0.0	2386	1255	1148
12	Alexandrium_tema_CAMNT_0028940359_frame1R_dino_ks_full	Alexandrium-monilatum-JR08_TRINITY_DN11248_c0_g1_i1|m.29052	80.13	1243	246	1	1	1243	38	1279	0.0	2102	1255	1281
13	Alexandrium_tema_CAMNT_0028987857_frame3r_dino_ks_full	Alexandrium-tamarense_TRINITY_DN100837_c0_g1_i1|m.270703	100.00	1056	0	0	1	1056	153	1208	0.0	2199	1100	1209
14	azadinium_CAMNT_0029612195_frame2_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN13411_c0_g1_i1|m.27715	100.00	1060	0	0	167	1226	1	1060	0.0	2221	1254	1061
15	azadinium_CAMNT_0029613309_frame3r_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN49854_c1_g1_i1|m.171584	91.76	1226	93	3	1	1225	6	1224	0.0	2336	1353	1242
16	azadinium_CAMNT_0029616779_frame1R_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN49895_c3_g2_i1|m.172155	82.93	1230	210	0	3	1232	2	1231	0.0	2144	1283	1244
17	azadinium_CAMNT_0029619413_frame2_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN76750_c0_g1_i1|m.239907	100.00	1186	0	0	14	1199	1	1186	0.0	2484	1211	1187
18	azadinium_CAMNT_0029646331_frame3_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50771_c6_g1_i1|m.188119	85.09	1100	155	4	13	1104	1	1099	0.0	1951	1281	1100
19	azadinium_CAMNT_0029646331_frame3_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50771_c6_g1_i2|m.188124	85.37	1094	157	3	13	1104	1	1093	0.0	1950	1281	1094
20	azadinium_CAMNT_0029646683_frame1_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50843_c0_g1_i3|m.188557	96.83	1261	40	0	38	1298	1	1261	0.0	2564	1298	1279
21	azadinium_CAMNT_0029646683_frame1_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50843_c0_g1_i1|m.188549	96.11	1261	49	0	38	1298	1	1261	0.0	2543	1298	1279
22	azadinium_CAMNT_0029646683_frame1_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50843_c0_g1_i5|m.188564	94.05	1261	75	0	38	1298	1	1261	0.0	2507	1298	1279
23	azadinium_CAMNT_0029646683_frame1_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50843_c0_g1_i2|m.188553	93.02	1261	88	0	38	1298	1	1261	0.0	2484	1298	1279
24	azadinium_CAMNT_0029647659_frame2_sl_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50938_c7_g1_i7|m.192097	88.89	945	91	4	17	947	3	947	0.0	1736	989	947
25	azadinium_CAMNT_0029647659_frame2_sl_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50938_c7_g1_i6|m.192092	87.13	948	105	5	17	947	3	950	0.0	1710	989	950
26	azadinium_CAMNT_0029647659_frame2_sl_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50938_c7_g1_i4|m.192084	85.29	945	125	4	17	947	3	947	0.0	1677	989	947
27	azadinium_CAMNT_0029647659_frame2_sl_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50938_c7_g1_i2|m.192075	83.65	948	138	5	17	947	3	950	0.0	1651	989	950
28	azadinium_CAMNT_0029648695_frame3_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50537_c7_g1_i1|m.182355	95.08	1362	65	2	31	1391	1	1361	0.0	2707	1438	1362
29	azadinium_CAMNT_0029648695_frame3_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50537_c7_g1_i2|m.182362	91.34	1362	116	2	31	1391	1	1361	0.0	2609	1438	1362
30	azadinium_CAMNT_0029664981_frame3r_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN49895_c3_g1_i1|m.172152	89.20	1102	113	2	132	1227	1	1102	0.0	2049	1347	1108
31	azadinium_CAMNT_0029680587_frame1R_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN16462_c0_g1_i1|m.36933	100.00	1201	0	0	20	1220	1	1201	0.0	2512	1239	1202
32	azadinium_CAMNT_0029681121_frame1R_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50592_c0_g3_i1|m.184170	95.02	943	47	0	1	943	39	981	0.0	1874	956	982
33	azadinium_CAMNT_0029681121_frame1R_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50592_c0_g1_i1|m.184163	86.95	912	119	0	1	912	39	950	0.0	1685	956	987
34	azadinium_CAMNT_0029681149_frame2_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50197_c0_g1_i1|m.176682	83.80	1191	182	2	29	1218	41	1221	0.0	2104	1284	1233
35	azadinium_CAMNT_0029681821_frame3r_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN49854_c3_g1_i1|m.171630	100.00	1231	0	0	1	1231	3	1233	0.0	2572	1321	1234
36	azadinium_CAMNT_0029685029_frame1_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN48673_c0_g2_i1|m.156850	100.00	1231	0	0	13	1243	1	1231	0.0	2579	1308	1232
37	azadinium_CAMNT_0029696489_frame1_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50938_c7_g1_i4|m.192084	92.73	880	60	1	1	876	68	947	0.0	1715	1039	947
38	azadinium_CAMNT_0029696489_frame1_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50938_c7_g1_i2|m.192075	91.73	883	66	2	1	876	68	950	0.0	1704	1039	950
39	azadinium_CAMNT_0029696489_frame1_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50938_c7_g1_i7|m.192097	88.18	880	100	1	1	876	68	947	0.0	1646	1039	947
40	azadinium_CAMNT_0029696489_frame1_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50938_c7_g1_i6|m.192092	86.86	883	109	2	1	876	68	950	0.0	1627	1039	950
41	azadinium_CAMNT_0029728931_frame3_dino_ks_full	Azadinium-spinosum-3D9_TRINITY_DN50197_c1_g1_i1|m.176687	97.70	1220	18	2	30	1239	34	1253	0.0	2477	1299	1254
42	azadinium_Contig	Azadinium-spinosum-3D9_TRINITY_DN50197_c0_g1_i1|m.176682	81.04	1118	200	8	1	1109	50	1164	0.0	1876	1109	1233
43	azadinium_Contig	Azadinium-spinosum-3D9_TRINITY_DN49854_c1_g2_i2|m.171596	80.41	1225	226	5	7	1228	13	1226	0.0	2054	1301	1244
44	lingulodinium_CAMNT_0033633585_frame3	Lingulodinium-polyedra-CCMP1738_TRINITY_DN23214_c0_g1_i1|m.75936	100.00	1066	0	0	1	1066	22	1087	0.0	2209	1264	1088
45	lingulodinium_CAMNT_0033658863_frame1R	Lingulodinium-polyedra-CCMP1738_TRINITY_DN24924_c0_g1_i1|m.85208	100.00	835	0	0	7	841	1	835	0.0	1714	859	836
46	lingulodinium_CAMNT_0033678165_frame1R	Lingulodinium-polyedra-CCMP1738_TRINITY_DN34294_c0_g2_i1|m.150741	96.57	961	33	0	1	961	70	1030	0.0	1940	1054	1031
47	lingulodinium_CAMNT_0033716467_frame3	Lingulodinium-polyedra-CCMP1738_TRINITY_DN34294_c0_g1_i1|m.150737	96.70	1030	34	0	6	1035	5	1034	0.0	2076	1087	1035
48	Pyrodinium_baha_CAMNT_0020831499_frame2	Pyrodinium-bahamense_TRINITY_DN29200_c0_g1_i1|m.50525	99.27	827	1	1	41	862	2	828	0.0	1712	864	829

	harmful for human transcriptomes	blastp alignments	
	transcriptome_id	Number of alignments	
		Total in h4h CC	PKS hits from (B)	PKS hits from (C)
1	Alexandrium-andersonii-CCMP2222	14 963	148	2
2	Alexandrium-catenella	22 331	96	NA
3	Alexandrium-fundyense	1366	NA	NA
4	Alexandrium-minutumHALIM-CCMP113	6529	11	NA
5	Alexandrium-monilatum-JR08	43104	59	3
6	Alexandrium-tamarense	43820	79	8
7	Azadinium-spinosum-3D9	23507	177	30
8	Dinophysis-acuminata-DAEP01	18528	2	NA
9	Gambierdiscus-australes-CAWD149	17872	172	NA
10	Gymnodinium-catenatum	15325	2	NA
11	Karenia-brevis	NA	NA	NA
12	Lingulodinium-polyedra-CCMP1738	35581	88	4
13	Prorocentrum-lima-CCMP684	377	NA	NA
14	Prorocentrum-minimum-CCMP1329	16294	1	NA
15	Protoceratium-reticulatum	18668	87	NA
16	Pyrodinium-bahamense	29053	102	1

SI Appendix Tab. S13 : comparison of PKS (literature) vs. harmful for human trait-CCs

	Number of hits (pident80 / qcov80 / scov80)	Number of unique sequences	Number of CC/sequences
D : PKS hits of non-h4h specific CC sequences onto PKS database from Kohli	1835	646	165/1144

	PKS hits on non-h4h specific CC sequences		Kohli et al. Screening experience – sup table 4				
	transcriptome_id	Number of hits (pident80 / qcov80 / scov80)	Number of hits				
			KS domains full	KS domains partial	active site conserved	KR domains full	KR domains partial
1	Alexandrium-tamarense	293	83	17	54	7	1
2	Alexandrium-monilatum-JR08	156	65	12	37	6	1
3	plg	152					
4	Alexandrium-margalefi-CS	128	61	38	38	3	1
5	Symbiodinium-sp--CCMP421	124	62	11	39	6	1
6	Symbiodinium-sp--C15	93	25	15	5	3	
7	Symbiodinium-sp--Mp	93	39	6	17	3	
8	Symbiodinium-goreaui-typeC1	85	31	0	19	3	1
9	Alexandrium-catenella	80					
10	Symbiodinium-sp--CCMP2430	72	24	10	15	3	
11	Lingulodinium-polyedra-CCMP1738	63	70	20	43	8	1
12	Gambierdiscus-australes-CAWD149	59	90	4	58	7	4
13	Pyrodinium-bahamense	52	53	14	35	2	
14	Scrippsiella-hangoei	43	36	1	23	4	
15	Gonyaulax-spinifera-CCMP409	39	13	29	9	4	
16	Scrippsiella-trochoidea-CCMP-3099	37	39	1	27	4	2
17	Amphidinium-carterae-CCMP1314	28	32	6	18	3	
18	Azadinium-spinosum-3D9	28	119	21	86	5	1
19	Pelagodinium-beii	24	19	6	14	5	
20	Symbiodinium-trenchi-D1a	24	8	16	6	3	1
21	Alexandrium-andersonii-CCMP2222	20	21	32	13	2	1
22	Kryptoperidinium-foliaceum-CCMP1326	18					
23	Symbiodinium-sp--A3	18	16	10	9	2	
24	bdt	16					
25	Heterocapsa-arctica	13	28	12	16	4	1
26	het	12					
27	Heterocapsa-triquetra	11	35	17	21	4	1
28	Glenodinium-foliaceum	10	10	18	7	3	1
29	Scrippsiella-nutricula	9	31	18	18	4	1
30	Protoceratium-reticulatum	8					
31	Durinskia-baltica	7	26	12	17	5	
32	Polarella-glacialis	5	25	13	16	5	2
33	Crypthecodinium-cohnii	4					
34	Heterocapsa-rotundata	4	17	20	11	2	
35	Karlodinium-micrum	4	36	13	24	2	
36	Prorocentrum-minimum-CCMP1329	2	69	33	45	3	1
37	Gymnodinium-catenatum	1	8		5	3	
38	Dinophysis-acuminata-DAEP01	lost after filtering hits					
39	Amphidinium-massartii	lost after filtering hits					
40	Noctiluca-scintillans	lost after filtering hits					
41	Ceratium-fusus	lost after filtering hits	7	2	4	2	2
42	dsr	lost after filtering hits					
43	Togula-jolla-CCCM725	lost after filtering hits	7		4	2	
44	Alexandrium-minutumHALIM-CCMP113	lost after filtering hits					
45	Oxyrrhis-marina	lost after filtering hits					
46	Gambierdiscus-sp	lost after filtering hits					

Interpro_desc of CC for PKS hits of (B) onto PKS database from Kohli : interpro_desc	frequencies
Thiolase-like	45

Interpro_desc of CC for PKS hits of (D) onto PKS database from Kohli : interpro_desc	frequencies
Thiolase-like	1 113
Thiolase-like, subgroup	1
NA	30

SI Appendix Tab. S14 : comparison of STX (literature) vs. harmful for human trait-CCs

							Murray et al. 2015 – table 1							
	order	family	genus	specie	presency	harmful for humans	documented To produce STX	sxtA1 (EST hits)	sxtA4 (EST hits)	sxtG (EST hits)	SxtA1-4 vs h4h (hits with pident>90)	SxtA1-4 vs non-h4h (hits with pident>90)	SxtG vs h4h (hits with pident>90)	SxtG vs non-h4h (hits with pident>90)
1	Dinophysiales	Dinophysiaceae	Dinophysis	acuminata	yes	DSP	no	2						
2	Gonyaulacales	Ceratiaceae	Ceratium	fusus	yes	No	no					2		
3	Gonyaulacales	Crypthecodiniaceae	Crypthecodinium	cohnii	yes	No	no					1		
4	Gonyaulacales	Goniodomataceae	Gambierdiscus	australes	yes	CFP	no	4						
5	Gonyaulacales	Goniodomataceae	Pyrodinium	bahamense	yes	PSP	yes	1	1	3		4		1
6	Gonyaulacales	Gonyaulacaceae	Alexandrium	andersonii	yes	PSP	no	1		1		1		
7	Gonyaulacales	Gonyaulacaceae	Alexandrium	catenella	yes	PSP								
8	Gonyaulacales	Gonyaulacaceae	Alexandrium	fundyense	no	PSP	yes	2	1	2				
9	Gonyaulacales	Gonyaulacaceae	Alexandrium	margalefi	yes	No	no							
10	Gonyaulacales	Gonyaulacaceae	Alexandrium	minutum	yes	PSP	yes	yes	yes					
11	Gonyaulacales	Gonyaulacaceae	Alexandrium	monilatum	yes	PSP	no	1				3		
12	Gonyaulacales	Gonyaulacaceae	Alexandrium	tamarense	yes	PSP	no	2				4		
13	Gonyaulacales	Gonyaulacaceae	Gonyaulax	spinifera	yes	No	no	1						
14	Gonyaulacales	Gonyaulacaceae	Lingulodinium	polyedra	yes	DSP	no	5				4		
15	Gonyaulacales	Gonyaulacaceae	Protoceratium	reticulatum	yes	DSP	no	1						
16	Gymnodiniales	Gymnodiniaceae	Akashiwo	sanguinea	no	No								
17	Gymnodiniales	Gymnodiniaceae	Amphidinium	carterae	yes	No	no					2		
18	Gymnodiniales	Gymnodiniaceae	Amphidinium	massartii	yes	No	no					1		
19	Gymnodiniales	Gymnodiniaceae	Gymnodinium	catenatum	yes	PSP	yes	3	5	2		2		1
20	Gymnodiniales	Gymnodiniaceae	Heterocapsa	sp. (add)	yes	No						5		
21	Gymnodiniales	Gymnodiniaceae	Gymnoxanthella	radiolariae (add)	yes	No						3		
22	Gymnodiniales	Gymnodiniaceae	Gyrodinium	dominans	no	No								
23	Gymnodiniales	Gymnodiniaceae	Togula	jolla	yes	No	no					1		
24	Gymnodiniales	Gymnodiniaceae	Karlodinium	micrum	yes	No	no	1				1		
25	Gymnodiniales	Kareniaceae	Karenia	brevis	no	NSP	no							
26	Noctilucales	Noctilucaceae	Noctiluca	scintillans	yes	No	no					1		
27	Oxyrrhinales	Oxyrrhinaceae	Oxyrrhis	marina	yes	No	no							
30	Peridiniales	Heterocapsaceae	Heterocapsa	arctica	yes	No	no	2				2		
31	Peridiniales	Heterocapsaceae	Heterocapsa	rotundata	yes	No	no	1				3		
32	Peridiniales	Heterocapsaceae	Heterocapsa	triquetra	yes	No	no	1				2		
33	Peridiniales	incertae sedis	Azadinium	spinosum	yes	AZP	no	1						
34	Peridiniales	Lessardiaceae	Lessardia	elongata	no	No								
35	Peridiniales	Peridiniaceae	Brandtodinium	nutricula	yes	No	no	2				2		
36	Peridiniales	Peridiniaceae	Brandtodinium	nutricula (add)	yes	No	no					4		
37	Peridiniales	Peridiniaceae	Durinskia	baltica	yes	No	no	2				4		
38	Peridiniales	Peridiniaceae	Glenodinium	foliaceum	yes	No	no					1		
39	Peridiniales	Peridiniaceae	Kryptoperidinium	foliaceum	yes	No	no	1				3		
40	Peridiniales	Peridiniaceae	Peridinium	aciculiferum	no	No	no	1						
41	Peridiniales	Peridiniaceae	Scrippsiella	hangoei	yes	No						1		
42	Peridiniales	Peridiniaceae	Scrippsiella	hangoei-like	no	No								
43	Peridiniales	Peridiniaceae	Scrippsiella	trochoidea	yes	No	no	1				3		
44	Prorocentrales	Prorocentraceae	Prorocentrum	lima	no	DSP	no							
45	Prorocentrales	Prorocentraceae	Prorocentrum	micans	no	No	no							
46	Prorocentrales	Prorocentraceae	Prorocentrum	minimum	yes	DSP	no	1				1		
47	Pyrocystales	Pyrocystaceae	Pyrocystis	lunula	no	No								
48	Suessiales	Suessiaceae	Pelagodinium	beii	yes	No	no	1				3		
49	Suessiales	Suessiaceae	Pelagodinium	beii (add)	yes	No	no					12		1
50	Suessiales	Suessiaceae	Polarella	glacialis	yes	No	no	2				3		
52	Suessiales	Symbiodiniaceae	Symbiodinium	sp.	yes	No	no	1				19		
59	Syndiniales	Amoebophryaceae	Amoebophrya	sp.	yes	No						1		
60	Thoracosphaerales	Thoracosphaeraceae	Thoracosphaera	heimii	no	No								
								42	7	7		99	0	3

SI Appendix Tab. S14 : comparison of STX (literature) vs. harmful for human trait-CCs

	Number of hits	Number of Unique sequences (in network)	Number of total CC/sequences in CC
sxtA1-4 hits onto h4h CC	358	n.d.	n.d.
sxtA1-4 hits onto h4h CC (pident>90)	0	0	0
sxtA1-4 hits onto non h4h CC	6 940	n.d.	n.d.
A : sxtA1-4 hits onto non h4h CC (pident>90)	394	99	20/166
sxtG hits onto h4h CC	8	a	n.d.
sxtG hits onto h4h CC (pident>90)	0	0	0
sxtG hits onto non h4h CC	251	n.d.	n.d.
B : sxtG hits onto non h4h CC (pident>90)	57	3	1/3

	interpro description	Number of Sequences	Notes
(A)	Pyridoxal phosphate-dependent transferase, Major region, Subdomain 1	110	sxtA protein of Alexandrium australiense contains domain that corresponds to PyrdxlP-dep_Trfase_major_sub1. (IPR015421) See : http://www.uniprot.org/uniprot/A0A0A1E788
	GNAT domain	22	Acyl-CoA N-acyltransferase (stxA2)
	Polyketide synthase, Phosphopantetheine-binding domain	15	prosthetic group of acyl carrier proteins (sxtA1)
	Acyl-CoA N-acyltransferase	5	Acyl-CoA N-acyltransferase (stxA2)
	S-adenosyl-L-methionine-dependent Methyltransferase	4	domain found in S-adenosyl-L-methionine-dependent Methyltransferases (SAM Mtases) (sxtA1)
	Death domain	2	cannot be related to sxtA domains
	DNA methylase, adenine-specific	2	cannot be related to sxtA domains
	Rieske [2Fe-2S] iron-sulphur domain	2	cannot be related to sxtA domains
	<NA>	4	
(B)	Amidinotransferase	1	(sxtG)
	<NA>	2	

		database	score	Interpro ID	Interpro description	Gene Ontology
12 hits with pident>90% to sxtA1-4 vs. non-h4h	plg_TRINITY_DN46666_c1_g2_i1|m.295389	Gene3D	1.9E-45	IPR015421	Pyridoxal phosphate-dependent transferase, major region, subdomain 1	GO:0003824|GO:0030170
		Pfam	1.8E-13	IPR004839	Aminotransferase, class I/classII	GO:0009058|GO:0030170
		SUPERFAMILY	1.08E-51	IPR015424	Pyridoxal phosphate-dependent transferase	NA
	plg_TRINITY_DN46666_c1_g2_i1|m.295389	Gene3D	3.1E-5	IPR009081	Acyl carrier protein-like	NA
		Pfam	4.2E-5	IPR009081	Acyl carrier protein-like	NA
		ProSiteProfiles	9 075	IPR009081	Acyl carrier protein-like	NA
		ProSiteProfiles	11.47	IPR000182	GNAT domain	GO:0008080
		SUPERFAMILY	6.91E-7	IPR009081	Acyl carrier protein-like	NA
		SUPERFAMILY	6.5E-10	IPR016181	Acyl-CoA N-acyltransferase	NA

SI Appendix Tab. S15 : comparison of symbiotic genes (literature) vs. symbiosis trait-CCs

	Number of hits	Number of Unique sequences (in network)	Number of Corresponding CC	Min CC size	Median CC size	Mean CC size	Max CC size
A1 : presumed symbiotic proteins hits onto symb CC	12 394	1 757	688	2,00	2,00	2,87	14,00
A2 : presumed symbiotic proteins hits onto symb CC (pident>60)	8	8	5	2,00	2,00	2,80	6,00
B1 : presumed symbiotic proteins hits onto _no_symb CC	46 913	8 324	1 363	2,00	3,00	5,99	464,00
B2 : presumed symbiotic proteins hits onto no_symb CC (pident>60)	107	71	21	2,00	3,00	10,10	64,00

	gene_name	found in A1	found in A2	found in B1	found in B2
1	P-type H+-ATPase	x		x	x
2	nodulation protein U	x		x	
3	nodulation protein nolO	x	x	x	
4	Phosphoadenosine phosphosulfate reductase	x	x	x	x
5	calumenin	x		x	
6	Phosphatidylinositol-3-phosphate kinase	x		x	
7	Host cell factor 2	x		x	
8	Fibronectin type III domain containing protein 3C1	x		x	
9	Spondin-1	x		x	
10	Merozoite surface protein 1	x	x	x	x
11	Avirulence protein AvrBs3	x		x	
12	Putative surface-exposed virulence protein BigA	x			
13	UDP-galactopyranose mutase				
14	Sporozoite developmental protein	x		x	
15	Mycocerosic acid synthase	x		x	
16	Transmembrane protein 151 homolog	x		x	
17	Chaoptin	x		x	
18	Uncharacterized MFS-type transporter ydgK	x		x	
19	Uncharacterized UDP-glucosyltransferase yojK	x		x	
20	Uncharacterized protein PF11_0213	x		x	
21	Protein dpy-19 homolog 1	x		x	
22	Superoxide dismutase	x		x	x
23	Catalase			x	x
24	Peroxiredoxin	x	x	x	x
25	Glutathione peroxidase	x		x	x
26	Glutathione reducase				
27	g-glutamylcysteine synthetase	x		x	x
28	Glutathione synthetase	x		x	
29	Glutathione S-transferase				
30	ferritin	x	x	x	
31	Solute carrier family 2, facilitated glucose transporter member 8	x		x	
32	Sodium/myo-inositol cotransporter 2	x		x	
33	Lipid storage droplets surface-binding protein 2				
34	Scavenger receptor class B member 1				
35	Organic cation transporter protein	x		x	
36	Epididymal secretory protein E1				
37	Sodium- and chloride-dependent taurine transporter	x		x	
38	Monocarboxylate transporter 10	x		x	
39	Vesicular inhibitory amino acid transporter	x		x	
40	Carbonic anhydrase 2	x		x	
41	Aquaporin-5	x		x	
42	Aquaporin-3	x		x	
43	Ammonium transporter Rh type B	x		x	
44	Putative ammonium transporter 1	x		x	

Blastp hits distribution						
genus	specie	symbiont	A1	A2	B1	B2
Brandtodinium	nutricula	Yes	301	3	485	17
Gymnoxanthella	radiolariae	Yes	186	0	232	5
Pelagodinium	beii	Yes	141	0	144	7
Symbiodinium	kawagutii	Yes			823	
Symbiodinium	sp.	Yes	1288	11		41
Akashiwo	sanguinea	No				
Alexandrium	andersonii	No			138	1
Alexandrium	catenella	No			194	4
Alexandrium	fundyense	No				
Alexandrium	margalefi	No			465	7
Alexandrium	minutum	No			37	
Alexandrium	monilatum	No			496	
Alexandrium	tamarense	No			528	6
Amoebophrya	sp.	No			15	
Amphidinium	carterae	No			209	5
Amphidinium	massartii	No			178	6
Azadinium	spinosum	No			280	1
Ceratium	fusus	No			268	4
Crypthecodinium	cohnii	No			205	1
Dinophysis	acuminata	No			183	3
Durinskia	baltica	No			299	7
Gambierdiscus	australes	No			207	4
Glenodinium	foliaceum	No			338	10
Gonyaulax	spinifera	No			142	3
Gymnodinium	catenatum	No			119	3
Gyrodinium	dominans	No				
Heterocapsa	sp.	No			530	14
Heterocapsa	arctica	No			134	2
Heterocapsa	rotundata	No			130	2
Heterocapsa	triquetra	No			198	3
Karenia	brevis	No				
Karlodinium	micrum	No			239	5
Kryptoperidinium	foliaceum	No			612	23
Lessardia	elongata	No				
Lingulodinium	polyedra	No			424	2
Noctiluca	scintillans	No			127	3
Oxyrrhis	marina	No			13	
Peridinium	aciculiferum	No				
Polarella	glacialis	No			158	1
Prorocentrum	lima	No				
Prorocentrum	micans	No				
Prorocentrum	minimum	No			122	5
Protoceratium	reticulatum	No			215	
Pyrocystis	lunula	No				
Pyrodinium	bahamense	No			300	2
Scrippsiella	hangoei	No			482	4
Scrippsiella	hangoei-like	No				
Scrippsiella	trochoidea	No			419	5
Thoracosphaera	heimii	No			1106	
Togula	jolla	No			105	2

SI Appendix Tab. S16 : exploring the 189 symbiosis trait-CCs that involve 7 or 8 dinoflagellates species

Goslim annotations found in 189 symbiosis associated CC composed Of at least 7 transcriptomes of symbiotic species				
Annotations	Goslim level	Freq	Global %	% over Annotated Sequences
NA	MF	1400	73,84 %	
protein binding	MF	96	5,06 %	19,35 %
nucleotide binding	MF	57	3,01 %	11,49 %
calcium ion binding	MF	53	2,80 %	10,69 %
nucleic acid binding	MF	43	2,27 %	8,67 %
aspartic-type endopeptidase activity	MF	29	1,53 %	5,85 %
phosphate ion transmembrane transporter activity	MF	28	1,48 %	5,65 %
cysteine-type peptidase activity	MF	26	1,37 %	5,24 %
oxidoreductase activity	MF	22	1,16 %	4,44 %
peptidase activity	MF	15	0,79 %	3,02 %
transporter activity	MF	12	0,63 %	2,42 %
iron ion binding	MF	10	0,53 %	2,02 %
heme binding	MF	9	0,47 %	1,81 %
metal ion transmembrane transporter activity	MF	9	0,47 %	1,81 %
microtubule motor activity	MF	9	0,47 %	1,81 %
ATP binding	MF	8	0,42 %	1,61 %
binding	MF	8	0,42 %	1,61 %
catalytic activity	MF	8	0,42 %	1,61 %
hydrolase activity, hydrolyzing O-glycosyl compounds	MF	8	0,42 %	1,61 %
magnesium ion binding	MF	8	0,42 %	1,61 %
metal ion binding	MF	8	0,42 %	1,61 %
RNA binding	MF	8	0,42 %	1,61 %
signal transducer activity	MF	8	0,42 %	1,61 %
zinc ion binding	MF	8	0,42 %	1,61 %
cAMP-dependent protein kinase regulator activity	MF	6	0,32 %	1,21 %
NA	BP	1602	84,49 %	
proteolysis	BP	70	3,69 %	23,81 %
transmembrane transport	BP	46	2,43 %	15,65 %
oxidation-reduction process	BP	32	1,69 %	10,88 %
phosphate ion transmembrane transport	BP	28	1,48 %	9,52 %
carbohydrate metabolic process	BP	18	0,95 %	6,12 %
photosynthesis	BP	13	0,69 %	4,42 %
cell redox homeostasis	BP	12	0,63 %	4,08 %
transport	BP	12	0,63 %	4,08 %
metal ion transport	BP	9	0,47 %	3,06 %
microtubule-based movement	BP	9	0,47 %	3,06 %
intracellular protein transport	BP	8	0,42 %	2,72 %
phosphorelay signal transduction system	BP	8	0,42 %	2,72 %
tRNA modification	BP	8	0,42 %	2,72 %
N-glycan processing	BP	7	0,37 %	2,38 %
RNA splicing	BP	7	0,37 %	2,38 %
regulation of protein phosphorylation	BP	6	0,32 %	2,04 %
pathogenesis	BP	1	0,05 %	0,34 %
NA	CC	1766	93,14 %	
integral component of membrane	CC	64	3,38 %	49,23 %
membrane	CC	32	1,69 %	24,62 %
photosystem I	CC	13	0,69 %	10,00 %
extracellular region	CC	8	0,42 %	6,15 %
spliceosomal complex	CC	7	0,37 %	5,38 %
cAMP-dependent protein kinase complex	CC	6	0,32 %	4,62 %

Transcriptome distribution : 187 CC of minimum 7 symbiotic transcriptomes		
Transcrptomes	Freq	%
Symbiodinium-sp--CCMP421	415	21,89 %
Symbiodinium-trenchi-D1a	301	15,88 %
Symbiodinium-goreaui-typeC1	255	13,45 %
Symbiodinium-sp--C15	237	12,50 %
Symbiodinium-sp--CCMP2430	233	12,29 %
Symbiodinium-sp--Mp	228	12,03 %
Pelagodinium-beii	211	11,13 %
Symbiodinium-sp--A3	10	0,53 %
dsr	5	0,26 %
bdt	1	0,05 %

Transcriptome distribution : 2 CC of minimum 8 symbiotic transcriptomes		
Transcrptomes	Freq	%
dsr	3	16,67 %
Symbiodinium-goreaui-typeC1	3	16,67 %
Pelagodinium-beii	2	11,11 %
Symbiodinium-sp--C15	2	11,11 %
Symbiodinium-sp--CCMP2430	2	11,11 %
Symbiodinium-sp--CCMP421	2	11,11 %
Symbiodinium-sp--Mp	2	11,11 %
Symbiodinium-trenchi-D1a	2	11,11 %

SI Appendix Tab. S17 : HSP70 in core CCs

Transcriptomes	HSP70	network presence	core transcriptome Status	cc_index
Symbiodinium-trenchi-D1a	25	yes	core	vBtLzyXFVpMb
dsr	19	yes	core	vBtLzyXFVpMb
Scrippsiella-hangoei	17	yes	core	vBtLzyXFVpMb
Alexandrium-tamarense	16	yes	core	vBtLzyXFVpMb
Karlodinium-micrum	14	yes	core	vBtLzyXFVpMb
Kryptoperidinium-foliaceum-CCMP1326	13	yes	core	vBtLzyXFVpMb
bdt	12	yes	core	vBtLzyXFVpMb
Symbiodinium-sp--C15	12	yes	core	vBtLzyXFVpMb
Durinskia-baltica	11	yes	core	vBtLzyXFVpMb
Pyrodinium-bahamense	11	yes	core	vBtLzyXFVpMb
Symbiodinium-goreaui-typeC1	10	yes	core	vBtLzyXFVpMb
Alexandrium-margalefi-CS	9	yes	core	vBtLzyXFVpMb
Glenodinium-foliaceum	9	yes	core	vBtLzyXFVpMb
Togula-jolla-CCCM725	9	yes	core	vBtLzyXFVpMb
Symbiodinium-sp--A3	8	yes	core	vBtLzyXFVpMb
Alexandrium-catenella	7	yes	core	vBtLzyXFVpMb
Prorocentrum-minimum-CCMP1329	7	yes	core	vBtLzyXFVpMb
Scrippsiella-nutricula	7	yes	core	vBtLzyXFVpMb
Scrippsiella-trochoidea-CCMP-3099	7	yes	core	vBtLzyXFVpMb
Alexandrium-minutumHALIM-CCMP113	6	yes	non core (<9000 transcripts)	vBtLzyXFVpMb
Azadinium-spinosum-3D9	6	yes	core	vBtLzyXFVpMb
het	6	yes	core	vBtLzyXFVpMb
Gymnodinium-catenatum	6	yes	core	vBtLzyXFVpMb
Alexandrium-monilatum-JR08	5	yes	core	vBtLzyXFVpMb
Amphidinium-carterae-CCMP1314	5	yes	core	vBtLzyXFVpMb
Pelagodinium-beii	5	yes	core	vBtLzyXFVpMb
Protoceratium-reticulatum	5	yes	core	vBtLzyXFVpMb
Symbiodinium-sp--CCMP421	5	yes	core	vBtLzyXFVpMb
Amphidinium-massartii	4	yes	core	vBtLzyXFVpMb
Ceratium-fusus	4	yes	core	vBtLzyXFVpMb
Gambierdiscus-australes-CAWD149	4	yes	core	vBtLzyXFVpMb
Gambierdiscus-sp	4	yes	non core (<9000 transcripts)	vBtLzyXFVpMb
Gonyaulax-spinifera-CCMP409	4	yes	core	vBtLzyXFVpMb
Heterocapsa-arctica	4	yes	core	vBtLzyXFVpMb
plg	4	yes	core	vBtLzyXFVpMb
Alexandrium-andersonii-CCMP2222	3	yes	core	vBtLzyXFVpMb
Crypthecodinium-cohnii	3	yes	core	vBtLzyXFVpMb
Lingulodinium-polyedra-CCMP1738	3	yes	core	vBtLzyXFVpMb
Noctiluca-scintillans	3	yes	core	vBtLzyXFVpMb
Oxyrrhis-marina	3	yes	non core (<9000 transcripts)	vBtLzyXFVpMb
Symbiodinium-sp--CCMP2430	3	yes	core	vBtLzyXFVpMb
Dinophysis-acuminata-DAEP01	2	yes	core	vBtLzyXFVpMb
Heterocapsa-rotundata	2	yes	core	vBtLzyXFVpMb
Heterocapsa-triquetra	2	yes	core	vBtLzyXFVpMb
Polarella-glacialis	2	yes	core	vBtLzyXFVpMb
Symbiodinium-sp--Mp	2	yes	core	vBtLzyXFVpMb
Akashiwo-sanguinea	0	no		
Alexandrium-fundyense	0	no		
Gyrodinium-dominans	0	no		
Lessardia-elongata	0	no		
Oxyrrhis-marina-CCMP788	0	no		
Peridinium-aciculiferum-PAER	0	no		
Prorocentrum-lima-CCMP684	0	no		
Prorocentrum-micans	0	no		
Pyrocystis-lunula-CCCM517	0	no		
Scrippsiella-cf-hangoei-SHHI	0	no		
Thoracosphaera-heimii	0	no		

	Number of sequences
Total number of sequences In VbtLzyXFVpMb	328
Number of core transcriptome sequences in VbtLzyXFVpMb	320
Number of non-core transcriptome sequences in VbtLzyXFVpMb	8

SI Appendix Tab. S18 : 101 multiprotein alignements (Janouskovec et al.) in core CCs

	Number of alignments of core sequences vs. 101 orthologous alignements of Janouskovec et al. 2016	Number of unique sequences	Number of CC
evalue 1e-3 + pident >= 90	39 607	5 965	46

SI Appendix Tab. S19 : symbiosis trait-CCs with no functional annotation

	# CC	# sequences
non annotated symbiosis associated CC (Interpro)	52 491	130 673
non annotated symbiosis associated CC (Interpro + nr)	52 193	129 754

SI Appendix Tab. S20 : STX sequences sources

Gene	source
sxtA1-4 and sxtG	**Murray et al. 2015 - M&M** sxtA GenBank accessions : JF343240-JF343356 sxtG GenBank accessions : JX995111-JX995130
PKS	**Kohli et al. 2015** - Supplementary materials 3 (http://dx.doi.org/10.1186/s12864-015-1625-y)

SI Appendix Tab. S21 : symbiosis sequences sources

gene_name	gene_id	organism	symbiosis_type	rôle	source
P-type H+-ATPase	SspPMA1	Symbiodinium spp.	Cnidarian-Algal		Symbiosis-dependent gene expression in coral–dinoflagellate association: cloning and characterization of a P-type H+-ATPase gene
nodulation protein U	Q53515 P31957 Q01990 P26027 Q07759	Symbiodinium kawagutti Symbiodinium minutum		symbiosis_establishment	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
nodulation protein nolO	AAB91692.1 ACP22653.1 CAD77047.1 AJX25945.1 ADV14890.1 Q45269 P18408 P94498	Symbiodinium kawagutti Symbiodinium minutum		symbiosis_establishment	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
Phosphoadenosine phosphosulfate reductase	P18408 P94498 1SUR Q9V0S1 AAM01462.1	Symbiodinium kawagutti Symbiodinium minutum		symbiosis_establishment	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
calumenin	G5EBH7 CAP33056.1 EFO82780.1	Symbiodinium kawagutti Symbiodinium minutum		cell_recognition	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
Phosphatidylinositol-3-phosphate kinase	ABX71159.1 O88763 Q6AZN6 Q5D891 Q8NEB9	Symbiodinium kawagutti Symbiodinium minutum		cell_recognition	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
Host cell factor 2	Q9Y5Z7 Q9D968 Q5RKG2 Q9Y5Z7.1 Q9D968.1 Q5RKG2.1 NP_001008358.1 NP_001074687.1 AAD27814.1 NP_037452.1	Symbiodinium kawagutti Symbiodinium minutum		cell_recognition	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
Fibronectin type III domain containing protein 3C1	Q6DFV6 NP_001007581.1 NP_001178651.1 ERE65388.1 AAH85176.1 AAH76625.1	Symbiodinium kawagutti Symbiodinium minutum		cell_recognition	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
Spondin-1	Q9HCB6 P35446 Q9GLX9 Q8VCC9 P35447 Q9W770 NP_990182.1	Symbiodinium kawagutti Symbiodinium minutum		cell_recognition	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
Merozoite surface protein 1	P04932 P04933 P50495 P04934 P19598 P08569 P13828 P13819 P13827 P13820	Symbiodinium kawagutti Symbiodinium minutum		cell_recognition	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
Avirulence protein AvrBs3	P14727 OFA08825.1 OEZ99879.1 KTD30443.1 KTD37687.1	Symbiodinium kawagutti Symbiodinium minutum		infection	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
Putative surface-exposed virulence protein BigA	P25927 EKJ63886.1 ESH74807.1	Symbiodinium kawagutti Symbiodinium minutum		infection	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
UDP-galactopyranose mutase	P37747 P9WIQ1 P9WIQ0 Q49398 P75499	Symbiodinium kawagutti Symbiodinium minutum		infection	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
Sporozoite developmental protein	P42789	Symbiodinium kawagutti Symbiodinium minutum		infection	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
Mycocerosic acid synthase	O53901 A0R1E8 Q02251 CKM94211.1 CKM94258.1 CKP55567.1 CMI61551.1	Symbiodinium kawagutti Symbiodinium minutum		infection	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
Transmembrane protein 151 homolog	NP_495963.2 Q626N3 Q23387	Symbiodinium kawagutti Symbiodinium minutum		unclear_function	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
Chaoptin	P12024 P82963 KKP35865.1	Symbiodinium kawagutti Symbiodinium minutum		unclear_function	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
Uncharacterized MFS-type transporter ydgK	P96709 BAS06723.1 BAR73475.1 AEN90579.1 BAV22866.1 GAD08910.1 BAN08010.1 CRH29135.1 CCB74084.1	Symbiodinium kawagutti Symbiodinium minutum		unclear_function	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
Uncharacterized UDP-glucosyltransferase yojK	O31853	Symbiodinium kawagutti Symbiodinium minutum		unclear_function	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
Uncharacterized protein PF11_0213	Q8IIG1 Q8IIG1.2	Symbiodinium kawagutti Symbiodinium minutum		unclear_function	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis
Protein dpy-19 homolog 1	Q2PZI1 Q2PZI1-2	Symbiodinium kawagutti Symbiodinium minutum		unclear_function	The symbiodinium kawagutti genome illuminates dinoflagellate gene expression and coral symbiosis

SI Appendix Tab. S21 : symbiosis sequences sources

gene_name	gene_id	organism	symbiosis_type	rôle	source
Superoxide dismutase	Q9WU84 Q8HXQ3 Q9JK72 P80174 Q52RN5	Nematostella vectensis Aiptasia pallida Anemonia viridis Acropora digitifera Acropora hyacinthus Acropora millepora Acropora palmata Montastreae faveolata Porites astreoides Pocillopora damicornis	Cnidarian-Algal		Study of Cnidarian-Algal Symbiosis in the “Omics” Age
Catalase	P17336 P90682 Q64405 Q9PT92 Q9PWF7 O62839	Nematostella vectensis Aiptasia pallida Anemonia viridis Acropora digitifera Acropora hyacinthus Acropora millepora Acropora palmata Montastreae faveolata Porites astreoides Pocillopora damicornis	Cnidarian-Algal		Study of Cnidarian-Algal Symbiosis in the “Omics” Age
Peroxiredoxin	Q90384 Q9BGI2 Q5ZJF4 Q9Z0V5	Nematostella vectensis Aiptasia pallida Anemonia viridis Acropora digitifera Acropora hyacinthus Acropora millepora Acropora palmata Montastreae faveolata Porites astreoides Pocillopora damicornis	Cnidarian-Algal		Study of Cnidarian-Algal Symbiosis in the “Omics” Age
Glutathione peroxidase	P00435 P11352 Q96SL4 O18994 Q00277 Q08BT9 P30710 Q95003 Q4AEI0	Nematostella vectensis Aiptasia pallida Anemonia viridis Acropora digitifera Acropora hyacinthus Acropora millepora Acropora palmata Montastreae faveolata Porites astreoides Pocillopora damicornis	Cnidarian-Algal		Study of Cnidarian-Algal Symbiosis in the “Omics” Age
Glutathione reducase	P70619 P47791 P00390 A2TIL1 Q6BPI1 Q60151	Nematostella vectensis Aiptasia pallida Anemonia viridis Acropora digitifera Acropora hyacinthus Acropora millepora Acropora palmata Montastreae faveolata Porites astreoides Pocillopora damicornis	Cnidarian-Algal		Study of Cnidarian-Algal Symbiosis in the “Omics” Age
g-glutamylcysteine synthetase	Q9W3K5 P19468 P97494 Q9NFN6 P48506	Nematostella vectensis Aiptasia pallida Anemonia viridis Acropora digitifera Acropora hyacinthus Acropora millepora Acropora palmata Montastreae faveolata Porites astreoides Pocillopora damicornis	Cnidarian-Algal		Study of Cnidarian-Algal Symbiosis in the “Omics” Age
Glutathione synthetase	Q54E83 P35668 P46413 O22494 P46416 P51855 Q5EAC2 P48637	Nematostella vectensis Aiptasia pallida Anemonia viridis Acropora digitifera Acropora hyacinthus Acropora millepora Acropora palmata Montastreae faveolata Porites astreoides Pocillopora damicornis	Cnidarian-Algal		Study of Cnidarian-Algal Symbiosis in the “Omics” Age
Glutathione S-transferase	P46434 P46436 P21266 P91253 P46427 P91252 O73888 O18598	Nematostella vectensis Aiptasia pallida Anemonia viridis Acropora digitifera Acropora hyacinthus Acropora millepora Acropora palmata Montastreae faveolata Porites astreoides Pocillopora damicornis	Cnidarian-Algal		Study of Cnidarian-Algal Symbiosis in the “Omics” Age
ferritin	P42577	Nematostella vectensis Aiptasia pallida Anemonia viridis Acropora digitifera Acropora hyacinthus Acropora millepora Acropora palmata Montastreae faveolata Porites astreoides Pocillopora damicornis	Cnidarian-Algal		Study of Cnidarian-Algal Symbiosis in the “Omics” Age
Solute carrier family 2, facilitated glucose transporter member 8	Q9NY64	Homo sapiens (Human)	Cnidarian-Algal	transport	Extensive Differences in Gene Expression Between Symbiotic and Aposymbiotic Cnidarians
Sodium/myo-inositol cotransporter 2	Q28728	Oryctolagus cuniculus (Rabbit)	Cnidarian-Algal	transport	Extensive Differences in Gene Expression Between Symbiotic and Aposymbiotic Cnidarians
Lipid storage droplets surface-binding protein 2	Q9VXY7	Drosophila melanogaster (Fruit fly)	Cnidarian-Algal	transport	Extensive Differences in Gene Expression Between Symbiotic and Aposymbiotic Cnidarians
Scavenger receptor class B member 1	Q8WTV0	Homo sapiens (Human)	Cnidarian-Algal	transport	Extensive Differences in Gene Expression Between Symbiotic and Aposymbiotic Cnidarians
Organic cation transporter protein	Q9VCA2	Drosophila melanogaster (Fruit fly)	Cnidarian-Algal	transport	Extensive Differences in Gene Expression Between Symbiotic and Aposymbiotic Cnidarians
Epididymal secretory protein E1	P61916	Homo sapiens (Human)	Cnidarian-Algal	transport	Extensive Differences in Gene Expression Between Symbiotic and Aposymbiotic Cnidarians
Sodium- and chloride-dependent taurine transporter	Q9MZ34	Bos taurus (Bovine)	Cnidarian-Algal	transport	Extensive Differences in Gene Expression Between Symbiotic and Aposymbiotic Cnidarians
Monocarboxylate transporter 10	Q3U9N9	Mus musculus (Mouse)	Cnidarian-Algal	transport	Extensive Differences in Gene Expression Between Symbiotic and Aposymbiotic Cnidarians
Vesicular inhibitory amino acid transporter	Q6PF45	Xenopus laevis (African clawed frog)	Cnidarian-Algal	transport	Extensive Differences in Gene Expression Between Symbiotic and Aposymbiotic Cnidarians
Carbonic anhydrase 2	Q8UWA5	Tribolodon hakonensis (Big-scaled redfin)	Cnidarian-Algal	transport	Extensive Differences in Gene Expression Between Symbiotic and Aposymbiotic Cnidarians
Aquaporin-5	Q866S3	Ovis aries (Sheep)	Cnidarian-Algal	transport	Extensive Differences in Gene Expression Between Symbiotic and Aposymbiotic Cnidarians
Aquaporin-3	A9Y006	Sus scrofa (Pig)	Cnidarian-Algal	transport	Extensive Differences in Gene Expression Between Symbiotic and Aposymbiotic Cnidarians
Ammonium transporter Rh type B	Q7T070	Danio rerio (Zebrafish) (Brachydanio rerio)	Cnidarian-Algal	transport	Extensive Differences in Gene Expression Between Symbiotic and Aposymbiotic Cnidarians
Putative ammonium transporter 1	P54145	Caenorhabditis elegans	Cnidarian-Algal	transport	Extensive Differences in Gene Expression Between Symbiotic and Aposymbiotic Cnidarians

SI Appendix Tab. S22 : MMETSP dataset metrics

id	species	taxon_id	taxonomy	assembly	library_count	library_list	peptides	lpi	kog	superfamily	pfam	pfam_%	tigrfam	tmhmm	transporter
MMETSP0223	Akashiwo sanguinea, Strain CCCM 885	143672	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Akashiwo;Akashiwo sanguinea, Strain CCCM 885	single	1	MMETSP0223	112	73	6	99	42	0,38	14	155	0
MMETSP1436	Alexandrium andersonii, Strain CCMP2222	327968	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Alexandrium;Alexandrium andersonii, Strain CCMP2222	single	1	MMETSP1436	38688	21060	5590	39663	16085	0,42	7234	56136	4
MMETSP0790	Alexandrium catenella, Strain OF101	2925	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Alexandrium;Alexandrium catenella, Strain OF101	single	1	MMETSP0790	69322	48618	14759	76134	36863	0,53	17348	127024	26
Alexandrium-fundyense-CCMP1719	Alexandrium fundyense CCMP1719	2932	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Alexandrium;Alexandrium fundyense CCMP1719	combined	2	MMETSP0196, MMETSP0347	9078	6830	1711	9275	4187	0,46	1860	11364	5
MMETSP0661	Alexandrium margalefi, Strain AMGDE01CS-322	109239	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Alexandrium;Alexandrium margalefi, Strain AMGDE01CS-322	single	1	MMETSP0661	50619	33177	9636	54567	25292	0,50	11403	95563	16
MMETSP0328	Alexandrium minutum	39455	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Alexandrium;Alexandrium minutum	single	1	MMETSP0328	12568	8790	2653	13256	6084	0,48	2932	20944	4
Alexandrium-monilatum-CCMP3105	Alexandrium monilatum, Strain CCMP3105	311494	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Alexandrium;Alexandrium monilatum, Strain CCMP3105	combined	4	MMETSP0096, MMETSP0095, MMETSP0093, MMETSP0097	93332	57328	17987	100671	47624	0,51	22889	173175	271
Alexandrium-temarense-CCMP1771	Alexandrium temarense CCMP1771	2926	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Alexandrium;Alexandrium temarense CCMP1771	combined	4	MMETSP0378, MMETSP0380, MMETSP0384, MMETSP0382	117461	78998	27884	134862	66365	0,56	31986	228814	567
Amphidinium-carterae-CCMP1314	Amphidinium carterae, Strain CCMP1314	2961	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Amphidinium;Amphidinium carterae, Strain CCMP1314	combined	3	MMETSP0258, MMETSP0398, MMETSP0259	36236	29701	10794	45139	25044	0,69	11568	92067	246
MMETSP0689	Amphidinium massartii	160604	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Amphidinium;Amphidinium massartii	single	1	MMETSP0689	49772	39311	15193	62474	34889	0,70	16357	132455	33
Azadinium-spinosum-3D9	Azadinium spinosum 3D9	632150	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Azadinium;Azadinium spinosum 3D9	combined	3	MMETSP1036, MMETSP1037, MMETSP1038	71170	50490	16675	83698	42849	0,60	20424	149112	322
MMETSP1462	Brandtodinium nutriculum, Strain RCC3387	1333877	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Brandtodinium;Brandtodinium nutriculum, Strain RCC3387	single	1	MMETSP1462	66441	37259	12844	73444	33233	0,50	14844	131906	36
Ceratium-fusus-PA161109	Ceratium fusus, Strain PA161109	2916	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Ceratium;Ceratium fusus, Strain PA161109	combined	2	MMETSP1075, MMETSP1074	69175	43573	15278	80207	42116	0,61	18995	149649	302
Crypthecodinium-cohnii-Seligo	Crypthecodinium cohnii Seligo	2866	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Crypthecodinium;Crypthecodinium cohnii Seligo	combined	4	MMETSP0325, MMETSP0326, MMETSP0324, MMETSP0323	48368	27832	11456	56086	29430	0,61	14302	105101	264
Durinskia-baltica-CSIRO_CS-38	Durinskia baltica, Strain CSIRO CS-38	400756	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Durinskia;Durinskia baltica, Strain CSIRO CS-38	combined	2	MMETSP0117, MMETSP0116	89242	51714	19573	100978	47084	0,53	21784	169890	26
MMETSP0766	Gambierdiscus australes	439317	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Gambierdiscus;Gambierdiscus australes	single	1	MMETSP0766	48911	31558	10055	54222	26530	0,54	11999	97846	15
MMETSP1439	Gonyaulax spinifera, Strain CCMP409	66791	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Gonyaulax;Gonyaulax spinifera, Strain CCMP409	single	1	MMETSP1439	39825	23739	7970	42496	19626	0,49	9375	71291	11
MMETSP0784	Gymnodinium catenatum	39447	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Gymnodinium;Gymnodinium catenatum	single	1	MMETSP0784	83508	56369	22906	99158	52256	0,63	25143	190401	46
MMETSP1148	Gyrodinium dominans, Strain SPMC 103	285985	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Gyrodinium;Gyrodinium dominans, Strain SPMC 103	single	1	MMETSP1148	117	69	18	91	38	0,32	16	146	0
MMETSP1441	Heterocapsa arctica, Strain CCMP445	192219	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Heterocapsa;Heterocapsa arctica, Strain CCMP445	single	1	MMETSP1441	45642	26449	7710	49075	21120	0,46	9787	75575	10
MMETSP0503	Heterocapsa rotundata, Strain SCCAPK-0483	89963	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Heterocapsa;Heterocapsa rotundata, Strain SCCAPK-0483	single	1	MMETSP0503	44207	36104	9831	50433	23506	0,53	11757	83181	11
MMETSP0448	Heterocapsa triquetra	66468	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Heterocapsa;Heterocapsa triquetra	single	1	MMETSP0448	57789	33844	10311	62799	27877	0,48	12898	104614	19
Karenia-brevis-CCMP2229	Karenia brevis CCMP2229	-311225	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Karenia;Karenia brevis CCMP2229	combined	4	MMETSP0027, MMETSP0031, MMETSP0029, MMETSP0030	80065	70968	18780	95421	50959	0,64	23514	200917	36
Karenia-brevis-SP1	Karenia brevis SP1	-311226	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Karenia;Karenia brevis SP1	combined	2	MMETSP0573, MMETSP0574	85537	75835	20967	101885	55865	0,65	25869	217254	41
Karenia-brevis-SP3	Karenia brevis SP3	-311227	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Karenia;Karenia brevis SP3	combined	2	MMETSP0528, MMETSP0527	70560	64372	16493	82798	45067	0,64	21092	179704	26
Karenia-brevis-Wilson	Karenia brevis Wilson	-311228	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Karenia;Karenia brevis Wilson	combined	4	MMETSP0202, MMETSP0201, MMETSP0648, MMETSP0649	93029	82026	21387	107874	58839	0,63	27358	227771	53
Karlodinium-micrum-CCMP2283	Karlodinium micrum, Strain CCMP2283	342587	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Karlodinium;Karlodinium micrum, Strain CCMP2283	combined	3	MMETSP1016, MMETSP1015, MMETSP1017	58585	40672	16595	71291	40034	0,68	19390	126876	39
Glenodinium-foliaceum-CCAP1116_3	Kryptoperidinium foliaceum CCAP1116_3	-311229	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Kryptoperidinium;Kryptoperidinium foliaceum CCAP1116_3	combined	2	MMETSP0118, MMETSP0119	107199	75552	18892	117709	50151	0,47	23305	179996	29
Kryptoperidinium-foliaceum-CCMP1326	Kryptoperidinium foliaceum CCMP1326	-311230	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Kryptoperidinium;Kryptoperidinium foliaceum CCMP1326	combined	2	MMETSP0120, MMETSP0121	162643	113945	32722	180904	82718	0,51	38044	294649	19
MMETSP1147	Lessardia elongata, Strain SPMC 104	210733	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Lessardia;Lessardia elongata, Strain SPMC 104	single	1	MMETSP1147	408	200	77	326	193	0,47	67	486	0
Lingulodinium-polyedra-CCMP1738	Lingulodinium polyedra CCMP1738	-311231	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Lingulodinium;Lingulodinium polyedra CCMP1738	combined	4	MMETSP1034, MMETSP1032, MMETSP1035, MMETSP1033	99361	57609	17251	109267	48080	0,48	22682	177070	31
Peridinium-aciculiferum-PAER_2	Peridinium aciculiferum PAER_2	268820	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Peridinium;Peridinium aciculiferum PAER_2	combined	2	MMETSP0370, MMETSP0371	53894	44275	12212	61394	30635	0,57	14424	100418	16
MMETSP0252	Prorocentrum lima, Strain CCMP684	39448	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Prorocentrum;Prorocentrum lima, Strain CCMP684	single	1	MMETSP0252	502	234	51	426	205	0,41	79	665	0
MMETSP0251	Prorocentrum micans, Strain CCCM 845	2945	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Prorocentrum;Prorocentrum micans, Strain CCCM 845	single	1	MMETSP0251	942	332	81	774	309	0,33	137	1131	0
Prorocentrum-minimum-CCMP1329	Prorocentrum minimum CCMP1329	-311218	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Prorocentrum;Prorocentrum minimum CCMP1329	combined	4	MMETSP0053, MMETSP0055, MMETSP0057, MMETSP0056	86751	60711	13908	94493	39161	0,45	17802	148748	30
Prorocentrum-minimum-CCMP2233	Prorocentrum minimum CCMP2233	39449	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Prorocentrum;Prorocentrum minimum CCMP2233	combined	3	MMETSP0269, MMETSP0267, MMETSP0268	80737	59739	13924	89555	38111	0,47	17525	151641	20
MMETSP0228	Protoceratium reticulatum, Strain CCCM 535 (=CCMP 1889)	133427	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Protoceratium;Protoceratium reticulatum, Strain CCCM 535 (=CCMP 1889)	single	1	MMETSP0228	75927	48692	16323	85658	41016	0,54	18957	155385	19
MMETSP0229	Pyrocystis lunula, Strain CCCM 517	2972	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Pyrocystis;Pyrocystis lunula, Strain CCCM 517	single	1	MMETSP0229	232	118	13	188	70	0,30	16	290	0
MMETSP0796	Pyrodinium bahamense	73915	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Pyrodinium;Pyrodinium bahamense	single	1	MMETSP0796	100308	67870	21865	115106	56221	0,56	26143	215662	45
Scrippsiella-hangoei-SHTV5	Scrippsiella hangoei SHTV5	268821	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Scrippsiella;Scrippsiella hangoei SHTV5	combined	3	MMETSP0361, MMETSP0359, MMETSP0360	75339	57668	17611	90213	44017	0,58	21121	168275	22
Scrippsiella-hangoei_like-SHHI_4	Scrippsiella hangoei-like, Strain SHHI-4	447022	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Scrippsiella;Scrippsiella hangoei-like, Strain SHHI-4	combined	3	MMETSP0368, MMETSP0369, MMETSP0367	75693	61946	17108	86442	43285	0,57	20958	151270	21
Scrippsiella-trochoidea-CCMP3099	Scrippsiella trochoidea CCMP3099	71861	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Scrippsiella;Scrippsiella trochoidea CCMP3099	combined	3	MMETSP0272, MMETSP0271, MMETSP0270	103400	62373	21825	121212	59693	0,58	28336	236983	47
MMETSP0225	Thoracosphaera heimii, Strain CCCM 670 (=CCMP 1069)	2923	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Thoracosphaera;Thoracosphaera heimii, Strain CCCM 670 (=CCMP 1069)	single	1	MMETSP0225	172	97	11	140	47	0,27	15	250	0
MMETSP0224	Togula jolla, Strain CCCM 725	285029	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophyceae_XX;Togula;Togula jolla, Strain CCCM 725	single	1	MMETSP0224	42755	29363	12742	50771	27777	0,65	13445	98675	19
MMETSP0797	Dinophysis acuminata	47934	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Dinophysiales;Dinophysis;Dinophysis acuminata	single	1	MMETSP0797	84397	44465	14518	92301	37711	0,45	17125	128594	23
MMETSP0253	Noctiluca scintillans	2966	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Noctilucales;Noctiluca;Noctiluca scintillans	single	1	MMETSP0253	40911	24994	11004	47555	25597	0,63	11591	92232	33
Oxyrrhis-marina	Oxyrrhis marina	2969	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Oxyrrhinaceae;Oxyrrhis;Oxyrrhis marina	combined	4	MMETSP0471, MMETSP0469, MMETSP0470, MMETSP0468	43350	14377	10761	51049	24567	0,57	11518	89114	154
Oxyrrhis-marina-LB1974	Oxyrrhis marina LB1974	-311155	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Oxyrrhinaceae;Oxyrrhis;Oxyrrhis marina LB1974	combined	3	MMETSP1426, MMETSP1424, MMETSP1425	34417	15405	8802	40852	20082	0,58	9355	72317	33
MMETSP0044	Oxyrrhis marina, Strain CCMP1788	-311153	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Oxyrrhinaceae;Oxyrrhis;Oxyrrhis marina, Strain CCMP1788	single	1	MMETSP0044	387	302	96	333	197	0,51	75	475	0
Oxyrrhis-marina-CCMP1795	Oxyrrhis marina, Strain CCMP1795	-311154	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Oxyrrhinaceae;Oxyrrhis;Oxyrrhis marina, Strain CCMP1795	combined	1	MMETSP0451	2013	1628	582	2038	1165	0,58	477	2399	0
MMETSP1338	Pelagodinium beii	43686	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Suessiales;Pelagodinium;Pelagodinium beii	single	1	MMETSP1338	48038	33165	12504	56579	30982	0,64	14451	106084	28
MMETSP0227	Polarella glacialis CCMP1383	348536	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Suessiales;Polarella;Polarella glacialis CCMP1383	single	1	MMETSP0227	59754	43818	15813	69055	36753	0,62	17692	120780	18
MMETSP1440	Polarella glacialis, Strain CCMP2088	89957	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Suessiales;Polarella;Polarella glacialis, Strain CCMP2088	single	1	MMETSP1440	33896	26399	7985	36745	19213	0,57	9087	62246	9
Symbiodinium-kawagutii-CCMP2468	Symbiodinium kawagutii CCMP2468	104179	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Suessiales;Symbiodinium;Symbiodinium kawagutii CCMP2468	combined	1	MMETSP0132	11679	7430	1856	11742	5911	0,51	2464	14199	1
MMETSP1374	Symbiodinium Mf12.5f/clade A3	993574	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Suessiales;Symbiodinium;Symbiodinium Mf12.5f/clade A3	single	1	MMETSP1374	37538	30038	9556	42225	23991	0,64	10781	80159	14
Symbiodinium-sp-C15	Symbiodinium sp C15	226949	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Suessiales;Symbiodinium;Symbiodinium sp C15	combined	2	MMETSP1370, MMETSP1371	37277	28961	10854	43828	24757	0,66	11534	82547	16
Symbiodinium-sp-CCMP2430	Symbiodinium sp CCMP2430	631057	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Suessiales;Symbiodinium;Symbiodinium sp CCMP2430	combined	3	MMETSP1117, MMETSP1115, MMETSP1116	44733	36637	11205	51859	28091	0,63	13094	113214	33
Symbiodinium-sp-Mp	Symbiodinium sp Mp	230996	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Suessiales;Symbiodinium;Symbiodinium sp Mp	combined	4	MMETSP1124, MMETSP1123, MMETSP1122, MMETSP1125	43951	32875	11242	52060	27762	0,63	13025	107760	23
MMETSP1377	Symbiodinium type D1a	462645	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Suessiales;Symbiodinium;Symbiodinium type D1a	single	1	MMETSP1377	45153	35715	12514	48124	27563	0,61	12793	80392	19
Symbiodinium-sp-C1	Symbiodinium, Strain C1	226962	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Suessiales;Symbiodinium;Symbiodinium, Strain C1	combined	2	MMETSP1367, MMETSP1369	45782	37569	13465	130	31216	0,68	14784	108658	271
MMETSP1110	Symbiodinium, Strain CCMP421	174948	Eukaryota;Alveolata;Dinophyta;Dinophyceae;Dinophyceae_X;Suessiales;Symbiodinium;Symbiodinium, Strain CCMP421	single	1	MMETSP1110	72821	60292	21314	91617	48810	0,67	23547	196349	36
MMETSP0795	Amoebophrya sp.	88552	Eukaryota;Alveolata;Dinophyta;Syndiniales;Dino-Group-II;Dino-Group-II-Clade-4;Amoebophrya;Amoebophrya sp.	single	1	MMETSP0795	14350	4387	2693	16286	7921	0,55	3475	28838	5

SI Appendix Tab. S19 : symbiosis trait-CCs with no functional annotation

	# CC	# sequences
non annotated h4h associated CC (Interpro)	25 550	56 918
non annotated h4h associated CC (Interpro + nr)	25 393	56 559

Le réseau de similarité de séquences m'a permis d'explorer le grand nombre de séquences peptidiques (1 275 911 séquences réparties dans 350 267 CCs) constituant l'ensemble des protéomes* de 46 espèces de dinoflagellés. Par ailleurs, 529 837 de ces séquences (41%) n'étaient pas annotées fonctionnellement et étaient donc *a priori* inexploitables, mais ont pu être prises en compte grâce à cette approche de réseau. Un ensemble de 37 842 domaines protéiques fonctionnels communs à 43 protéomes* de dinoflagellés et réparti dans 252 CCs a été identifié. 16 de ces CCs restent à ce jour sans aucune annotation fonctionnelle. L'approche par réseau a facilité le croisement des nombreuses données de domaines protéiques avec les traits fonctionnels des organismes étudiés. Ainsi, 90 794 CCs, composées de domaines protéiques fonctionnels conservés appartenant exclusivement à des dinoflagellés symbiotiques, ont été mise en évidence. Parmis ces CCs, j'ai identifié un nombre important de domaines protéiques dont la fonction est liée au transport d'ions et qui sont localisés au niveau membranaire. Grâce à l'utilisation des SSN*, nous savons que les domaines protéiques annotées comme "transporteurs d'ions" ne sont cependant pas exclusives aux 90 794 CCs symbiotiques. Ce qui m'a conduit à réaliser une estimation basée sur le nombre de domaines observés avec cette annotation et m'a permis de montrer que ces protéines semblaient cependant deux fois plus abondantes chez les espèces symbiotiques que chez les espèces non-symbiotiques. De fait, je suggère que ces processus liées aux échanges d'ions sont particulièrement importants chez les lignées symbiotiques de dinoflagellés et que de plus amples études focalisée sur ces domaines pourraient nous permettre de mieux comprendre la physiologie des organismes impliqués dans ces relations de symbioses. De plus, 57% des CCs (et donc des domaines protéiques qui les composent) sont non-annotés et pourraient correspondre à de nouvelles fonctions et potentiels marqueurs génomiques encore inconnus chez les dinoflagellés symbiotiques. De tels marqueurs pourraient être à la base de nouvelles hypothèses concernant la physiologie des dinoflagellés et pourraient être validés au travers d'expériences *in vitro*. D'autre part, les domaines protéiques appartenant aux CCs symbiotiques peuvent constituer des marqueurs intéressants de la présence de dinoflagellés symbiotiques au sein de données méta-génomique*, méta-transcriptomiques* ou méta-protéomiques. L'ajout de nouvelles souches de di-

noflagellés (symbiotiques ou non-symbiotiques) au sein du SSN* déjà existant pourrait permettre de confirmer les marqueurs issus des CCs identifiés dans cette étude. Sur la base de ces résultats, la recherche de ces potentiels marqueurs dans le cadre de l'étude de transcriptomes d'holobiontes, mélangeant séquences de radiolaires et de dinoflagellés, est alors possible et pourrait permettre de proposer des pistes de la caractérisation fonctionnelle de cette symbiose.

Chapitre 4

Assemblage et analyse fonctionnelle des transcriptomes d'holobiontes

L'observation de la symbiose entre radiolaires et dinoflagellés date du 19 ème siècle, mais ce n'est seulement que depuis récemment que les acteurs de ces associations commencent à être identifiés plus précisément, grâce notamment à l'évolution des méthodes d'échantillonnage et de séquençage du matériel génétique de ces organismes [Decelle, Suzuki et al. 2012 ; Probert et al. 2014]. L'obtention de séquences génomiques (*i.e.* 18S et 28S rDNA) a permis l'utilisation de phylogénie moléculaire pour identifier les partenaires de ces symbioses, mais les études de génomique fonctionnelle permettant de caractériser les mécanismes fonctionnels impliqués dans ces interactions sont encore rares [Balzano et al. 2015]. Ces associations symbiotiques sont particulièrement fréquentes dans les océans [Decelle, Probert et al. 2012] et il apparaît aujourd'hui important de mieux comprendre les mécanismes qui les régissent. L'étude des holobiontes composés d'un hôte radiolaire et de micro-algues symbiontes dinoflagellés constitue ainsi le deuxième volet de cette thèse. Pour aborder cette étude, j'ai choisi de travailler dans un premier temps sur un exemple théorique afin d'élaborer une stratégie d'analyse adaptée à un transcriptome d'holobionte (Figure 4.1). J'aborde donc ici les défis bio-informatiques et computationnelles que représentent l'étude des transcriptomes d'holobiontes impliquant des organismes non-modèles*. Pour palier à ces difficultés, je présenterai la mise en place d'une nouvelle approche bioinformatique réalisée en collaboration avec l'équipe GenScale de l'INRIA de Rennes (https://team.inria.fr/genscale/). À ce jour cette approche innovante a été valorisée par 2 communications internationales (conférence internationale sur les holobiontes à Paris en avril 2017, conférence RCAM *Recent Computational Advances in Metagenomics* à Paris en octobre 2017) et une communication nationale (colloque de Génomique Environnementale à Marseille, 13-15 septembre 2017). De plus, un article sera soumis pour un numéro spécial de la revue *Microbiome* (https://link.springer.com/journal/40168) courant novembre 2017.

En combinant cette nouvelle approche avec la chaîne d'analyses présentée au chapitre 3, j'ai dans un deuxième temps analysé les données de 3 transcriptomes d'holobiontes radiolaire-dinoflagellés. Dans cette étude, je tente d'identifier les gènes candidats potentiellement importants dans les mécanismes de la mise en place et du

maintien de ces associations symbiotiques.

4.1 Développement d'une chaîne d'analyses adaptée à l'étude de transcriptomes d'holobiontes

Les transcriptomes d'holobiontes disponibles correspondent à des mélanges contenant des séquences d'hôte et de symbiontes qui sont, certes composés d'un nombre restreint d'acteurs par rapport à un méta-transcriptome environnementale, mais présente des ARN de diverses origines et donc une complexité accrue en comparaison d'un transcriptome monospécifique (Figure 4.1). Dans un exemple théorique de transcriptome d'holobionte, j'ai choisi de définir 4 sous-ensembles : (1) les séquences résultant uniquement de l'expression des gènes de l'hôte, (2) les séquences résultant de l'expression des gènes des partenaires symbiotiques, (3) les séquences dites *core* correspondant soit à des séquences exprimées à la fois par l'hôte et les symbiontes (expression de gènes communs) soit à des séquences produites par un des deux partenaires mais dont on ne peut pas distinguer l'appartenance spécifique (du fait de la similitude des séquences), et (4) des séquences qui ne peuvent être assignées à aucune des trois autres catégories (hôte, symbiontes ou *core*). Afin de déterminer ces 4 catégories de séquences, j'ai utilisé l'outil bioinformatique `Short Read Connector` (`SRC`) (Marchet et al. 2016).

4.1.1 `Short Read Connector` (`SRC`) : un outil adapté à la comparaison de très larges jeux de données de séquences

J'ai souhaité orienter mes travaux vers le développement d'une stratégie permettant de réduire le nombre de contigs chimériques* assemblés tout en assignant les séquences aux différents partenaires d'un holobionte avant l'étape d'assemblage. Le développement de cette stratégie s'est fait dans le cadre d'une collaboration avec Camille Marchet (doctorante en 2$^{\text{ème}}$ année) et Pierre Peterlongo (CR INRIA) de l'équipe GenScale à l'INRIA de Rennes qui ont développé le programme `Short Read Connector` (`SRC`) [Marchet et al. 2016]. `SRC` est un outil développé initialement

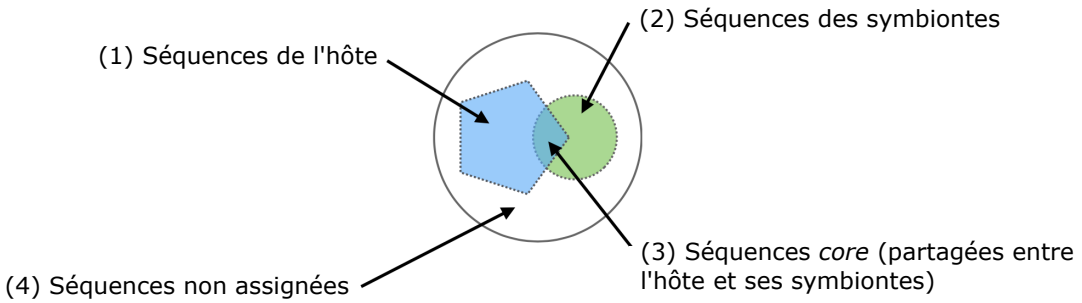


FIGURE 4.1 – Exemple théorique d'un transcriptome d'holobionte. Les quatres sous-ensembles correspondent à (1) les séquences de l'hôte, (2) les séquences des symbiontes, (3) les séquences *core* et (4) les séquences non-assignées. Figure tirée de ma communication orale dans le cadre du congrès international *Holobiont* au Muséum National d'Histoire Naturelle en mai 2017. Annexe D

pour estimer les similarités entre séquences de plusieurs jeux de méta-génomique. Les avantages de cet outil sont (1) sa capacité à comparer des jeux méta-génomiques* et/ou méta-transcriptomiques* composés de plusieurs millions de séquences (*e.g.* `SRC` est capable d'analyser un jeu de 189 207 003 lectures alors que les autres outils testés (Bowtie2 [LANGMEAD et SALZBERG 2012], BWA [H. LI et DURBIN 2010], starcode [ZORITA, CUSCÓ et FILION 2015]) échouent) et (2) un algorithme de comparaison de séquences basé sur la comparaison des k-mers* issus de séquences, différent des alignements de type Smith-Waterman utilisés par BLAST [ALTSCHUL et al. 1990]. En comparaison à BLAST, `SRC` peut comparer 1 millions de lectures en 13 secondes contre 13 minutes pour BLAST sur le même jeu de données [MARCHET et al. 2016].

Le programme `SRC` se décompose en deux versions : `Short Read Connector Linker` (`SRC_l`) et `Short Read Connector Counter` (`SRC_c`). La première phase de ces deux approches emploie une structure de données appelée quasi-dictionnaire dans lequel sont indexés tous les k-mers existants dans les séquences d'un jeu B. La seconde phase diverge ensuite pour les deux versions (`linker` et `counter`). `SRC_l` mesure une similarité entre séquences de deux ou plusieurs jeux de données, et identifie la liste de séquences similaires. Pour faire cela, la mesure de similarité entre une séquence b d'un jeu B et une séquence q d'un jeu Q correspond au nombre

de k-mers non-chevauchants de la séquence b qui apparaissent également dans la séquence q. Lorsque cette mesure dépasse un seuil arbitraire défini par l'utilisateur, les séquences b et q sont déclarées similaires (Figure 4.2). L'autre version de `SRC`, `SRC_c`, qui est la version que nous avons employée dans notre étude, consiste à estimer le nombre d'occurrences d'une séquence b issue d'un jeu B au sein d'un jeu Q. `SRC_c` recherche les k-mers du jeu Q parmi les k-mers indexés du jeu B dans le quasi-dictionnaire, afin d'estimer l'abondance d'un k-mer q dans le jeu B (Figure 4.2). L'abondance d'un k-mer q est approximée par rapport au nombre moyen d'occurrences de ce dernier dans le jeu B. Lorsque cette valeur d'abondance dépasse un seuil fixé par l'utilisateur, on considère que les k-mers b et q sont partagés entre les jeux B et Q. Lorsque deux séquences de B et Q partagent un minimum de x k-mers (x étant un seuil arbitraire défini par l'utilisateur) alors on estime que ces deux séquences sont partagées entre B et Q. Par exemple, on peut imposer à `SRC_c` la recherche de similarité à partir de 2 k-mers non-chevauchants partagés.

Les possibilités de comparaison de séquences entre jeux de données massifs offertes par `SRC` nous ont mené à l'élaboration d'un protocole permettant d'identifier les séquences communes entre un transcriptome d'holobionte et une ou plusieurs banques de référence. Ces banques sont construites afin d'inclure les séquences appartenant aux organismes potentiellement présents dans l'holobionte ou, à défaut, à des lignées phylogénétiquement proches. Le protocole permet ainsi d'identifier, à partir du transcriptome de l'holobionte, quatre catégories de séquences (Figure 4.1). Notre collaboration avec l'équipe de Rennes a permis la mise en place d'une version personnalisée de `SRC_c` (disponible sur le dépôt GitHub, https://github.com/GATB/short_read_connector) dans laquelle seules les informations d'abondance des k-mers d'un premier jeu dans un second jeu sont conservées et les informations de positions sont omises. Cette modification permet de réduire la quantité de données stockées et d'accélérer procédure pour chaque analyse de `SRC_c` et est adaptée à notre problématique de détection des hôtes et des symbiontes.

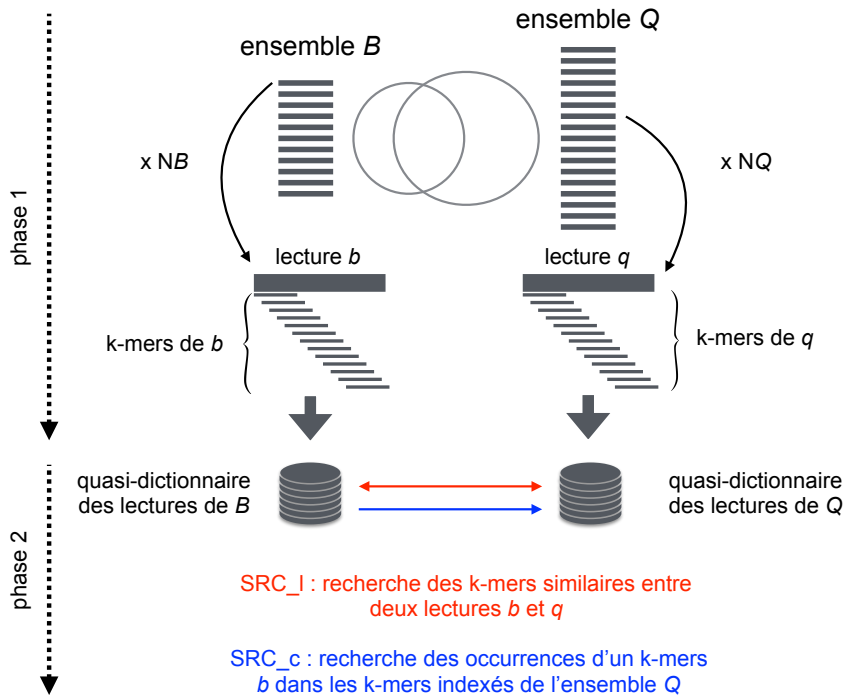


FIGURE 4.2 – Représentation schématique d'une comparaison de deux ensembles de lectures B et Q avec `SRC linker` (rouge) et `counter` (bleu). La phase commune aux deux versions de `SRC` (phase 1) ainsi que la phase spécifique à chaque version (phase 2) sont indiquées. Les deux ensembles de lectures B et Q sont respectivement composés de NB et NQ lectures.

4.1.2 Évaluation de `SRC_c` avec des données contrôlées

Afin d'évaluer `SRC_c` à trier des séquences d'holobionte, un transcriptome artificiel d'holobionte Rhizaria-dinoflagellés a été créé. Il s'agit d'un jeu de données créé à partir d'échantillons de lectures de plusieurs jeux de données de RNA-seq de Rhizaria et de dinoflagellés. J'ai utilisé 4 échantillons de 1 million de lectures chacun provenant d'un transcriptome de Rhizaria non-symbiotique (*Protocystis ornithocephala*, Cercozoa, Phaeodaria) et de trois transcriptomes de dinoflagellés non-symbiotiques (*Alexandrium fundyense*, *Prorocentrum lima* et *Heterocapsa triquetra*, respectivement notés SYMB1, SYMB2, SYMB3 (Figure 4.3)). Les 4 échantillons formant un total de 4 millions de lectures ont été groupés en un seul fichier et les lectures ont été mélangées. Deux banques de référence ont été créées : l'une composée des 1 mil-

lions de lectures de Rhizaria incluses dans le transcriptome d'holobionte artificiel, l'autre composée des 4 millions de lectures de dinoflagellés également incluses dans le transcriptome artificiel. Deux analyses indépendantes de `SRC_c` ont été menées pour trouver les similarités entre les lectures du transcriptome artificiel et (1) la banque de lectures de référence de Rhizaria et (2) la banque de lectures de référence de dinoflagellés. Les proportions de lectures du transcriptome artificiel similaires aux banques de référence ont été calculées, ainsi qu'une mesure de précision pour chacune des analyses. La mesure de précision correspond au pourcentage de lectures du transcriptome artificiel correctement assignées à chacune des banques (Figure 4.3).

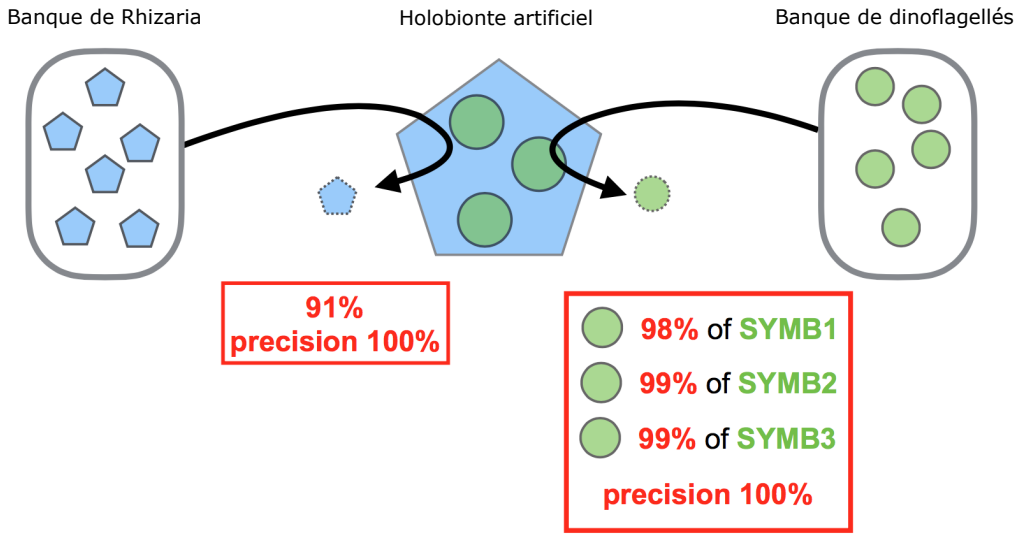


FIGURE 4.3 – Test de `SRC_c` sur un transcriptome d'holobionte artificiel. Les pourcentages de lectures de l'holobionte artificiel retrouvés dans la banque de Rhizaria d'une part et de dinoflagellés d'autre part sont indiqués dans les encadrés rouges, ainsi que les mesures de précision. Par exemple, 91% des reads de Rhizaria de l'holobionte trouvent des séquences similaires dans la banque de référence Rhizaria. Figure tirée de ma communication orale dans le cadre du congrès *Holobiont* au Muséum National d'Histoire Naturelle en mai 2017, (Annexe D)

Cette analyse démontre que `SRC_c` est capable de retrouver les similarités des lectures au sein d'un transcriptome d'holobionte artificiel et ce avec une précision de 100%. Néanmoins, cette évaluation de `SRC_c` n'est valable que dans le cadre d'une analyse où les contenus du transcriptome d'holobionte ainsi que des banques de

référence sont contrôlés. Pour valider cette stratégie et l'utilisation de `SRC_c` dans le cadre d'études de transcriptomes d'holobiontes, l'analyse des jeux de données publiés dans la littérature a été mise en oeuvre.

4.1.3 Mise en application de l'approche `SRC_c` avec des données réelles (analyse de transcriptomes d'holobiontes)

Il existe des modèles de symbioses marines dont les partenaires ainsi que leur interactions sont relativement bien définis. J'ai choisi deux modèles de symbiose marine parmi les plus étudiés et dont l'assemblage et l'analyse du transcriptome ont été menés dans des études antérieures. Le premier modèle sélectionné correspond à une symbiose corallienne entre deux lignées d'eucaryotes : un corail de l'espèce *Orbicella faveolata* (Metazoa, Cnidaria, Scleractinia - eucaryote multicellulaire) vivant en symbiose avec des dinoflagellés appartenant au genre *Symbiodinium* spp. (Dinophyta, Suessiales - eucaryotes unicellulaires) [Pinzón et al. 2015]. Le deuxième modèle est une symbiose eucaryote-procaryote réunissant une éponge de l'espèce *Xestospongia muta* (Metazoa, Porifera - eucaryote multicellulaire) et un mélange de bactéries symbiotiques qui ne sont pas clairement identifiées [Fiore et al. 2015]. Un troisième modèle est une symbiose entre un radiolaire et des microalgues dinoflagellés. L'objectif de ce troisième modèle est d'appliquer notre approche dans le cadre d'une symbiose encore non étudiée incluant deux lignées d'organismes (eucaryotes unicellulaires) non-modèles*. Les résultats obtenus sont présentés dans l'article ci-dessous.

A *de novo* approach to disentangle/decouple partner identity and function in holobiont systems

List of authors

Arnaud Meng[1*†], Camille Marchet[2†], Erwan Corre[3], Pierre Peterlongo[2], Adriana Alberti[4,5], Corinne Da Silva[4,5], Patrick Wincker[4,5], Eric Pelletier[4,5], Ian Probert[6], Johan Decelle[7], Stéphane Le Crom[1], Fabrice Not[6] and Lucie Bittner[1*]

* Correspondence:
arnaud.meng@gmail.com; lucie.bittner@upmc.fr
1 Institut de Biologie Paris Seine, University Pierre and Marie Curie, Quai Saint Bernard, 75005 Paris, France
Full list of author information is available at the end of the article
† Equal contributor

Keywords

holobiont; transcriptomic; *de novo* assembly; marine plankton

Background

In its scientific acceptation, symbiosis is defined as the living together of unlike organisms whatever the nature of their relationship [1], going from parasitism to mutualism. Symbiosis is a widespread phenomenon in the biosphere and plays crucial roles in evolution and ecology. One of the most popular examples of mutualism is the interaction between fungi and land plants, where fungi form mycorrhizae that help land plants to retrieve nutrients from soil [2]. In the ocean, benthic coastal ecosystems can be structured and supported by symbiotic associations involving multipartners such as corals (Cnidaria, i.e. multicellular eukaryotes), microalgae (Dinophyceae, *Symbiodinium*, i.e. unicellular eukaryotes), and Bacteria. Breakdown of this symbiosis ultimately leads to coral bleaching, dramatically affecting the whole reef ecosystems [3]. While coral bleaching has been largely studied, numerous symbiotic associations, involving other partners, also contribute to make coral reef persisting in oligotrophic seas. For instance, symbiotic association between sponges (Porifera, i.e. multicellular eukaryotes) and Bacteria (prokaryotes) allows Bacteria to grow within the mesohyl matrix of the sponge where they can be metabolically active and persist in a highly

oligotrophic habitat. The symbiotic interactions between sponges and bacteria are currently poorly understood from the genomic point of view [4]. Symbiotic associations involving two unicellular eukaryotes are also widespread in the oceanic plankton [5–7]. For instance, the cosmopolitan mutualistic associations between heterotroph Radiolaria (host) and endosymbiotic microalgae play significant ecological and biogeochemical roles in the oceans [8] but the underlying genomic basis of such associations remained uncharacterized. Although not cultivable *in vitro,* nucleic acids extraction is nevertheless possible on such symbiotic partnerships, and this recently allowed shedding light on the identity of the partners and their co-evolutionary history [6, 7]. Although several microalgal taxa have been identified most of the symbiont belong to the Dinophyta [9]. The lack of reference genomes for both Dinophyta and Radiolaria make their study challenging for *de novo* assembly and functional annotation [10, 11]. The study of the RNA mixture from a holobiont system, being composed of the host and its symbiotic microbial communities offers the opportunity to characterized functional aspects through their expressed genes, and so in different abiotic conditions/decoupling the functional/metabolic role of each partner.

Currently, RNA-seq approaches are the best available tools to obtain large amount of genomic information for uncultured organisms isolated from the environment [12, 13]. RNA sequencing for a whole holobiont is now possible [14–16] and has promoted the development of sequencing projects [17] for non-model organisms. Non-model holobiont RNA-seq datasets corresponds to a mixture of data coming simultaneously from the host and from the symbiont(s). studying such datasets share similarities with meta-transcriptomics and requires *de novo* assembly of transcripts sequences, which implies large computational resources and has the potential to introduce biases such as generating numerous chimeric sequences resulting from the mis-assembly of RNA fragments from the host and from the symbiont(s) [18, 19]. A variety of analysis strategies has been developed to address meta-transcriptomic challenges. Some of these strategies avoid the assembly step to focus on identifying abundant species and significant functional differences between meta-transcriptomes directly from raw data [20, 21]. Other strategies use statistical tools and machine learning algorithms

to improve the quality of *de novo* assembly of meta-transcriptome by learning from their abundance information [22].

Here we developed an original strategy aiming at improving *de novo* assembly for newly generated holobiont sequence dataset. We chose to use the Short Read Connector software in its Counter version (SRC_c) [23]. SRC_c is a fast kmer-based method initially developed to estimate the similarity between numerous (meta)genomic datasets by extracting their common sequences. We focused on holobiont transcriptomes for which *a priori* no or little genomic knowledge has been previously produced for host and symbionts, and we used SRC_c to compare these holobiont sequences to publicly available databases. Our strategy is to use SRC_c to assign at best holobiont sequences either to the host or to the symbionts before the *de novo* assembly step (Fig. 1). It allows then independent assembly of the datasets and prevents the potential mis-assemblies of reads from diverse origin.

As a proof of concept, we analyzed two public and one unpublished holobiont transcriptomes datasets, and we compared quantitatively and qualitatively results obtained when involving SRC_c or not.

Results

Choice of holobiont models and building of host and symbiont reference libraries

We applied our strategy to disentangle the sequences and then *de novo* assemble the transcriptome of three distinct marine holobiont systems (Fig 2). Two of them were already assembled and published. The first model (M1) involves a Cnidaria host (*Orbicella faveolata*, belonging to the Metazoa) and Dinophyta symbionts (*Symbiodinium* spp., a unicellular eukaryote belonging to the Alveolata) forming a mutualistic association via nutrient exchanges [24, 25]. This symbiotic association represents likely the best-known example of symbiosis in marine ecosystems, and many studies have been made trying to understand coral bleaching events (*i.e.* the loss of symbionts) [26, 27]. The coral holobiont also encompass other microorganisms consisting of bacteria, archaea, fungi, viruses [28, 29]. In the second holobiont model (M2) the marine sponge *Xestospongia muta*

(Porifera) harbors a dense (~40% of its volume) and diverse microbial community including marine protists (*e.g.* fungi), archaea and mainly bacteria [30–32]. The symbiotic associations between sponges and bacteria (suggested to be commensalism [33]) have become a major research focus to understand how sponges and their microbial communities can perform a variety of functional roles such as nutrition, cycling of metabolites and host defense allowing them to proliferate in oligotrophic conditions [34, 35]. We chose a third, yet unpublished, holobiont dataset (M3) involving two distinct lineages of protists (unicellular eukaryotes): the radiolarian *Collozoum* sp. as host and Dinophyta symbionts belonging to the *Brandtodinium nutricula* species [6]). In this association, the radiolarian host forms a gelatinous matrix of several centimeters, which contains hundreds of host cells and thousands of symbiotic microalgae (refer to image). Recent studies showed that this symbiosis is widely distributed in the ocean and significantly contribute to biomass and carbon export in the open ocean [36, 37].

For each of the three holobiont models (Fig. 2), we built reference sequences libraries representing host and symbiont(s) by selecting the taxonomically closest organisms available in public datasets (see Methods, Additional files 1). The M1 host reference library encompasses 22 assembled transcriptomes from Cnidaria (including data from the host species *Orbicella faveolata* itself) and the M1 symbiont reference library encompasses 123 RNA-seq reads datasets (including the presumed major symbiont *Symbiodinium* spp. [38]). The M2 host reference library involves 4 RNA-seq reads datasets from distinct Porifera genera (and differ from the *Xestospongia* genus) whereas the M2 symbiont reference library corresponds to the *Tara* Oceans metagenomic gene catalogue (OM-RGC) assembled from the pico-planktonic fractions (< 3 µm) including bacteria or Archaea [39]. For M3, we used the four Rhizaria transcriptomes published so far to create the reference host library whereas the same library as for M2 has been used for symbiont references. All reference libraries described above include assembled transcriptomes, genomes or RNA-seq raw reads datasets for eukaryotic or prokaryotic holobiont partners (Additional files 1). Their sizes vary from 4.5 Mbp to 25 Gbp with sequences length from 100 bp to 84 Kbp (Additional files 1).

Disentangling the holobiont sequences

Disentangling the holobiont sequences for all three models (M1, M2 and M3), the SRC_c memory footprint was far lower than our cluster's capacity (Tab. 1), even for the biggest data set to index (M2 symbiont library of 25 Gbp has been built with 58.9G of RAM). This induces that any addition of data can be considered.

The comparison of holobiont reads to reference host and symbiont sequence libraries enabled to identify and classify them into four categories (Fig. 1): (1) reads specific to the host, (2) reads specific to the symbionts (including microalgae, bacteria….), (3) reads which can be assigned to both reference libraries and (4) reads which do not match any reference library (refereed as 'unassigned'). For the three holobiont models, the distribution within the four categories is reported in Tab. 2.

With M1, SRC_c assigned 64.3% of the holobiont reads to the cnidarian host and 7.2% to the Dinophyta symbiont full library (analysis M1a, Tab. 2). Restricting the symbiont library to the genus *Symbiodinium* spp. sequences allowed obtaining similar results with 64.5% of the reads identified as specific to the host library and 7.1% as specific to the symbiont library (analysis M1b, Tab. 2). On the contrary, when *Symbiodinium* spp. is removed from the library, only 0.6% of the holobiont reads could be assigned to the symbionts and the proportion of reads assigned to the host increases up to 67.3% (analysis M1c, Tab. 2). Our tests on the symbionts library showed that the library content impacted drastically the reads retrieval by SRC_c and demonstrated the sensitivity of the strategy. Considering these results, we focused on the M1a dataset for downstream analyses. We also noticed that shared reads (i.e. found in both host and symbiont libraries) always represent the lowest proportion of holobiont reads (M1a, M2 and M3).

De novo assembly, contigs evaluation and downstream analyses for M1 and M2

For each holobiont transcriptome, four subsets of reads were independently *de novo* assembled, producing contigs from which protein domains were then predicted and functionally annotated (Fig. 1). For holobiont models M1a and M2, the assembly metrics, statistics and functional annotations

from our contigs are summarized in Tab. 3, and comparison with previous studies are shown in Fig. 3.

Compared to the studies where these datasets were initially published, our strategy allows considering more reads (16,818,599 reads for M2) in the assembly step as well as obtaining more assembled contigs (136,039 contigs for M1a and 78,567 contigs for M2) (Fig. 3). The contigs metrics show shorter lengths of N50 (580 bp shorter for M1a and 219 bp shorter for M2) (Fig. 3) compared to the original publication analyses. The M1a contigs display high remapping rates (>80%) while M2 contigs show mixed results (25% < x < 86%) (Tab. 3). With M1a, a total of 255,223 protein coding domains were predicted for 44.1% of the assembled contigs and functional annotations were found for nearly 30% of these protein coding domains (Tab. 3). With M2, protein coding domains were predicted for 39.6% of the contigs, and 54.9% of the domains were functionally annotated (Tab. 3). In comparison with statistics available in previous studies, we obtained 1.6 times more functionally annotated contigs for M1a (Fig. 3). This comparison for M2 could not be made since the exact number of annotated contigs in the holobiont assembly has not been reported by the authors.

To further test the usefulness of the reads sorting before the *de novo* assembly step, we compared the contigs assignment of M1a and M2 (column 1 in Tab. 3) with a taxonomic assignment performed with MEGAN6 [40]. For M1a, MEGAN6 assigned 71,143 contigs to the host *Orbicella faveolata* and 148,409 contigs to the symbiont *Symbiodinium* spp. (Additional files 2). All the contigs assigned to *Orbicella faveolata* with MEGAN6 were also found with the SRC_c strategy (Tab. 3) but we assigned 19,415 more contigs to the host category. On the contrary, MEGAN6 assigned 21,197 additional contigs to *Symbiodinium* spp. compared to our categorization strategy (Tab. 3, Additional files 2). With M2, MEGAN6 assigned 11 contigs to the host *Xestospongia muta* (Additional files 2) which is far less than the 2,654 contigs defined with the SRC_c strategy (Tab. 3). However, MEGAN6 assigned also 33,810 contigs to *Amphimedon queenslandica,* a distinct sponge species which is not supposed to be the host in this holobiont system. MEGAN6 also succeeded to assign more contigs to Bacteria (21,318 contigs) than the SRC_c strategy (2,431 contigs) (Tab. 3).

Our functional annotations were compared to initial studies having generated these datasets. As previous publications do not provide exhaustive lists of the functional annotations and their corresponding abundance, these comparisons are essentially qualitative. For the *O. faveolata* host (M1), we only found similarities in the most abundant annotations (Additional file 3). At biological processes level, both our study and Pinzón et al. 2015 found abundant metabolic process GO term (GO:0008152; 819 CDs (coding sequences) and 5,278 genes respectively). At the molecular function level, our host contigs mainly corresponded to binding protein (GO:0005515; 36,349 CDs) while Pinzón et al. 2015 mainly found catalytic activity functions (GO:0003824; 3,361 genes). For M2, rare overlaps are found between Fiore et al. 2015 and our annotations (Additional file 3): at the biological processes level, 1 of the top 15 host annotations is identical (signal transduction (GO:0007165)) and 3 of the top 15 symbiont annotations are in common (metabolic process (GO:0008152); proton transport (GO:0015992) and protein folding (GO:0006457)).

Benchmark comparisons on M3: what difference does it make to use SRC_c?

For the holobiont model M3, assembly metrics, abundance of chimera and functional contents were compared between the SRC_c contig sets (host, symbiont, shared and unassigned) and a direct *de novo* assembled transcriptome obtained from holobiont reads considered all together (this strategy is hereafter called *noSRC*).

The assembly metrics appear very similar between SRC and noSCR (Tab. 4). A comparable number of reads were used for the assembly step and a comparable number of assembled contigs were obtained. The N50 value for the *noSRC* strategy is slightly longer while the remapping rates are 5% better with the SRC strategy. Calculation times performed on the same bioinformatic cluster revealed that the SRC strategy was 40 hours longer. The SRC strategy showed 50% less chimeras (418 contigs) than the *noSRC* strategy (777 contigs) with most chimeras contained in the unassigned set (Tab. 4). We noticed slightly less annotated CDs with the SRC strategy (45,768 against 47,260 (data not shown)), however the number and the composition in GO annotations were very similar (Fig 1 from Additional files 4). We found 253 different biological processes with SRC against 255 with the *noSRC* strategy, and the top 5 functional annotations in the 3 Gene Ontology

levels (Molecular Function, Biological Process and Cellular Component) are strictly identical (Fig 2 from Additional files 4). Considering all GO annotations, 686 are common to both strategies while 52 are exclusive to the SRC strategy and 42 to the *noSRC* strategy (Fig 3 from Additional files 4). To test the usefulness of the categorization step, all M3 contigs from the SRC strategy were taxonomically assigned using MEGAN6 (Additional files 5). MEGAN6 assigned 10 contigs to Collodaria whereas the SRC strategy assigned 683 contigs to the host category. MEGAN6 assigned 1,383 contigs to Dinophyceae compared to the 5,207 contigs categorized as symbionts. The leftover MEGAN6 contigs were assigned to Bacteria and Archeae (3,799 contigs), Viruses (76 contigs), other-eukaryotes (29,524 contigs) and 127,447 contigs remained unassigned (162,947 unassigned contigs with the categorization strategy).

Discussions

The use of SRC_c to tackle meta-transcriptomic challenges

The strategy proposed here is a practical and scalable solution for transcriptomic assembly of non-model holobiont organisms, from which no or limited genomic information is available. The present implementation of SRC_c [23] based on reference databases of putative partners involved in the holobiont consortium, and our analysis strategy, which includes de novo assembly, protein domain prediction and functional annotation, enabled the categorization of holobiont reads into 4 subsets. These subsets can then be independently assembled, limiting potential creation of chimeras while generating more assembled contigs (Fig. 1). The newly defined shared reads category represents an added value compared to other holobiont transcriptomic studies and has been later processed with the same methodology than other categories (Fig. 1).

With respect to the reference libraries, with M1, when the expected symbiotic partner (i.e. *Symbiodinium* spp.) is missing from the reference library, the number of reads assigned to the symbiont category decreases drastically from 50M reads to nearly 5M reads (Tab. 2). In addition, the M2 and M3 libraries do not contain reference data for the expected host partner, and

consequently only a low proportion of the holobiont reads are assigned to the host (19% and 3%, respectively). Accordingly, to this observation, the proportion of unassigned reads is directly linked to both host and symbiont libraries content with respect to the studied holobiont: less unassigned reads were observed when the “correct” actors are involved (M1a: 24.4%) compared to the poorly studied models (M2: 61.6% and M3: 72.5%)

These results highlight the sensitivity and specificity of the SRC_c requests that relies on the completeness of the database to accurately sort the reads of the holobiont. The SRC_c assignation step could be further improved for all models by adding more sequences (i.e. reads, assembled genes or transcripts) from taxonomically close species to the host and symbiont reference libraries, but also from parasites and viruses that are common in multicellular and unicellular host cells.

And can be compared to a metabarcoding approach for correction and improvement of reference databases prior to SRC_c

We also compared the metrics of our SRC_c contigs to those from previous studies (M1a and M2) [24, 30]. With the SRC_c strategy, the amount of reads used for *de novo* assembly of M2 was higher than for previous studies (Fig. 3). We found that, not only our strategy allowed defining a new category of contigs (the “shared” contigs), but also allowed assembling more contigs than previous studies (Fig. 3). Our contigs metrics showed lower N50 for both models compared to previous studies, but showed remapping rates overall for M1a (up to 90%, (Tab. 3)). Differences of the number of contigs as well as contigs metrics can be due to the use of distinct *de novo* assembly software: e.g. M2 data were processed with the CLC workbench [CLC bio, Boston, MA, USA; (https://www.qiagenbioinformatics.com/)] in the original publication while we choose the Trinity software [41] and we suggest that SRC_c do not significantly impact transcriptome assembly. In fact, previous studies had shown that Trinity is able to generate more assembled contigs than the CLC assembler when applied on the same dataset. It is also known that assembled contigs from Trinity are shorter than those assembled by CLC but provided similar proportion of significant hits to the nr database [42].

With M1a, our strategy produced 1.5 times more CDs with a functional annotation (Fig. 3). We can not exclude that this observation can be the consequence of a better suited assembly strategy (SRC_c treatment and / or assembly software), and / or the use of a different annotation pipeline, and / or the supplementation of reference annotation databases between 2015 [24] and now.

With M3 analyses we can estimate how SRC_c impacts the *de novo* assembly step and downstream analyses compared to a more conventional protocol (here called the *noSRC* strategy) (Tab. 4). Minimal differences were found between the two protocols concerning the number of assembled contigs and, as for M1a and M2, the SRC_c strategy produces shorter contigs sequences with higher remapping rates. However a significant diminution of the number of potential chimeras was observed. We conclude that the read assignation performed before the assembly step largely contributes to limit the production of chimeras. This shows that the use of SRC_c impacts the *de novo* assembled transcriptome quality and contributes to address one of the *de novo* assembly challenge yet unsolved [43]. The MEGAN6 contigs assignation from M2 shows more contigs than SRC_c could assign to host and symbiont (Tab. 3 and Additional files 2). In contrast, MEGAN6 assigned less contigs to host and symbiont than SRC_c for the M3. We suggest that (1) SRC_c required libraries containing close organisms reference sequences and that (2) SRC_c performs well in non-model context for which reference sequences are missing in public databases. Finally, the calculation time for the two protocols showed that the SRC_c strategy increases the total time with nearly 40 additional hours compared to a classic assembly strategy (Tab. 4). However, compared to classic strategies, the SRC_c strategy has the tremendous benefit to create directly 4 independent subsets (two of which are directly assigned to holobionts partners).

Does SRC_c help us to make new biological assumptions?

For all models, the SRC_c strategy leads to a higher number of annotated contigs, however as only partial information on the annotation content were provided for only the host or the symbionts in previous publications [24, 30], we were mainly restricted to qualitative comparisons.

Comparing the M1a host transcriptomes to the previous study transcriptome, very few similarities were found for the most occurring functions, even if the most annotated function is common (i.e. metabolic process GO). Our 20 most occurring functions include signal transduction functions (14% of the total annotations) and molecule transport functions (8% of the total annotations) that do not appear in the most occurring function from [24]. These newly highlighted functions could help in better understanding the *Orbicella faveolata* host with respect to communication and cellular exchanges with its partners. We were not able to perform a similar analysis for the symbiont transcriptome since authors of previous studies focused on the host transcriptome. For M2, only 1/15 and 3/15 common annotations for host and symbiont respectively, were found. We suggest that the divergences in terms of analyses pipeline (Trinity versus CLC for *de novo* assembly, followed by InterProScan versus FastAnnotator for functional annotation) make the functional annotations contents hard to compare between the studies. Even though, both studies differ in term of methodology to assemble and annotate *de novo* transcriptome, both results must be considered as potentially valuable and must be verify with genome alignment when available or through *in sillico* validation for restricted group of functions (*e.g.* PCR).

In recent years, new cell-cell holobionts, symbioses involving heterotrophic hosts and photosynthetic symbionts have been described in the oceanic plankton using morphological and molecular data [5–7, 15]. Radiolarians and their microalgae (e.g. Haptophytes, Dinoflagellates) have a significant ecological and biogeochemical importance [44–47], but little is known about symbiosis establishment and maintenance conditions. Moreover, while microalgal lineages can be easily grown in a laboratory [Meng et al. *submitted*], the study of radiolarians can only rely on single-cell isolation on the field [36, 48]. In this study, the radiolarian host belongs to the Collodaria order which is ubiquitous and abundant in the open ocean, and which has an important role as active predator and host of symbiotic microalgae [36, 49]. Our knowledge about their ecology and evolution is limited and hence our analyses represent an opportunity to learn more about the genetic repertoire of such uncultivable, non-model lineage. Regarding functional annotations, the SRC and the *noSRC* strategies provided very similar results but the SRC strategy categorized the GO annotations among

4 subsets (host, symbiont, shared and unassigned) (Additional files 5), which can be explored independently, allowing group specific interpretations and biological hypothesis building for each partner from the holobiont. For instance, in the M3, we have detected symbiont CDs linked to the photosystem I and II processes which confirmed that SRC_c succeeded to assign reads to the good actors in the holobiont system, here the photosynthetic symbionts in which functions linked to photosystem were expected (Additional files 4).

Strategies regarding the use of SRC_c and future perspectives

SRC_c successfully compared different holobiont read sets to huge reference libraries in a short amount of time (less than 24h), with reasonable computational resources (i.e. 10 CPUs and less than 20Go of RAM). By setting parameters, we could make SRC adaptable to heterogeneous nature of sequences in libraries (*i.e.* length, row reads or assembled genes/transcripts, data volume, k-mers distribution). When studying reads, selecting the abundant k-mers helps to remove the one corresponding potentially to sequencing errors; however rare sequence k-mers are consequently lost. On the contrary, when indexing assembled genomes or transcriptomes, we do not expect a redundancy of the k-mers such as in high-throughput sequencing experiments, and we assume that any k-mer is relevant when it comes from a reference sequence. A solidity threshold of 1 is in this case recommended in order to keep all the k-mers. Hence in this study, we kept the default k-mer solidity threshold value that was appropriate when indexing reads (i.e. sequences shorter than 300 bp, with a relatively high coverage), and lowered it to 1 when indexing longer sequences as ESTs or assembled genes.

Due to the presence of small reads (50 bp) in our holobiont datasets, we modified the default k-mer size value of 31 to a value of 25, so that any read contains at least a few k-mers. A read smaller than the k-mer size could not be retrieved by our approach. Usually the k-mer size is higher [50], however 25 base pairs corresponds to an descent value to ensure the uniqueness of the read [51]. During the query phase of SRC_c, a query sequence (from a dataset Q) must contain at least *s*% positions covered by at least one indexed k-mers (from a dataset B), to be considered similar to data from the set B. We set this value *s* to 50%. This means a read of size l should have at least l×*s*

positions covered by (overlapping or nonoverlapping) indexed k-mers. For instance, when a large majority of the reads could not be assigned, our strategy was to decrease this parameter *s* from 50 to 40 in order to increase the quantity of reads recalled. For M2 we set s to 40 and for other models we used *s*=50.

SRC_c implements a heuristic computing a k-mer based similarity. Contrary to BLAST-like methods, SRC_c relies uniquely on shared k-mers for its similarity computation. It means that a certain amount of error-free k-mers (i.e. k-mers that do not contain sequencing errors) must be found in common in order to output sequences, which can make SRC_c less sensitive compared to alignment methods which authorize mismatches. However contrary to alignment methods, SRC_c was tailored to scale to very high-volume data sets and comparisons presented in [23] showed that SRC_c could handle sets of order of magnitudes higher volumes than BLAST (Additional files 7). SRC_c's efficiency relies on its particular probabilistic data structure. The lightweight indexing and query of k-mers is made at the price of rare false positives. In our case, false positives correspond to k-mers that are not contained in the original indexed library. Such a false positive rate is controled and low (Additional files 7). As in this work, the k-mer size was relatively low (i.e. 25), the default value for this parameter was kept ensuring a low false positives rate. For longer k-mers (sizes > 31), we recommend to increase the size of the fingerprint if more precision is needed. SRC_c can also be used in a no-false positive mode that requires more memory, but that is still less costly than a hash table as demonstrated in [23]. In our experiments, SRC_c helps to retrieve holobiont reads similar to host or symbiont close species. Previous tools like COMMET [50] already proposed such computation, although their data structure makes difficult the use of k-mers of small size, as computation time would be drastically impacted. SRC_c was chosen for its simple output and its adaptability to the heterogeneous nature of the libraries studied. This is simply made by adapting the k-mer lowest occurrence and size parameters.

Future works could include more extensive exploration of the impact of the similarity threshold parameter on the sensitivity of our approach. In this regard, if the reads similarity rate to the libraries could be relaxed, it may decrease the number of unassigned reads in particular for poorly studied

models. A second strategy to explore in the near future would be to implement an iterative enriching strategy to maximize the proportion of holobiont reads assigned to the host or to the symbiont. This strategy can allow to assign more sequences in the case of non-model organisms. After a first assignment round with SRC_c, holobiont reads linked to an identified group (host/symbiont) can be added to the reference libraries. Then, based on these new enriched libraries, a second run of SRC_c can be performed on the holobiont reads. This can be implemented as an iterative pipeline: at each round, more reads will be assigned to the host or symbiont categories and will then be used as reference libraries.

Finally, the approach proposed here has been applied to holobiont systems (between 2 partners) but it could be used to address larger metatranscriptomic datasets composed of more complex assemblages. Depending on the SRC_c library content, the user can choose to target either one or more specific species among the variety that composed such metatranscriptomic datasets. Coupled to our assembly and downstream analysis strategy, the subsets resulting of the used of SRC_c are processed *de novo* allowing the potential discovery of newly assembled transcripts and the exploration of the functional their functional feature contents without reference genome.

Conclusions

SRC_c successfully processed a variety of large-scale datasets and offered a pragmatic way to classify before assembly sequences from the different holobiont partners. Moreover, we showed that our strategy allows to improve some assembly metrics, and also helped to reduce drastically the proportion of chimeras in the newly *de novo* assembled sequences. Our strategy offers an efficient, large scale, comparison strategy to assemble and study holobionts involving non-model organisms. This de novo approach, allowing a taxonomic categorization of functionalities, can reveal and link both identity and function, which is necessary to better understand the functioning and contribution of each partners in holobiont systems.

Methods

Radiolaria-Dinophyta holobiont model (M3) sampling, RNA-seq library and sequencing

The Collodaria colony was sampled in the south Pacific Ocean at the station 112.01 (coordinates in decimal degrees: latitude -23.3, longitude -133.9) during the *Tara* Oceans expedition in 2011 [52]. The radiolarian colony of few centimeters diameter was collected *in situ* at the subsurface (1m deep) with a plastic jar, preventing disruption of the colony and aggregation of other planktonic organisms. Live observations through the binocular were performed to verify that no organisms were accidentally attached to the colony before preservation. The collected colony was directly isolated in 15 mL of RNAlater (ThermoFisher Scientific, Waltham, MA) and preserved at -20°C. Total RNA extraction was performed using NucleoSpin RNA kit (Macherey-Nagel, Düren, Germany) starting from a slice (about 1 cm diameter) of Collodaria PAC 37 colony. Briefly, frozen cells were transferred in a 1.5 mL tube containing 100 µL RA1 lysis buffer and grinded for 1 min with a motor driven pellet pestle previously refrigerated in liquid nitrogen. Then 250 µL RA1 lysis buffer, previously mixed with 3,5 µL β-mercaptoethanol (1% of total RA1 volume), were added to the lysed cells and the total volume was transferred to a Nucleospin filter. After centrifugation and addition of an equal volume of 70% ethanol, the RNA was purified following the manufacturer's instructions and finally eluted in 40 µL nuclease-free water. Quantity and quality of extracted RNA were assessed by capillary electrophoresis on an Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA).

Finally, in order to reduce as far as possible the risk of residual genomic DNA, a further DNase treatment was applied on the total RNA using Turbo DNA-free kit (Thermo Fisher Scientific), according to the manufacturer's protocol. After purification with the RNA Clean and Concentrator-5 kit (ZymoResearch, Irvine, CA), RNA was eluted in 10 µL nuclease-free water and used to synthetize cDNA with the Ovation RNA-seq System Version 2 (NuGEN, San Carlos, CA), following the manufacturer's protocol. After cDNA shearing by Covaris E210 instrument (Covaris, Woburn, MA), Illumina library was prepared using the SPRIWorks Library Preparation System on a SPRI TE instrument (Beckmann Coulter Genomics, Danvers, MA), according to the manufacturer's protocol without size selection. Ligation products were PCR-amplified using Illumina adapter-specific primers and Platinum Pfx DNA polymerase (ThermoFisher Scientific). After library profile analysis by Agilent 2100 Bioanalyzer and qPCR quantification (MxPro, Agilent Technologies), the library was

sequenced using 101 base-length read chemistry in a paired-end flow cell on HiSeq2000 Illumina sequencer (Illumina, San Diego, CA), in order to obtain nearly 50 million paired end reads.

Data retrieval and sequence libraries construction

For each holobiont model, sequence libraries were created based on published data from taxonomically close organisms to host and symbiont species. Detailed statistics of these reference libraries can be found in Additional files 1.

For the Cnidaria-Dinophyta holobiont model (M1), the host library includes 20 assembled transcriptomes (466,582 contigs) of cnidarian organisms [53] and 2 genome-derived ESTs (201,677 ESTs) of *Nematostella vectensis* and *Orbicella faveolata* [54]. The symbiont library is composed of 123 RNA-seq reads datasets (a total of 5,563,498,607 reads) of Dinophyta from the MMETSP project [55]. We built 3 versions of the symbiont reference library, one composed of all Dinophyta (M1 a), the second exclusively composed of *Symbiodinium* spp. (15 RNA-seq datasets, a total of 123,122,726 reads) (M1 b) and the third composed of all Dinophyta except *Symbiodinium* spp. (108 RNA-seq datasets, a total of 5,440,375,881 reads) (M1 c).

For the Porifera-Bacteria holobiont model (M2), 4 RNA-seq datasets of poriferan species were included in the host library (642,229,924 total reads): *Amphimedon queenslandica* [56] *Crella elegans* [57] and both *Haliclona amboinensis* and *Haliclona tubifera* [58]. The complete bacterial gene catalog (40,154,822 assembled gene sequences) derived from the first stations from the *Tara* Oceans expedition [39] has been downloaded to constitute the symbiont reference library (OM-RGC).

For the Radiolaria-Dinophyta holobiont model (M3), we gathered Rhizaria sequences from 4 *de novo* assembled holobionts: 7,215 presumed host transcripts were extracted among a total 15,404 *de novo* assembled transcripts [15]. Host specific sequences were extracted from holobionts assemblies removing first sequences from prokaryotic origin with a blastn (e-value 1e-3) against the OM-RGC database, and second, removing symbionts sequences with a blastx (e-value 1e-3)

against Dinophyta *de novo* assembled transcriptomes [Meng et al. *submitted*]. The exhaustive Dinophyta library created for the M1a was used for the reference symbiont library.

Comparing meta-transcriptomes (i.e. holobiont reads) to reference libraries using Short Read Counter (SRC_c)

> Presentation of SRC_c

Short Read Connector Counter (SRC_c) [23] relies on a very lightweight data structure called a quasi-dictionary that enables to work with voluminous sequence sets. The quasi-dictionary enables to associate a piece of information to any element from a static set composed of N distinct elements. It is composed of two parts: a minimal perfect hash function (MPHF) [59] and a fingerprint table. The MPHF allows to index very efficiently the elements of the set in memory, such that each element can be associated to any piece of information (*i.e.* k-mer coverage, location in reads, …). The fingerprint table is used to verify the membership of an element to the indexed set of elements using the MPHF. This way, stranger elements to the MPHF can be filtered out. The quasi-dictionary is a probabilistic structure with a controlled false positive rate that depends on the size of the fingerprint. SRC_c needs as input two sets of sequences (that can be identical). To compare sequences from a query set Q to those from a target set T, the set indexed in the quasi-dictionary is a set of k-mers from T. Finally, for each sequence *S* from Q, the number of k-mers of *S* shared with T provides a similarity measure of *S* with the set T. This implies that the similarity measure given is asymmetrical: it depends on the placement of the k-mers on the reads of Q, not of those of T. SRC_c is available at https://github.com/GATB/ short read connector, the commit 94aa6a65b5ddf61eba95108069fae29c41e51fb0 was used for this study.

> Application on data

In this study, SRC_c is used to assign reads from an holobiont transcriptome either to the host or to the symbionts. We divided the query of the holobiont data set Q in two parts, one that consists in the comparison of Q reads to a bank (i.e. reference library) of host sequences, and another that performs the comparison to a bank of symbiont sequences. The sets to index are composed of k-mers from the sequences. In each comparison, two sequence sets are considered. The whole

holobiont set Q and the target bank set B. First, the set B, which contains reads or assembled sequences and represents sequences close to the host (resp. symbiont), is indexed. During the indexation phase, the solid set of k-mers (i.e. the set composed of any k-mer which occurrence is above a user-fixed threshold (the solidity threshold) in the data set) from T is computed using the DSK [60] method. This set is next indexed in the quasi-dictionary previously described. Then the reads from the holobiont data set (Q) are queried. For each read, the query phase reports the abundance of its indexed k-mers. In the meantime, reads are checked to have enough positions (i.e. more than a given threshold which can be parameterized) for which an indexed k-mer starts over their length. This enables to add stringency to the query: a read that shares only a few k-mers with the index is considered not enough similar to the index. Finally, each read from Q (the holobiont) which was found similar to T (the host or the symbionts) during the query are returned in a binary vector and can be extracted to a FASTA format.

> Parameters choice

Parameters from SRC_c must be carefully chosen. First, the solidity threshold is adapted according to the nature of the sequences in the bank data set. For libraries which sequences are reads (symbiont libraries for model 1) the default value for the solidity threshold (= 2) was kept. For longer sequences (host libraries for model 1, sequences of models 2 and 3) the threshold was adapted and set to 1 when using libraries of assembled sequences or EST (host libraries for model 1, sequences of models 2 and 3). We chose a k-mer length of 25 according to the smaller input read length. We set the similarity value s to 50% for models 1 and 3, and decreased it to 40% for model 2. Both query and indexation phases are parallelized in SRC_c. For this study analyses were performed on a Linux system with 40 cores, with the option -t 0 (maximal number of available threads is used) and 250 GB of memory.

Read filtering, de novo assembly and downstream analysis

All read subsets resulting from the SRC_c step were first filtered (sequences trimming and cleaning) with the Trimmomatic program [61] (v0.36) and custom parameter SLIDINGWINDOW:10:20.

441 Filtered reads were assembled using the *de novo* transcriptome assembly program Trinity [41]
442 (v2.4.0) with default parameters. The newly assembled contigs metrics were calculated with the
443 Transrate program [62] (v1.0.3). Additional downstream analyses include protein coding domain
444 prediction using Transdecoder [63] (v3.0.1) and functional annotation with InterProScan 5 [64]
445 (v5.24-63), both with default parameters. The pipeline used for the steps described above is publicly
446 available on a GitHub repository https://github.com/arnaudmeng/dntap [53, Meng et al. *submitted*].

Taxonomic assignment with MEGAN6

447 The contigs sequences were compared to the nr database (August 2017 version) with the DIAMOND
448 software [66] (v0.28.22.84) using default parameters for BLASTx comparison and a e-value of $1e^{-3}$.
449 The resulting alignments were processed with the *daa2rma* tool script provided with MEGAN6 and
450 GeneInfo Identifier (GI) were mapped to alignments using the gi_taxid.bin file (version of May 2017).
451 Finally, taxonomic assignment has been calculated with default parameters using the MEGAN LCA
452 (Last Common Ancestor) algorithm and were visualized through the MEGAN6 software.

Chimeras identification

453 We followed the protocol described in [67]. 50,000 randomly sampled *de novo* assembled contigs
454 for the M3 (with the SRC strategy and without SRC strategy) were compared to the 7,215 Rhizaria
455 presumed contigs from [15] and 3,494,295 coding domains from *de novo* assembled contigs of 54
456 dinoflagellates transcriptomes [Meng et al. *submitted*]. The comparison was made using the
457 BLASTx program [68] (e-value $1e^{-3}$). The tools scripts *detect_chimera_from_blastx.py* from [67]
458 was applied to resulting alignments to detect potential chimeras.

Competing interests
The authors declare that they have no competing interests.

Author's contributions:
LB and AM designed the analysis, and LB guided the study.
IP, JD and FN performed sampling and culture steps.
AA and CDS optimized the molecular protocoles and performed the sequencing analysis.
AM and CM performed the computational analyses, with the help of PP and EC.
AM, CM, PP, SLC, FN and LB wrote the manuscript.
EP and PW provided critical discussions.
All authors read and approved the final manuscript.

Acknowledgements
We thank the RCC staff for providing the dinoflagellates cultures as well as ABIMS staff for the help on computational facilities. This work was supported by a 3-year Ph.D. grant from the "Interface pour le Vivant" (IPV) program at the University of Pierre et Marie Curie (UPMC), Paris, France. This project was supported by Région Ile-de-France.

Author details
1 Sorbonne Universités, UPMC Univ Paris 06, Univ Antilles, Univ Nice Sophia Antipolis, CNRS, Evolution Paris Seine - Institut de Biologie Paris Seine (EPS - IBPS), 75005 Paris, France .
2 Institut de Recherche en Informatique et Systèmes Aléatoires, INRIA, Campus de Beaulieu, 263 avenue du Généeral Leclerc, 35042 Rennes, France.
3 ABiMS, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France.
4 Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France.
5 UMR8030, CNRS, Evry, France.
6 UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France.
7 Helmholtz Centre for Environmental Research – UFZ, Department of Isotope Biogeochemistry, Permoserstraße 15, 04318 Leipzig, Germany. johan.decelle@ufz.de

Figure 1 Theoretical overview on the application of SRC_c on holobiont transcriptome. The comparisons to (1) host and (2) symbiont reads/sequences library are done against the entire holobiont dataset to retrieve host and symbiont similar reads. The 4 resulting subsets (host, symbiont, shared and unassigned reads) are then processed independently (de novo assembly and downstream analyses).

Figure 2 Pictures of the 3 holobiont models. (A) the *Orbicella faveolata* holobiont in symbiosis (unbleached) in 2010 at reefs of La Parguera, Puerto Rico (credits: [24]). (B) A *Xestospongia muta* specimen in symbiosis on a coral reef near Little Cayman in the Caribbean (credits: Cara Fiore, january 14, 2015 http://feedthedatamonster.com). (C) A Collodaria colony with symbionts sampled in South Pacific Ocean at station 112.01 of the Tara Pacific expedition in 2011 (credits: Johan Decelle).

Figure 3 Overview and comparison of the total assembled contigs for holobiont model M1a and M2 compared to the assembled meta-transcriptomes from (A) Pinzon et al. 2015 [24] and (B) Fiore et al. 2015 [30] respectively. General details about *de novo* assembly and functional annotation (termed FA) features are presented in corresponding tables for (A) holobiont model M1a versus Pinzon et al. 2015 [24] meta-transcriptome, and (B) holobiont model M2 versus Fiore et al. 2015 [30]. NC means that exact number is not communicated.

Table 1 Performances (memory peak and wallclock time) of SRC indexing and query steps on the several data sets for models 1 and 2.

Table 2 SRC_c assignment results for the Cnidaria-Dinophyta holobiont model (M1) against the complete Dinophyta library (M1a), the *Symbiodinium* spp. exclusive library (M1b) and the Dinophyta library excluding *Symbiodinium* spp. (M1c), the Porifera-Bacteria holobiont model (M2) and the Radiolaria-Dinophyta holobiont model (M3).

Table 3 *De novo assembly* metrics and downstream analysis of SRC_c resulting subsets for holobiont models M1a, M2 and M3.

Table 4 SRC_c impact on assembled contigs quality and calculation times of Radiolaria-Dinophyta holobiont model (M3) compared to a direct meta-transcriptome assembly strategy. In grey are displayed the details for SRC_c holobiont categories (host, symbiont, shared and unassigned). The "total" values for N50 and remapping rates of the SRC_c strategy were re-calculated on pooled contigs from host, symbiont, shared and unassigned subsets.

Additional Files

Additional file 1 SRC_c library content information and data sources. Table with detailed information of SRC_c libraries contents. The type of data and the total library sizes are displayed. It includes taxonomic contents and links to data repositories for holobiont models M1, M2 and M3 and data that constitute SRC_c reads/sequences libraries.

Additional file 2 Taxonomic assignment of SRC assembled contigs with MEGAN6 for the holobiont models M1 and M2.

Additional file 3 details of common GO annotations M1 and M2 our contigs versus previous studies

Additional file 4 Comparison of functional annotations between SRC assembled transcriptomes and a *de novo* assembled transcriptome without the use of SRC_c in the case of holobiont model M3. Details of the functional annotations results for the SRC strategy applied to M3, the tables displayed correspond to the top 15 GO annotations found in host, symbiont, shared and unassigned transcriptomes for the three levels of annotations (MF: Molecular Functions, BP: Biological Process and CC: Cellular Component).

Additional file 5 Radiolaria-Dinophyta meta-transcriptome taxonomic assignment with MEGAN6. Table of taxonomic assignation of the 167,023 *de novo* assembled contigs from the assembly without SRC reads sorting of the holobiont model M3.

Bibliography

1. De Bary A. De la symbiose. Rev Int Sci. 1879;3:301–9.

2. Selosse M-A, Strullu-Derrien C. Origins of the terrestrial flora: A symbiosis with fungi? BIO Web Conf. 2015;4:00009.

3. Davy SK, Allemand D, Weis VM. Cell Biology of Cnidarian-Dinoflagellate Symbiosis. Microbiol Mol Biol Rev MMBR. 2012;76:229–61.

4. Hentschel U, Usher KM, Taylor MW. Marine sponges as microbial fermenters. FEMS Microbiol Ecol. 2006;55:167–77.

5. Decelle J, Probert I, Bittner L, Desdevises Y, Colin S, Vargas C de, et al. An original mode of symbiosis in open ocean plankton. Proc Natl Acad Sci. 2012;109:18000–5.

6. Probert I, Siano R, Poirier C, Decelle J, Biard T, Tuji A, et al. Brandtodinium gen. nov. and B. nutricula comb. Nov. (Dinophyceae), a dinoflagellate commonly found in symbiosis with polycystine radiolarians. J Phycol. 2014;50:388–99.

7. Mordret S, Romac S, Henry N, Colin S, Carmichael M, Berney C, et al. The symbiotic life of Symbiodinium in the open ocean within a new species of calcifying ciliate (Tiarina sp.). ISME J. 2016;10:1424–36.

8. Decelle J, Siano R, Probert I, Poirier C, Not F. Multiple microalgal partners in symbiosis with the acantharian Acanthochiasma sp. (Radiolaria). Symbiosis. 2012;58:233–44.

9. Decelle J, Colin S, Foster RA. Photosymbiosis in Marine Planktonic Protists. In: Marine Protists. Springer, Tokyo; 2015. p. 465–500. doi:10.1007/978-4-431-55130-0_19.

10. Sibbald SJ, Archibald JM. More protist genomes needed. Nat Ecol Evol. 2017;1:0145.

11. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat Microbiol. 2016;1:nmicrobiol201648.

12. Reuter JA, Spacek DV, Snyder MP. High-Throughput Sequencing Technologies. Mol Cell. 2015;58:586–97.

13. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, et al. The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol. 2016;17:53.

14. Shinzato C, Inoue M, Kusakabe M. A Snapshot of a Coral "Holobiont": A Transcriptome Assembly of the Scleractinian Coral, Porites, Captures a Wide Variety of Genes from Both the Host and Symbiotic Zooxanthellae. PLOS ONE. 2014;9:e85182.

15. Balzano S, Corre E, Decelle J, Sierra R, Wincker P, Da Silva C, et al. Transcriptome analyses to investigate symbiotic relationships between marine protists. Microb Physiol Metab. 2015;6:98.

16. Daniels C, Baumgarten S, Yum LK, MIchell CT, Bayer T, Arif C, et al. Metatranscriptome analysis of the reef-building coral Orbicella faveolata indicates holobiont response to coral disease. Front Mar Sci. 2015;2. doi:10.3389/fmars.2015.00062.

17. Bashiardes S, Zilberman-Schapira G, Elinav E. Use of Metatranscriptomics in Microbiome Research. Bioinforma Biol Insights. 2016;10:19–25.

18. Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, et al. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol. 2014;15:553.

19. Sangwan N, Xia F, Gilbert JA. Recovering complete and draft population genomes from metagenome datasets. Microbiome. 2016;4:8.

20. Westreich ST, Korf I, Mills DA, Lemay DG. SAMSA: a comprehensive metatranscriptome analysis pipeline. BMC Bioinformatics. 2016;17:399.

21. Martinez X, Pozuelo M, Pascal V, Campos D, Gut I, Gut M, et al. MetaTrans: an open-source pipeline for metatranscriptomics. Sci Rep. 2016;6:srep26447.

22. Mohsen H, Tang H, Ye Y. Improving de novo metatranscriptome assembly via machine learning algorithms. Int J Comput Biol Drug Des. 2017;10:91–107.

23. Marchet C, Limasset A, Bittner L, Peterlongo P. A resource-frugal probabilistic dictionary and applications in (meta)genomics. ArXiv160508319 Cs Q-Bio. 2016. http://arxiv.org/abs/1605.08319. Accessed 27 Jul 2017.

24. Pinzón JH, Kamel B, Burge CA, Harvell CD, Medina M, Weil E, et al. Whole transcriptome analysis reveals changes in expression of immune-related genes during and after bleaching in a reef-building coral. R Soc Open Sci. 2015;2. doi:10.1098/rsos.140214.

25. Davy SK, Allemand D, Weis VM. Cell Biology of Cnidarian-Dinoflagellate Symbiosis. Microbiol Mol Biol Rev. 2012;76:229–61.

26. Hoegh-Guldberg O. Climate change, coral bleaching and the future of the world's coral reefs. Mar Freshw Res. 1999;50:839–66.

27. Muller-Parker G, D'Elia CF, Cook CB. Interactions Between Corals and Their Symbiotic Algae. In: Coral Reefs in the Anthropocene. Springer, Dordrecht; 2015. p. 99–116. doi:10.1007/978-94-017-7249-5_5.

28. Rohwer F, Seguritan V, Azam F, Knowlton N. Diversity and distribution of coral-associated bacteria. Mar Ecol Prog Ser. 2002;243:1–10.

29. Thompson JR, Rivera HE, Closek CJ, Medina M. Microbes in the coral holobiont: partners through evolution, development, and ecological interactions. Front Cell Infect Microbiol. 2015;4. doi:10.3389/fcimb.2014.00176.

30. Fiore CL, Labrie M, Jarett JK, Lesser MP. Transcriptional activity of the giant barrel sponge, Xestospongia muta Holobiont: molecular evidence for metabolic interchange. Front Microbiol. 2015;6. doi:10.3389/fmicb.2015.00364.

31. Webster NS, Taylor MW. Marine sponges and their microbial symbionts: love and other relationships. Environ Microbiol. 2012;14:335–46.

32. Simister RL, Deines P, Botté ES, Webster NS, Taylor MW. Sponge-specific clusters revisited: a comprehensive phylogeny of sponge-associated microorganisms. Environ Microbiol. 2012;14:517–24.

33. Siegl A, Kamke J, Hochmuth T, Piel J, Richter M, Liang C, et al. Single-cell genomics reveals the lifestyle of Poribacteria, a candidate phylum symbiotically associated with marine sponges. ISME J. 2011;5:61–70.

34. Webster NS, Luter HM, Soo RM, Botté ES, Simister RL, Abdo D, et al. Same, same but different: symbiotic bacterial associations in GBR sponges. Front Microbiol. 2013;3. doi:10.3389/fmicb.2012.00444.

35. Hentschel U, Piel J, Degnan SM, Taylor MW. Genomic insights into the marine sponge microbiome. Nat Rev Microbiol. 2012;10:641–54.

36. Biard T, Pillet L, Decelle J, Poirier C, Suzuki N, Not F. Towards an Integrative Morpho-molecular Classification of the Collodaria (Polycystinea, Radiolaria). Protist. 2015;166:374–88.

37. Guidi L, Chaffron S, Bittner L, Eveillard D, Larhlimi A, Roux S, et al. Plankton networks driving carbon export in the oligotrophic ocean. Nature. 2016;532:465–70.

38. Schwarz JA, Brokstein PB, Voolstra C, Terry AY, Miller DJ, Szmant AM, et al. Coral life history and symbiosis: Functional genomic resources for two reef building Caribbean corals, Acropora palmata and Montastraea faveolata. BMC Genomics. 2008;9:97.

39. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science. 2015;348:1261359.

40. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007;17:377–86.

41. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome (Trinity). Nat Biotechnol. 2011;29:644–52.

42. Thanh NM, Jung H, Lyons RE, Njaci I, Yoon B-H, Chand V, et al. Optimizing de novo transcriptome assembly and extending genomic resources for striped catfish (Pangasianodon hypophthalmus). Mar Genomics. 2015;23:87–97.

43. Ungaro A, Pech N, Martin J-F, McCairns SR, Mevy J-P, Chappaz R, et al. Challenges and advances for transcriptome assembly in non-model species. bioRxiv. 2017;:084145.

558 44. Anderson OR. Radiolaria. Springer Science & Business Media; 2012.

559 45. Murray SA, Suggett DJ, Doblin MA, Kohli GS, Seymour JR, Fabris M, et al. Unravelling the
560 functional genetics of dinoflagellates: a review of approaches and opportunities. Perspect Phycol.
561 2016;:37–52.

562 46. Le Bescot N, Mahé F, Audic S, Dimier C, Garet M-J, Poulain J, et al. Global patterns of pelagic
563 dinoflagellate diversity across protist size classes unveiled by metabarcoding. Environ Microbiol.
564 2016;18:609–26.

565 47. Biard T, Bigeard E, Audic S, Poulain J, Gutierrez-Rodriguez A, Pesant S, et al. Biogeography
566 and diversity of Collodaria (Radiolaria) in the global ocean. ISME J. 2017;11:1331–44.

567 48. Decelle J, Suzuki N, Mahé F, de Vargas C, Not F. Molecular Phylogeny and Morphological
568 Evolution of the Acantharia (Radiolaria). Protist. 2012;163:435–50.

569 49. Biard T, Stemmann L, Picheral M, Mayot N, Vandromme P, Hauss H, et al. In situ imaging
570 reveals the biomass of giant protists in the global ocean. Nature. 2016;advance online publication.
571 doi:10.1038/nature17652.

572 50. Maillet N, Collet G, Vannier T, Lavenier D, Peterlongo P. Commet: Comparing and combining
573 multiple metagenomic datasets. In: 2014 IEEE International Conference on Bioinformatics and
574 Biomedicine (BIBM). 2014. p. 94–8.

575 51. Fofanov Y, Pettitt B, Li T, Tchoumakov S. Process and apparatus for using the sets of pseudo
576 random subsequences present in genomes for identification of species. 2005.
577 http://www.google.ch/patents/US20050255459.

578 52. Pesant S, Not F, Picheral M, Kandels-Lewis S, Bescot NL, Gorsky G, et al. Open science
579 resources for the discovery and analysis of *Tara* Oceans data. Sci Data. 2015;2:sdata201523.

580 53. Bhattacharya D, Agrawal S, Aranda M, Baumgarten S, Belcaid M, Drake JL, et al. Comparative
581 genomics explains the evolutionary success of reef-forming corals. eLife. 2016;5.

582 54. Shinzato C, Shoguchi E, Kawashima T, Hamada M, Hisata K, Tanaka M, et al. Using the
583 Acropora digitifera genome to understand coral responses to environmental change. Nature.
584 2011;476:320–3.

585 55. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, et al. The Marine
586 Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional
587 Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. PLOS Biol.
588 2014;12:e1001889.

589 56. Fernandez-Valverde SL, Calcino AD, Degnan BM. Deep developmental transcriptome
590 sequencing uncovers numerous new genes and enhances gene annotation in the sponge
591 Amphimedon queenslandica. BMC Genomics. 2015;16:387.

592 57. Pérez-Porro AR, Navarro-Gómez D, Uriz MJ, Giribet G. A NGS approach to the encrusting
593 Mediterranean sponge Crella elegans (Porifera, Demospongiae, Poecilosclerida): transcriptome
594 sequencing, characterization and overview of the gene expression along three life cycle stages. Mol
595 Ecol Resour. 2013;13:494–509.

596 58. Guzman C, Conaco C. Comparative transcriptome analysis reveals insights into the streamlined
597 genomes of haplosclerid demosponges. Sci Rep. 2016;6. doi:10.1038/srep18774.

598 59. Limasset A, Rizk G, Chikhi R, Peterlongo P. Fast and scalable minimal perfect hashing for
599 massive key sets. ArXiv170203154 Cs. 2017. http://arxiv.org/abs/1702.03154.

600 60. Rizk G, Lavenier D, Chikhi R. DSK: k-mer counting with very low memory usage.
601 Bioinformatics. 2013;29:652–3.

602 61. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina Sequence Data.
603 Bioinformatics. 2014;:btu170.

604 62. Smith-Unna R, Boursnell C, Patro R, Hibberd J, Kelly S. TransRate: reference free quality
605 assessment of de novo transcriptome assemblies. Genome Res. 2016;:gr.196469.115.

606 63. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo
607 transcript sequence reconstruction from RNA-seq using the Trinity platform for reference
608 generation and analysis. Nat Protoc. 2013;8:1494–512.

609 64. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-
610 scale protein function classification. Bioinforma Oxf Engl. 2014;30:1236–40.

611 65. Botebol H, Lelandais G, Six C, Lesuisse E, Meng A, Bittner L, et al. Acclimation of a low iron
612 adapted Ostreococcus strain to iron limitation through cell biomass lowering. Sci Rep. 2017;7:327.

613 66. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat
614 Methods. 2015;12:59–60.

615 67. Yang Y, Smith SA. Optimizing de novo assembly of short-read RNA-seq data for
616 phylogenomics. BMC Genomics. 2013;14:328.

617 68. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J
618 Mol Biol. 1990;215:403–10.

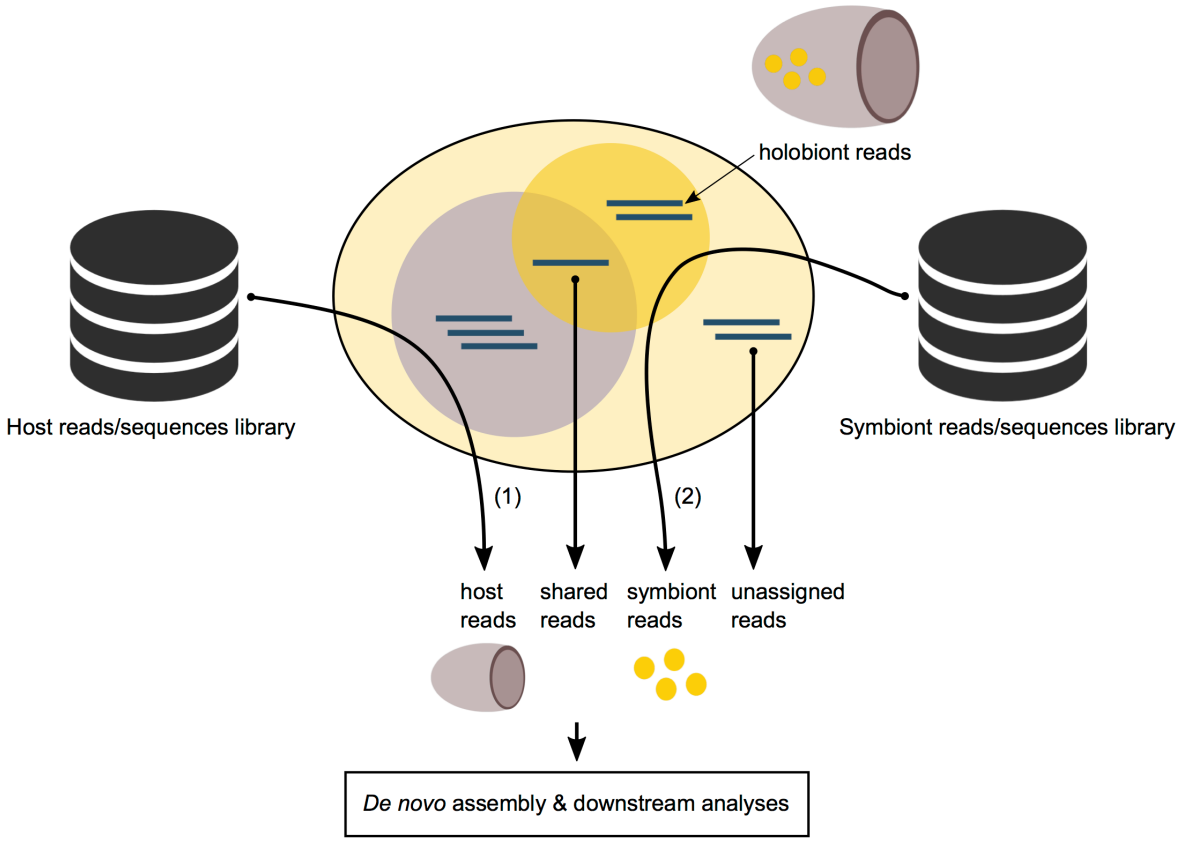


Fig. 1 Theoretical overview on the application of SRC_c on holobiont transcriptome. The comparisons to (1) host and (2) symbiont reads/sequences library are done against the entire holobiont dataset to retrieve host and symbiont similar reads. The 4 resulting subsets (host, symbiont, shared and unassigned reads) are then processed independently (*de novo* assembly and downstream analyses).

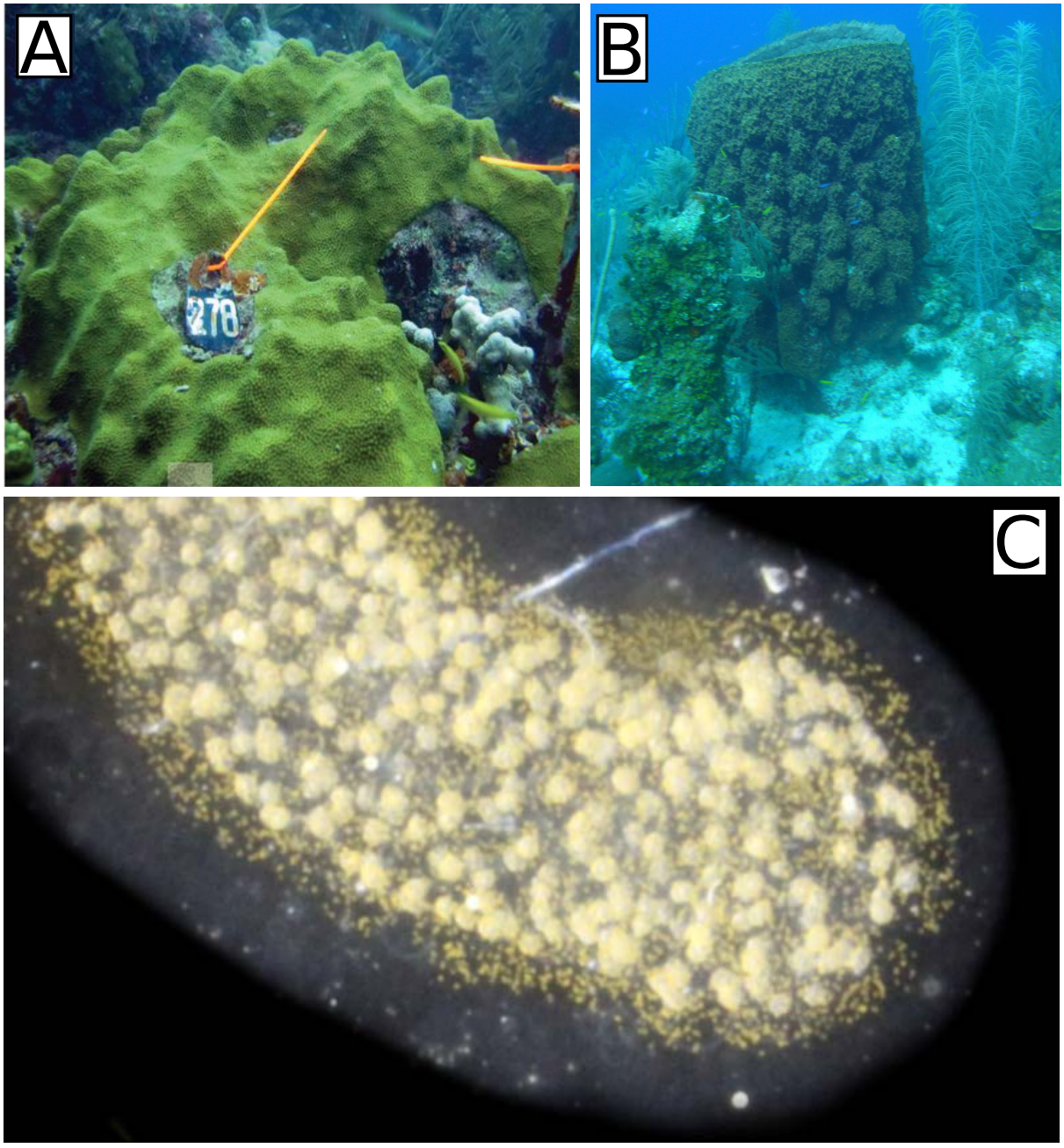
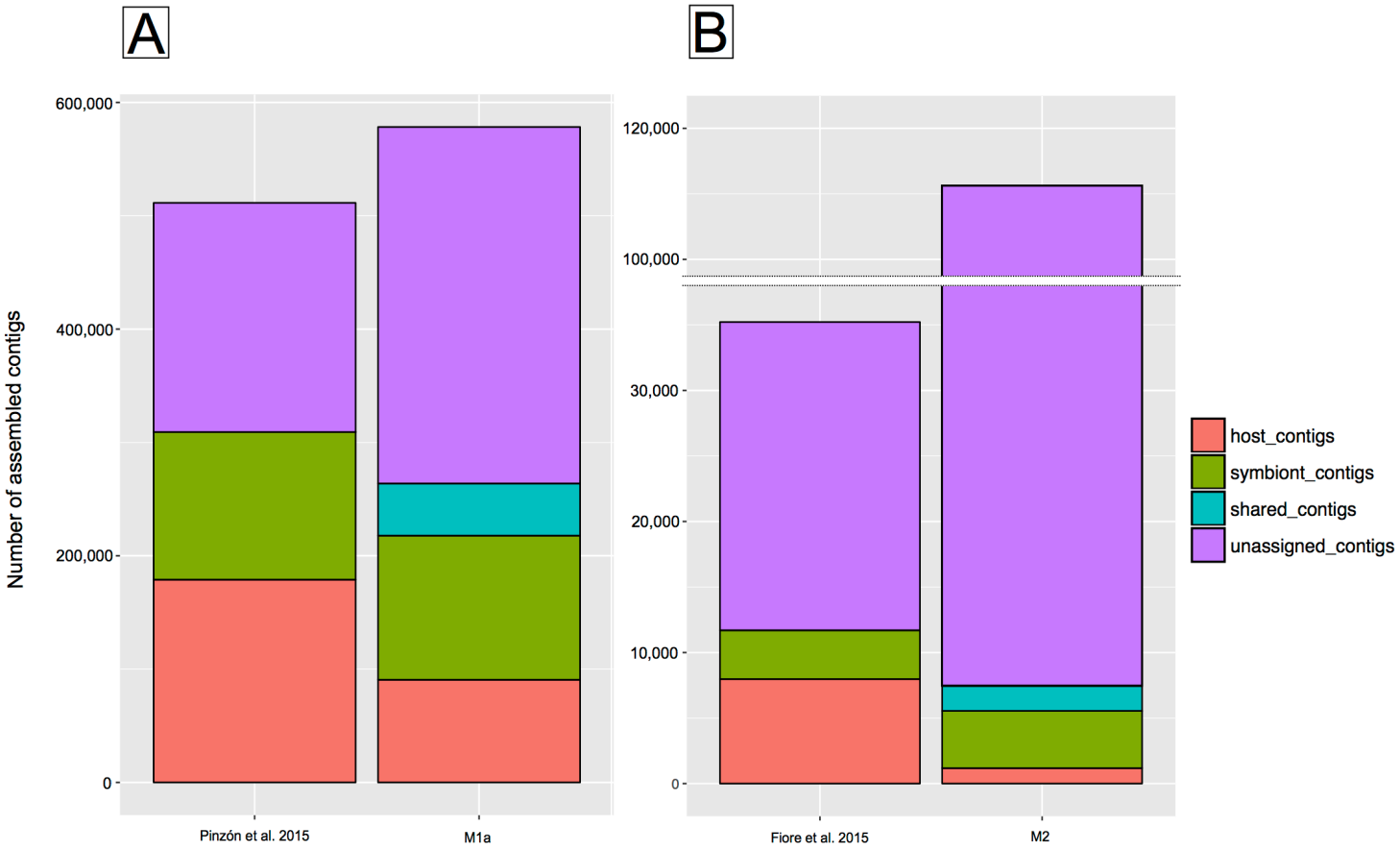


Fig. 2 Pictures of the 3 holobiont models. (A) the *Orbicella faveolata* holobiont in symbiosis (unbleached) in 2010 at reefs of La Parguera, Puerto Rico (credits: Pinzón et al. 2015). (B) A *Xestospongia muta* specimen in symbiosis on a coral reef near Little Cayman in the Caribbean (credits: Cara Fiore, january 14, 2015 http://feedthedatamonster.com). (C) A Collodaria colony with symbionts sampled in South Pacific Ocean at station 112.01 of the Tara Pacific expedition in 2011 (credits: Johan Decelle).



	Pinzón et al. 2015	M1a
# reads for assembly	775,025,024	775,025,024
# assembled contigs	442,294	578,333
N50 (bp)	1551	971
# cds with FA	108,409	171,931

	Fiore et al. 2015	M2
# reads for assembly	13,723,211	30,541,810
# assembled contigs	35,219	113,786
N50 (bp)	1010	719
# cds with FA	NC	24,738

Fig. 3 Overview and comparison of the total assembled contigs for holobiont model M1a and M2 compared to the assembled meta-transcriptomes from (A) Pinzon et al. 2015 and (B) Fiore et al. 2015 respectively. Details about *de novo* assembly and functional annotation (termed FA) features are presented in corresponding tables for (A) holobiont model M1a versus Pinzon et al. 2015 meta-transcriptome, and (B) holobiont model M2 versus Fiore et al. 2015. NC means that exact number is not communicated.

		Time(hh:mm:ss)	Memory (Gb)
Cnidaria-Dinophyta holobiont (M1)	all symbionts library (M1a)	15:40:42	34,2
	Symbiodinium spp. library (M1b)	01:34:57	6,96
	other symbionts library (M1c)	15:08:45	33,7
	host library	01:06:56	3,9
Porifera-Bacteria holobiont (M2)	symbionts library	21:04:47	58,9
	host library	02:46:06	9,60
Radiolaria-Dinophyta holobiont (M3)	symbionts library	07:05:28	4,10
	host library	00:05:57	3,9

Tab. 1 Performances (memory peak and wallclock time) of SRC indexing and query steps on the several data sets for models M1 and M2.

		# reads	% reads from holobiont
Orbicella faveolata holobiont (M1a)	**total**	**775 025 024**	
	assigned to host library	498 008 661	64.26%
	assigned to symbiont library	56 011 798	7.23%
	shared	32 133 818	4.15%
	unassigned	188 870 747	24.37%
Orbicella faveolata holobiont (M1b)	assigned to host library	500 145 229	64.53%
	assigned to symbiont library	54 850 148	7.08%
	shared	29 997 250	3.87%
	unassigned	190 032 397	24.52%
Orbicella faveolata holobiont (M1c)	assigned to host library	521 591 231	67.30%
	assigned to symbiont library	4 817 450	0.62%
	shared	8 551 248	1.10%
	unassigned	240 065 095	30.98%
Xestospongia muta holobiont (M2)	**total**	**33 220 038**	
	assigned to host library	6 193 678	19.04%
	assigned to symbiont library	825 154	10.64%
	shared	5 112 031	8.63%
	unassigned	21 090 174	61.69%
Collozoum sp. holobiont (M3)	**total**	**97 957 794**	
	assigned to host library	3 188 944	3.26%
	assigned to symbiont library	23 234 402	23.72%
	shared	531 432	0.54%
	unassigned	71 003 016	72.48%

Tab. 2 SRC_c assignment results for the Cnidaria-Dinophyta holobiont model (M1) against the complete Dinophyta library (M1a), the *Symbiodinium* spp. exclusive library (M1b) and the Dinophyta library excluding *Symbiodinium* spp. (M1c), the Porifera-Bacteria holobiont model (M2) and the Radiolaria-Dinophyta holobiont model (M3).

		# contigs	% contigs in holobiont	smallest	longest	N50	mean length	%GC	remapping rate (%)	# with ORFs	% of contigs with ORFs	remapping rate of holobiont reads (%)	# predicted cds	% contigs with predicted cds	# annotated cds	% cds with functional annotations
Cnidaria-Dinophyta holobiont (M1a)	host	90 558	15.66%	201	29 214	1 840	949	42%	97.8%	31 105	34.3%	71.6%	42 992	47.5%	35 358	39%
	symbiont	127 212	22%	201	13 093	1 091	719	57%	90.4%	58 286	45.8%	72.3%	84 151	66.2%	53 011	41.7%
	shared	46 017	7.96%	201	7 727	1 067	796	55%	82.3%	28 075	61%	41.4%	38 547	83.8%	25 382	55.2%
	unassigned	314 546	54.39%	201	19 174	732	558	46%	83.6%	67 509	21.5%	25.9%	89 533	28.5%	58 188	18.5%
	total	**578 333**								**184 975**			**255 223**		**171 939**	
Porifera-Bacteria holobiont (M2)	host	2 654	2.33%	201	1 921	299	311	42%	44.4%	215	8.1%	17.6%	707	26.6%	593	83.9%
	symbiont	2 431	2.14%	201	5 001	406	396	46%	25%	411	16.9%	4.7%	1 072	44.1%	988	92.2%
	shared	2 324	2.04%	201	751	301	299	54%	86.4%	8	0.3%	22.3%	163	7%	30	18.4%
	unassigned	106 377	93.49%	201	8 811	748	572	39%	73.2%	29 520	27.8%	59.1%	43 150	40.6%	23 127	53.6%
	total	**113 786**								**30 154**			**45 092**		**24 738**	
Radiolaria-Dinophyta holobiont (M3)	host	693	0.41%	201	1 209	277	303	42%	65.2%	44	6.3%	10.6%	123	17.7%	49	7.1%
	symbiont	5 207	3.08%	201	1 777	324	328	54%	76.2%	618	11.9%	32%	1 468	28.2%	942	18.1%
	shared	52	0.03%	201	639	298	308	39%	81.3%	0	0%	18.6%	6	11.5%	5	9.6%
	unassigned	162 947	96.48%	201	10 569	714	580	41%	89.7%	49 032	30.1%	73.2%	72 420	44.4%	44 772	27.5%
	total	**168 899**								**49 694**			**74 017**		**45 768**	

Tab. 3 *De novo assembly* metrics and downstream analysis of SRC_c resulting subsets for holobiont models M1a, M2 and M3.

		no SRC	SRC
# reads used in assembly		48 733 956	48 660 697
# assembled contigs		167 023	168 899
N50 (bp)	**total**	818	702
	host		277
	symbiont		324
	shared		298
	unassigned		714
remapping rates (%)	**total**	85,6	90,5
	host		65,2
	symbiont		76,2
	shared		81,3
	unassigned		89,7
# chimera	**total**	777	418
	host		4
	symbiont		47
	shared		0
	unassigned		367
Calculation time (min)	**total**	330	2 783
	SRC		2 460
	assembly	330	323

Tab. 4 SRC_c impact on assembled contigs quality and calculation times of Radiolaria-Dinophyta holobiont model (M3) compared to a direct meta-transcriptome assembly strategy. In grey are displayed the details for SRC_c holobiont categories (host, symbiont, shared and unassigned). The “total” values for N50 and remapping rates of the SRC_c strategy were re-calculated on pooled contigs from host, symbiont, shared and unassigned subsets.

Additional file 1 SRC_c library content informations and data sources. Table with detailed informations of SRC_c libraries contents. The type of data and the total library sizes are displayed. It includes taxonomic contents and links to data repositories for holobiont models M1, M2 and M3 and data that constitute SRC_c reads/sequences libraries.

model		species (strain)	source	note	link to data	status	Number of sequences	sequence length
M1	Holobiont	*Orbicella faveolata* +Dinophyt	Whole transcriptome analysis reveals changes in expression of immune-related genes during and after bleaching in a reef-building coral. Jorge Pinzòn et al. 2015	Cnidaria-Dinophya holobiont (RNA-seq raw reads)	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA236103	holobiont (M1)	775,025,024 reads	25-85 bp
	Cnidaria	*Acropora_digitifera* *Montastraea_faveolata* *Acropora_hyacinthu* *Platygyra_carnosus* *Acropora_millepora* *Pocillopora_damicornis* *Acropora_palmata* *Porites_astreoides* *Acropora_tenuis* *Porites_australiensis* *Astreopora_sp* *Porites_lobata* *Favia_sp* *Pseudodiploria_strigosa* *Fungia_scutaria* *Seriatopora_hystrix* *Madracis_auretenra* *Stylophora_pistillata* *Montastraea_cavernosa*	Comparative genomics explains the evolutionary success of reef-forming corals. Debashish Bhattacharya et al. 2016	coral assembled transcriptomes "Coral genomic and transcriptome data compiled in this study are summarized in Figure 1—source data 1. All data were filtered to remove assembled contigs <300 bp" http://comparative.reefgenomics.org/ (transcriptome assembly)	https://elifesciences.org/articles/13288	host library	668,259 assembled contigs	
	Dinophyta		The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. Keeling et al. 2014	Dinophyta (RNA-seq raw reads)	https://imicrobe.us/project/view/104	symbiont library	5,563,498,607 reads (all Dinophyta) 123,122,726 reads (only *Symbiodinium* sp.) 5,440,375,881 reads (Dinophyta less Symbiodinium sp.)	
M2	Holobiont	*Xestonspongia muta* +Bacteri	Transcriptional activity of the giant barrel sponge, Xestospongia muta Holobiont: molecular evidence for metabolic interchange. Cara Fiore et al. 2015	Porifera-Bacteria holobiont (RNA-seq raw reads)	https://imicrobe.us/project/view/128	holobiont (M2)	33,220,038 reads	25-99 bp
	Porifera	*Amphimedon queenslandica*	Deep developmental transcriptome sequencing uncovers numerous new genes and enhances gene annotation in the sponge Amphimedon queenslandica. Selene L. Fernandez-Valverde et al. 2015	Porifera(RNA-seq raw reads)	https://trace.ncbi.nlm.nih.gov/Traces/study/?acc=SRP044247	host library	642,229,924 reads	
	Porifera	*Haliclona amboinensis*	Comparative transcriptome analysis reveals insights into the streamlined genomes of haplosclerid demosponges. Christine Guzman et al. 2016	Porifera(RNA-seq raw reads)	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA264137			
	Porifera	*Haliclona tubifera*	Comparative transcriptome analysis reveals insights into the streamlined genomes of haplosclerid demosponges. Christine Guzman et al. 2016	Porifera(RNA-seq raw reads)	https://www.ncbi.nlm.nih.gov/bioproject/?term=Haliclona%20tub			
	Porifera	*Crella elegans*	A NGS approach to the encrusting Mediterranean sponge Crella elegans (Porifera, Demospongiae, Poecilosclerida): transcriptome sequencing, characterization and overview of the gene expression along three life cycle stages. A. R. Perez-Porro et al. 2013	Porifera(RNA-seq raw reads)	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA182019			
	Bacteria	Bacteria		bacterial gene catalog (assembled genes	http://ocean-microbiome.embl.de/companion.html#OM-RGC	symbiont library	40,154,822 assembled genes	
M3	Holobiont	*Collozoum* sp. +Dinophyta	unpublished (Roscoff, FR)	Radiolaria-Dinophya holobiont (RNA-seq raw reads)		holobiont (M3)	97,957,794 reads	32-101 bp
	Rhizaria	*A. elongata* *Collozoum sp.* *S. streptacantha* *A. scolymantha*	Transcriptome analyses to investigate symbiotic relationships between marine protists. Balzano et al. 2015	Rhizaria assembled transcriptomes	None. Roscoff, FR	host library	7,215 assembled contigs	
	Dinophyta		The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. Keeling et al. 2014	Dinophyta (RNA-seq raw reads)	https://imicrobe.us/project/view/104	symbiont library	5,563,498,607 reads (all Dinophyta)	

M1

Phylum	Order	species	# contigs from meta-transcriptome	% contigs from meta-transcriptome
Cnidaria				
	Scleractinia			
		Orbicella faveolata	**71 143**	**12%**
		Acropora digitifera	9 537	1.6%
Dinoflagellata				
	Suessiales			
		Symbiodinium microadriaticum	148 409	25.6%
		Symbiodinium sp. Clade C	210	0.03%
		unclassified *Symbdiodinium*	33	0.005%
Bacteria and Archeae			1 417	0.24%
Viruses			31	0.005%
Other Eukaryotes			80 481	13.9%
Unassigned			276 609	47.8%

M2

Phylum	Order	species	# contigs from meta-transcriptome	% contigs from meta-transcriptome
Porifera				
	Haplosclerida			
		Xestospongia muta	**11**	**0.009%**
		Amphimedon queenslandica	33 810	29.7%
Bacteria and Archeae			**21 318**	**18.7%**
Viruses			58	0.05%
Other Eukaryotes			8 923	7.8%
Unassigned			49 666	43.6%

Additional file 2 Taxonomic assignment of SRC assembled contigs with MEGAN6 for the holobiont models M1 and M2.

Host (BP)

term_ID	description	value
GO:0008152	metabolic process	819
GO:0035556	intracellular signal transduction	779
GO:0007165	signal transduction	527
GO:0006355	regulation of transcription, DNA-templated	474
GO:0006811	ion transport	467
GO:0006508	proteolysis	423
GO:0006396	RNA processing	266
GO:0006751	glutathione catabolic process	257
GO:0007154	cell communication	256
GO:0009190	cyclic nucleotide biosynthetic process	250
GO:0006886	intracellular protein transport	230
GO:0005975	carbohydrate metabolic process	213
GO:0045454	cell redox homeostasis	206
GO:0042981	regulation of apoptotic process	203
GO:0000413	protein peptidyl-prolyl isomerization	186

Host (MF)

term_ID	description	value
GO:0005515	protein binding	36349
GO:0004930	G-protein coupled receptor activity	9794
GO:0005509	calcium ion binding	8372
GO:0003676	nucleic acid binding	8351
GO:0004672	protein kinase activity	3152
GO:0003677	DNA binding	2833
GO:0005524	ATP binding	1920
GO:0003824	catalytic activity	1780
GO:0003700	transcription factor activity, sequence-specific DNA binding	1531
GO:0005506	iron ion binding	1389
GO:0016491	oxidoreductase activity	1378
GO:0004252	serine-type endopeptidase activity	1281
GO:0005249	voltage-gated potassium channel activity	1138
GO:0005044	scavenger receptor activity	1092
GO:0005216	ion channel activity	989

Host (CC)

term_ID	description	value
GO:0016021	integral component of membrane	2464
GO:0016020	membrane	2308
GO:0005622	intracellular	622
GO:0005576	extracellular region	523
GO:0005634	nucleus	479
GO:0005886	plasma membrane	201
GO:0005874	microtubule	198
GO:0005737	cytoplasm	162
GO:0000786	nucleosome	155
GO:0000139	Golgi membrane	97
GO:0005856	cytoskeleton	84
GO:0005813	centrosome	57
GO:0005789	endoplasmic reticulum membrane	51
GO:0005887	integral component of plasma membrane	44
GO:0005783	endoplasmic reticulum	39

Symbiont (BP)

term_ID	description	value
GO:0008152	metabolic process	1065
GO:0006508	proteolysis	604
GO:0000413	protein peptidyl-prolyl isomerization	572
GO:0009765	photosynthesis, light harvesting	415
GO:0005975	carbohydrate metabolic process	378
GO:0045454	cell redox homeostasis	326
GO:0001522	pseudouridine synthesis	323
GO:0009190	cyclic nucleotide biosynthetic process	267
GO:0015986	ATP synthesis coupled proton transport	219
GO:0006629	lipid metabolic process	212
GO:0006810	transport	200
GO:0055114	oxidation-reduction process	181
GO:0006886	intracellular protein transport	179
GO:0006807	nitrogen compound metabolic process	172
GO:0006470	protein dephosphorylation	170

Symbiont (MF)

term_ID	description	value
GO:0005515	protein binding	29906
GO:0005509	calcium ion binding	10452
GO:0003676	nucleic acid binding	4617
GO:0016491	oxidoreductase activity	3859
GO:0003824	catalytic activity	3805
GO:0004672	protein kinase activity	3781
GO:0005524	ATP binding	3244
GO:0003723	RNA binding	2520
GO:0003735	structural constituent of ribosome	2280
GO:0046872	metal ion binding	2032
GO:0003677	DNA binding	1758
GO:0003777	microtubule motor activity	1688
GO:0005216	ion channel activity	1416
GO:0016787	hydrolase activity	1164
GO:0003924	GTPase activity	1151

Symbiont (CC)

term_ID	description	value
GO:0016021	integral component of membrane	1154
GO:0005874	microtubule	1045
GO:0016020	membrane	704
GO:0005737	cytoplasm	301
GO:0000015	phosphopyruvate hydratase complex	217
GO:0005634	nucleus	216
GO:0009523	photosystem II	208
GO:0000786	nucleosome	176
GO:0005622	intracellular	146
GO:0000139	Golgi membrane	142
GO:0009522	photosystem I	140
GO:0005956	protein kinase CK2 complex	48
GO:0005643	nuclear pore	45
GO:0000148	1,3-beta-D-glucan synthase complex	41
GO:0005576	extracellular region	37

Shared (BP)

term_ID	description	value
GO:0008152	metabolic process	465
GO:0006508	proteolysis	327
GO:0000413	protein peptidyl-prolyl isomerization	302
GO:0045454	cell redox homeostasis	179
GO:0009190	cyclic nucleotide biosynthetic process	169
GO:0005975	carbohydrate metabolic process	163
GO:0006886	intracellular protein transport	151
GO:0009765	photosynthesis, light harvesting	128
GO:0018298	protein-chromophore linkage	128
GO:0015986	ATP synthesis coupled proton transport	115
GO:0006810	transport	107
GO:0055114	oxidation-reduction process	104
GO:0055085	transmembrane transport	102
GO:0006621	protein retention in ER lumen	99
GO:0006464	cellular protein modification process	94

Shared (MF)

term_ID	description	value
GO:0005515	protein binding	12607
GO:0005509	calcium ion binding	5043
GO:0003676	nucleic acid binding	2430
GO:0004672	protein kinase activity	2205
GO:0003824	catalytic activity	1913
GO:0005524	ATP binding	1574
GO:0016491	oxidoreductase activity	1537
GO:0003723	RNA binding	1359
GO:0003677	DNA binding	1029
GO:0003777	microtubule motor activity	969
GO:0046872	metal ion binding	782
GO:0003735	structural constituent of ribosome	781
GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds	617
GO:0016787	hydrolase activity	533
GO:0005216	ion channel activity	532

Unassigned (BP)

term_ID	description	value
GO:0008152	metabolic process	993
GO:0006508	proteolysis	674
GO:0045454	cell redox homeostasis	473
GO:0000413	protein peptidyl-prolyl isomerization	431
GO:0005975	carbohydrate metabolic process	373
GO:0015986	ATP synthesis coupled proton transport	275
GO:0009765	photosynthesis, light harvesting	241
GO:0006629	lipid metabolic process	241
GO:0015074	DNA integration	235
GO:0006886	intracellular protein transport	229
GO:0009190	cyclic nucleotide biosynthetic process	223
GO:0055085	transmembrane transport	210
GO:0001522	pseudouridine synthesis	204
GO:0055114	oxidation-reduction process	180
GO:0006810	transport	172

Unassigned (MF)

term_ID	description	value
GO:0005515	protein binding	33923
GO:0005509	calcium ion binding	9506
GO:0003676	nucleic acid binding	5885
GO:0004672	protein kinase activity	4781
GO:0003824	catalytic activity	3983
GO:0016491	oxidoreductase activity	3336
GO:0005524	ATP binding	3012
GO:0003723	RNA binding	2564
GO:0003735	structural constituent of ribosome	2532
GO:0003677	DNA binding	1995
GO:0046872	metal ion binding	1921
GO:0003777	microtubule motor activity	1666
GO:0005506	iron ion binding	1387
GO:0004252	serine-type endopeptidase activity	1264
GO:0005216	ion channel activity	1062

Unassigned (CC)

term_ID	description	value
GO:0016021	integral component of membrane	1385
GO:0016020	membrane	1036
GO:0000786	nucleosome	479
GO:0005737	cytoplasm	440
GO:0005874	microtubule	397
GO:0005622	intracellular	273
GO:0005634	nucleus	267
GO:0005576	extracellular region	148
GO:0000139	Golgi membrane	112
GO:0000015	phosphopyruvate hydratase complex	105
GO:0009522	photosystem I	99
GO:0009523	photosystem II	68
GO:0005643	nuclear pore	58
GO:0005783	endoplasmic reticulum	53
GO:0005887	integral component of plasma membrane	40

Additional file 3-1 details of common GO annotations M1, our contigs versus previous studies.

Host (BP)

term_ID	description	value
GO:0006886	intracellular protein transport	16
GO:0006508	proteolysis	15
GO:0016311	dephosphorylation	10
GO:0000398	mRNA splicing, via spliceosome	8
GO:0000413	protein peptidyl-prolyl isomerization	7
GO:0055114	oxidation-reduction process	6
GO:0035556	intracellular signal transduction	5
GO:0007155	cell adhesion	4
GO:0008152	metabolic process	4
GO:0046034	ATP metabolic process	4
GO:0009772	photosynthetic electron transport in photosystem II	3
GO:0006596	polyamine biosynthetic process	3
GO:0045454	cell redox homeostasis	3
GO:0006351	transcription, DNA-templated	3
GO:0016579	protein deubiquitination	3

Host (MF)

term_ID	description	value
GO:0005044	scavenger receptor activity	595
GO:0005515	protein binding	414
GO:0004672	protein kinase activity	191
GO:0005509	calcium ion binding	135
GO:0005200	structural constituent of cytoskeleton	79
GO:0004725	protein tyrosine phosphatase activity	76
GO:0003735	structural constituent of ribosome	54
GO:0003924	GTPase activity	51
GO:0005525	GTP binding	50
GO:0005524	ATP binding	37
GO:0016787	hydrolase activity	26
GO:0003774	motor activity	22
GO:0004096	catalase activity	21
GO:0004842	ubiquitin-protein transferase activity	20
GO:0004713	protein tyrosine kinase activity	20

Host (CC)

term_ID	description	value
GO:0016020	membrane	329
GO:0000786	nucleosome	64
GO:0005874	microtubule	45
GO:0005956	protein kinase CK2 complex	9
GO:0005622	intracellular	7
GO:0005885	Arp2/3 protein complex	4
GO:0008290	F-actin capping protein complex	4
GO:0016021	integral component of membrane	3
GO:0009522	photosystem I	3
GO:0005887	integral component of plasma membrane	1
GO:0005856	cytoskeleton	1

Symbiont (BP)

term_ID	description	value
GO:0015986	ATP synthesis coupled proton transport	111
GO:0009772	photosynthetic electron transport in photosystem II	42
GO:0015031	protein transport	35
GO:0006826	iron ion transport	31
GO:0008152	metabolic process	28
GO:0015992	proton transport	22
GO:0008033	tRNA processing	20
GO:0006457	protein folding	19
GO:0006412	translation	19
GO:0000160	phosphorelay signal transduction system	14
GO:0055114	oxidation-reduction process	13
GO:0006520	cellular amino acid metabolic process	13
GO:0009306	protein secretion	12
GO:0005975	carbohydrate metabolic process	11
GO:0006810	transport	9

Symbiont (MF)

term_ID	description	value
GO:0003735	structural constituent of ribosome	742
GO:0003677	DNA binding	157
GO:0005524	ATP binding	154
GO:0003924	GTPase activity	147
GO:0003723	RNA binding	138
GO:0003676	nucleic acid binding	111
GO:0003824	catalytic activity	109
GO:0005525	GTP binding	85
GO:0000166	nucleotide binding	83
GO:0016491	oxidoreductase activity	59
GO:0004872	receptor activity	55
GO:0009055	electron carrier activity	53
GO:0005515	protein binding	48
GO:0003746	translation elongation factor activity	48
GO:0003899	DNA-directed 5'-3' RNA polymerase activity	44

Symbiont (CC)

term_ID	description	value
GO:0005737	cytoplasm	49
GO:0016021	integral component of membrane	25
GO:0005622	intracellular	23
GO:0016020	membrane	22
GO:0005840	ribosome	13
GO:0009279	cell outer membrane	11
GO:0000015	phosphopyruvate hydratase complex	10
GO:0019867	outer membrane	3
GO:0009276	Gram-negative-bacterium-type cell wall	2
GO:0005960	glycine cleavage complex	2
GO:0005856	cytoskeleton	1

Shared (BP)

term_ID	description	value
GO:0009772	photosynthetic electron transport in photosystem II	7

Shared (MF)

term_ID	description	value
GO:0005515	protein binding	24
GO:0003924	GTPase activity	15
GO:0003677	DNA binding	5
GO:0005524	ATP binding	4
GO:0005525	GTP binding	1

Unassigned (BP)

term_ID	description	value
GO:0007165	signal transduction	558
GO:0008152	metabolic process	523
GO:0035556	intracellular signal transduction	431
GO:0007154	cell communication	372
GO:0006508	proteolysis	310
GO:0006629	lipid metabolic process	249
GO:0006886	intracellular protein transport	198
GO:0007155	cell adhesion	169
GO:0006810	transport	155
GO:0006355	regulation of transcription, DNA-templated	155
GO:0005975	carbohydrate metabolic process	134
GO:0042981	regulation of apoptotic process	120
GO:0030001	metal ion transport	119
GO:0016311	dephosphorylation	109
GO:0000398	mRNA splicing, via spliceosome	103

Unassigned (MF)

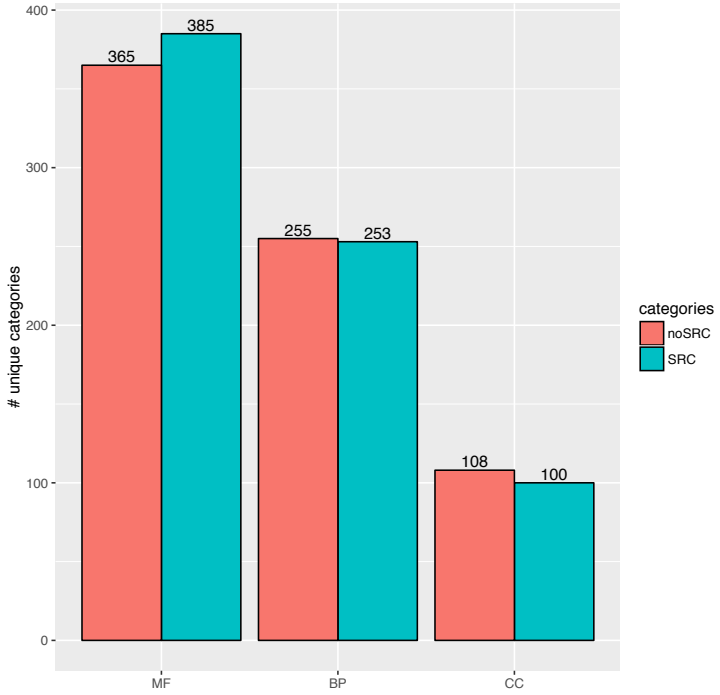
term_ID	description	value
GO:0005515	protein binding	24644
GO:0005044	scavenger receptor activity	3416
GO:0005509	calcium ion binding	2904
GO:0004672	protein kinase activity	2360
GO:0003676	nucleic acid binding	2223
GO:0003824	catalytic activity	1151
GO:0003677	DNA binding	1147
GO:0005524	ATP binding	1035
GO:0016491	oxidoreductase activity	1009
GO:0004725	protein tyrosine phosphatase activity	966
GO:0004930	G-protein coupled receptor activity	854
GO:0008270	zinc ion binding	752
GO:0003700	transcription factor activity, sequence-specific DNA binding	657
GO:0004867	serine-type endopeptidase inhibitor activity	606
GO:0003723	RNA binding	577

Unassigned (CC)

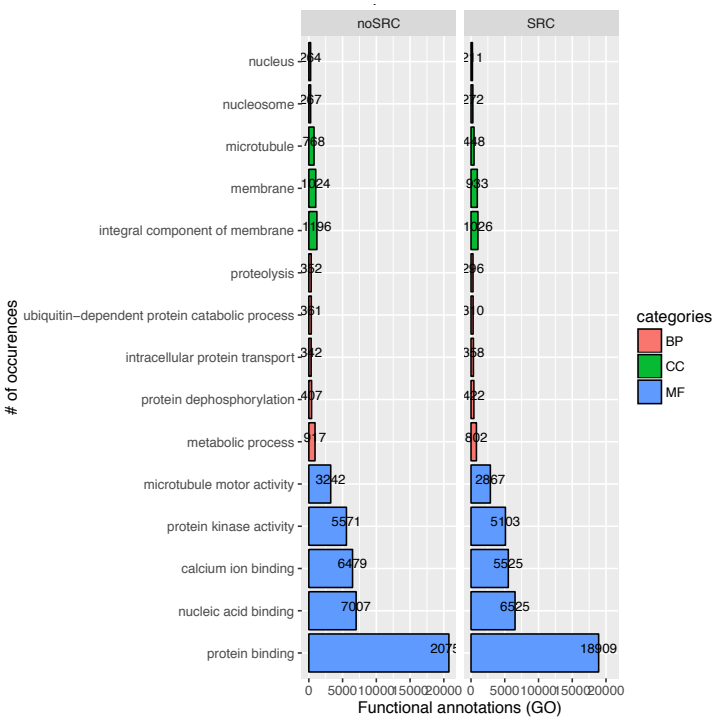
term_ID	description	value
GO:0016020	membrane	3409
GO:0016021	integral component of membrane	792
GO:0005622	intracellular	605
GO:0005634	nucleus	255
GO:0005737	cytoplasm	126
GO:0005886	plasma membrane	122
GO:0005856	cytoskeleton	105
GO:0005576	extracellular region	69
GO:0005783	endoplasmic reticulum	47
GO:0005789	endoplasmic reticulum membrane	40
GO:0000145	exocyst	37
GO:0000159	protein phosphatase type 2A complex	28
GO:0005643	nuclear pore	28
GO:0000786	nucleosome	27
GO:0005887	integral component of plasma membrane	25

Additional file 3-2 details of common GO annotations M2, our contigs versus previous studies.

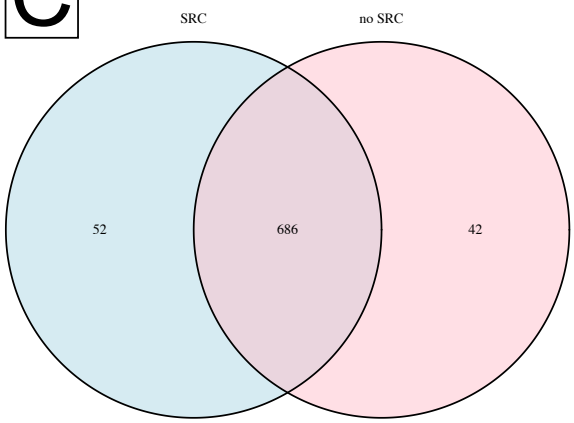
A


400
300
200
100
0
unique categories
365
385
255
253
108
100
MF
BP
CC
categories
noSRC
SRC

B


noSRC
SRC
nucleus
nucleosome
microtubule
membrane
integral component of membrane
proteolysis
ubiquitin−dependent protein catabolic process
intracellular protein transport
protein dephosphorylation
metabolic process
microtubule motor activity
protein kinase activity
calcium ion binding
nucleic acid binding
protein binding
3242
5571
6479
7007
2867
5103
5525
6525
18909
0
5000
10000
15000
20000
Functional annotations (GO)
# of occurences
categories
BP
CC
MF

C


SRC
no SRC
52
686
42

Host (BP)

term_ID	description	value
GO:6813	potassium ion transport	2

Host (MF)

term_ID	description	value
GO:5515	protein binding	11
GO:4672	protein kinase activity	7
GO:3676	nucleic acid binding	2
GO:5524	ATP binding	2
GO:46872	metal ion binding	1
GO:3677	DNA binding	1

Symbiont (BP)

term_ID	description	value
GO:9772	photosynthetic electron transport in photosystem II	24
GO:6886	intracellular protein transport	19
GO:919	cyclic nucleotide biosynthetic process	7
GO:6396	RNA processing	6
GO:652	cellular amino acid metabolic process	6
GO:681	transport	6
GO:6511	ubiquitin-dependent protein catabolic process	5
GO:55114	oxidation-reduction process	4
GO:398	mRNA splicing via spliceosome	4
GO:945	pathogenesis	3
GO:5585	transmembrane transport	3
GO:1531	protein transport	2
GO:6457	protein folding	2
GO:694	vesicle docking involved in exocytosis	2

Symbiont (MF)

term_ID	description	value
GO:0005515	protein binding	473
GO:0005509	calcium ion binding	188
GO:0005200	structural constituent of cytoskeleton	153
GO:0003677	DNA binding	111
GO:0003924	GTPase activity	88
GO:0005524	ATP binding	82
GO:0004672	protein kinase activity	81
GO:0003676	nucleic acid binding	64
GO:0003777	microtubule motor activity	60
GO:0015002	heme-copper terminal oxidase activity	38
GO:0005525	GTP binding	25
GO:0009055	electron carrier activity	23
GO:0016491	oxidoreductase activity	23
GO:0003743	translation initiation factor activity	22

Symbiont (CC)

term_ID	description	value
GO:0005874	microtubule	103
GO:0009522	photosystem I	29
GO:0016021	integral component of membrane	13
GO:0016020	membrane	11
GO:0009521	photosystem	9
GO:0005634	nucleus	9
GO:0009523	photosystem II	2
GO:0031514	motile cilium	2
GO:0036157	outer dynein arm	1
GO:0000015	phosphopyruvate hydratase complex	1
GO:0005737	cytoplasm	1
GO:0005838	proteasome regulatory particle	1

Shared (BP)

term_ID	description	value
GO:0034220	ion transmembrane transport	3

Shared (MF)

term_ID	description	value
GO:0046872	metal ion binding	6

Shared (CC)

term_ID	description	value
GO:0016020	membrane	5

Unassigned (BP)

term_ID	description	value
GO:0008152	metabolic process	802
GO:0006470	protein dephosphorylation	422
GO:0006886	intracellular protein transport	339
GO:0006511	ubiquitin-dependent protein catabolic process	305
GO:0006508	proteolysis	296
GO:0009190	cyclic nucleotide biosynthetic process	249
GO:0007165	signal transduction	237
GO:0016311	dephosphorylation	183
GO:0006629	lipid metabolic process	182
GO:0000413	protein peptidyl-prolyl isomerization	179
GO:0005975	carbohydrate metabolic process	165
GO:0055085	transmembrane transport	159
GO:0016579	protein deubiquitination	159
GO:0045454	cell redox homeostasis	155

Unassigned (MF)

term_ID	description	value
GO:0005515	protein binding	18425
GO:0003676	nucleic acid binding	6459
GO:0005509	calcium ion binding	5337
GO:0004672	protein kinase activity	5015
GO:0003777	microtubule motor activity	2807
GO:0003824	catalytic activity	2231
GO:0005524	ATP binding	2177
GO:0003677	DNA binding	2001
GO:0046872	metal ion binding	2001
GO:0003723	RNA binding	1920
GO:0016491	oxidoreductase activity	1359
GO:0003924	GTPase activity	1045
GO:0005089	Rho guanyl-nucleotide exchange factor activity	1001
GO:0005525	GTP binding	933

Unassigned (CC)

term_ID	description	value
GO:0016021	integral component of membrane	1013
GO:0016020	membrane	922
GO:0005874	microtubule	345
GO:0000786	nucleosome	272
GO:0005634	nucleus	202
GO:0005956	protein kinase CK2 complex	193
GO:0005737	cytoplasm	180
GO:0000159	protein phosphatase type 2A complex	148
GO:0005885	Arp2/3 protein complex	55
GO:0005783	endoplasmic reticulum	42
GO:0005672	transcription factor TFIIA complex	35
GO:0008290	F-actin capping protein complex	34
GO:0071203	WASH complex	33
GO:0005643	nuclear pore	33

Additional file 4 : Comparison of functional annotations between SRC assembled transcriptomes and a *de novo* assembled transcriptome without the use of SRC_c in the case of holobiont model M3.

Details of the functional annotations results for the SRC strategy applied to M3. The tables displayed correspond to the top 15 GO annotations found in host, symbiont, shared and unassigned transcriptomes for the three levels of annotations (MF: Molecular Functions, BP: Biological Process and CC: Cellular Component).

kingdom	Group/Class	order	# contigs from meta-transcriptome	% contigs from meta-transcriptome
Rhizaria			3 910	2,34%
	Polycystinea		103	0,06%
		Collodaria	**10**	**0,01%**
		Nasselaria	55	
Alveolata			2 267	1,36%
	Dinophyceae		**1 383**	**0,83%**
		Gonyaulacales	6	
		Gymniodinales	13	
		Peridiniales	14	
		Prorocentrales	3	
		Suessiales	1 157	
		Syndiniales	7	
Bacteria and Archeae			3 799	2,27%
Viruses			76	0,05%
Other Eukaryotes			29 524	17,68%
Unassigned			127 447	76,31%

Additional file 5 Radiolaria-Dinophyta meta-transcriptome taxonomic assignment with MEGAN6. Table of taxonomic assignation of the 167,023 *de novo* assembled contigs from the assembly without SRC reads sorting of the holobiont model M3.

4.2 Identification des fonctions clefs dans les processus fonctionnels régissant la symbiose radiolaires et dinoflagellés

Une fois validée sur des modèles de symbiose publiées ainsi que sur un modèle d'association symbiotique radiolaire-dinoflagellés, j'ai employé l'approche `SRC_c` pour étudier trois transcriptomes d'holobionte radiolaire-dinoflagellés dont deux sont encore non publiés (Figures 4.4 et 4.5) (cf. détails d'échantillonnage ci-après). J'ai également enrichi la banque de séquences de référence pour l'hôte avec des lectures provenant de séquençage Illumina de (1) 4 transcriptomes de radiolaires et de 1 transcriptome de cercozoaire non publiés (Figure 4.2) dont l'échantillonnage et le séquençage ont été réalisés conjointement entre la station biologique de Roscoff et le Genoscope à Evry entre 2011 et 2013, et (2) 2 transcriptomes de Radiolaires non-symbiotiques disponibles publiquement depuis août [Krabberød, Orr et al. 2017]. L'objectif de cette étude est de fournir les « meilleurs » jeux de transcriptomes assemblés de Rhizaria symbiotiques et non-symbiotiques, et ce afin d'en déduire et d'en extraire les séquences (et potentiellement les fonctions associées) en lien avec les symbioses.

Code	Séquençage				Taxonomie				Trait fonctionnel	
	souche	tech.	banque	# cellules	classe	ordre	genre	espèce	photo-symbiotique	symbionte(s) présumé(s)
BDQ	PAC37	Illumina	OVATION	colony	Polycystinea	Collodaria	*Collozoum*	sp.	oui	dinoflagellates (B. nutricula, G. radiolaroae)
BFT	SES43	Illumina	OVATION	68 cells	Polycystinea	Spumellaria	*Euchitonia*	*furcata*	oui	dinoflagellates (B. nutricula, G. radiolaroae)
BRX	LIPID2	Illumina		colony	Polycystinea	Collodaria	*Raphidozoum*	sp.	oui	dinoflagellates (B. nutricula, G. radiolaroae)
BRW	SES20	Illumina		Few cells	Acantharea	Acantharia	*Actinelius*	*primordialis*	non	
BGU	OSH121	Illumina	SMARTER	160 cells	Filosa	Phaeodaria	*Protocystis*	*ornithocephala*	non	
BRY	OSH202	Illumina		33 cells	Polycystinea	Spumellaria	*Amphisphaera*	*tanzhiyuani*	non	
BGX	VIL377	Illumina	SMARTER	49 cells	Polycystinea	Nassellaria	*Eucyrtidium*	*acuminatum*	non	
BRR	OSH117	Illumina		37 cells	Polycystinea	Nassellaria	*Cycladophora*	*davisiana*	non	
ASRAAA		454	SMARTER	50 cells	Filosa	Phaeodaria	*Aulacantha*	*scolymantha*	non	
ASRAAB		454	SMARTER	colony	Polycystinea	Collodaria	*Collozoum*	sp.	oui	dinoflagellates (B. nutricula, G. radiolaroae)
ASRAAC		454	SMARTER	50 cells	Polycystinea	Spumellaria	*Spongosphaera*	*streptacantha*	oui	?
ASRAAD		454	SMARTER	150 cells	Acantharea	Acantharia	*Amphilonche*	*elongata*	oui	Haptophyta (Phaeocystis sp.)
SRR5929440		single cell			Polycystinea	Nassellaria	*Lithomelissa*	*setosa*	non	
SRR5929439		single cell			Sticholonchea	Sticholonchida	*Sticholonche*	*zanclea*	non	

Figure 4.4 – Tableau récapitulatif des échantillons de Rhizaria symbiotiques et non-symbiotiques analysés dans l'article Meng et al *in prep.*, p.185. Les informations liées au séquençage des ARNs, à la taxonomie des échantillons ainsi qu'au trait fonctionnel de la symbiose sont indiqués.

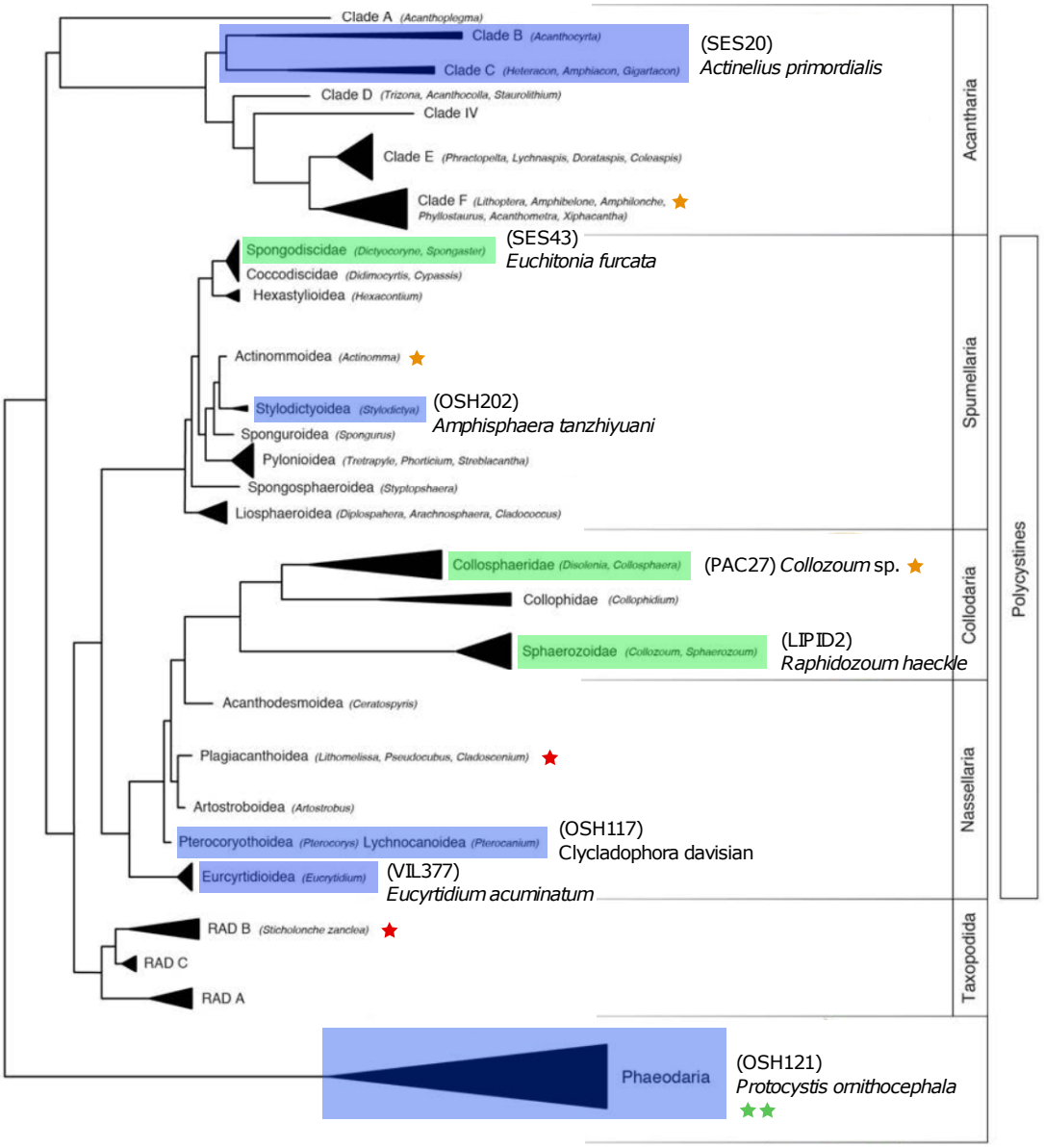


FIGURE 4.5 – Phylogénie schématique des Radiolaria représentant les 3 lignées majeurs de radiolaires (Acantharia, Polycystinea, Taxopodia) [SUZUKI et NOT 2015]. Les spécimens échantillonnés et analysés durant ma thèse sont indiqués : les encadrés verts correspondent aux lignées symbiotiques (holobiontes) et les encadrés bleus aux lignées non-symbiotiques. Le groupe des Phaeodaria (Cercozoa) est également représenté (en tant que groupe externe). Les spécimens de l'étude de [BALZANO et al. 2015] sont indiqués par une étoile orange, ceux étudiés dans [KRABBERØD, ORR et al. 2017] sont indiqués par une étoile rouge et les deux spécimens Phaeodaria sont indiqués par une étoile verte. (article Meng et al *in prep.*, p.185)

Key functions involved in the establishment and the maintenance of marine plankton symbiosis revealed by a meta-transcriptome approach

Arnaud Meng[1*], Erwan Corre[3], Pierre Peterlongo[2], Camille Marchet[2], Adriana Alberti[4,5], Corinne Da Silva[4,5], Patrick Wincker[4,5], Ian Probert[6], Noritoshi Suzuki[7], Stéphane Le Crom[1], Lucie Bittner[1*]† and Fabrice Not[6*]†

*** Correspondence:**
arnaud.meng@gmail.com; lucie.bittner@upmc.fr; not@sb-roscoff.fr
1 Institut de Biologie Paris Seine, University Pierre and Marie Curie, Quai Saint Bernard, 75005 Paris, France
Full list of author information is available at the end of the article
† Equal contributor

Author details
1 Sorbonne Universités, UPMC Univ Paris 06, Univ Antilles, Univ Nice Sophia Antipolis, CNRS, Evolution Paris Seine - Institut de Biologie Paris Seine (EPS - IBPS), 75005 Paris, France.
2 Institut de Recherche en Informatique et Systèmes Aléatoires, INRIA, Campus de Beaulieu, 263 avenue du Général Leclerc, 35042 Rennes, France.
3 ABiMS, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France.
4 Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France.
5 UMR 8030, CNRS, Evry, France.
6 UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France.
7 Institute of Geology and Paleontology, Graduate School of Science, Tohoku University, 6–3Aoba, Aramaki, Aoba-ku, Sendai, 980–8578 Japan

Introduction

Symbiosis is a widespread phenomenon in the biosphere. In the water column, planktonic organisms are key component of pelagic ecosystems and many species form mutualistic association with microalgae forming photosymbioses [1]. Here we intended to investigate the genetic basis of photosymbiosis, through a transcriptomic approach on marine plankton. We focused on associations occurring between radiolarian host (heterotrophic protist) and dinoflagellates symbionts (autotrophic or mixotrophic protist) living inside the host cell. Despite the fact that Radiolaria are mostly known for their exceptional and complex skeleton structure sketched by Ernst Haeckel [2], they exhibit ecologically significant symbiotic associations that remain poorly described at the genomic level. For many radiolarian species harboring microalgal symbionts, recently identified as dinoflagellates (*e.g.* *Gymnoxanthella Radiolarae*, *Brandtodinium nutricula*) [3, 4] and haptophytes (*Phaeocystis* spp.) [5], photosymbiosis is fully part of their functional capacities. These holobionts are widespread in the oligotrophic open oceans and have fundamental implications in biogeochemical carbon, silica and strontium cycles [6, 7]. In contrast with other ecologically significant marine plankton symbioses (*e.g.* coral or sponge symbiosis), little is known about the symbiosis between Radiolaria and Dinophyta.

This is essentially due to the lack of genomic information for most of non-model marine protists [8–10].

We used Short_Read_Connector_counter (SRC_c) [11] to disentangle RNA-seq reads of three holobionts transcriptomes into distinct subsets, among which, one is composed of reads from the symbionts and another of reads from the host. This step is crucial to limit mis-assemblies and the creation of chimeric sequences (Meng et al. *in prep. for Microbiome*). Independent analyses were performed in parallel on each subset: *de novo* assembly, followed by protein domains prediction and sequences functional annotation. Our strategy produced a comprehensive genomic dataset for Radiolaria since [12] and [9]. Moreover, our study generates newly assembled transcriptomes from holobiont systems, which enable the search for molecular sequences involved in the establishment and maintenance of symbiosis. These new sequences will be used for 1- phylogenomics investigation 2- reference for environmental metagenomic studies 3- characterizing the molecular basis of photosymbiotic relationships in the plankton.

Results

Holobiont reads sorting with SRC_c

Using SRC_c, holobiont reads were successfully compared to reference libraries involving either potential host or potential symbiont sequences. Two subsets of reads were subsequently created: a host subset involving reads similar to rhizarian sequences, and a symbiont subset involving reads similar to dinoflagellate sequences. An additional category (called "shared") corresponds to holobiont reads that find similarities with both rhizarian and dinoflagellates reference sequences. Finally, the remaining holobiont reads were gathered in the 'unassigned' category.

For the *Collozoum* sp. holobiont transcriptome (a total of 97,857,794 reads), 44% of the reads can be assigned either to the host, the symbiont or the shared category (Fig. 1a). A total of 23 % of the reads were assigned to the dinoflagellates, while 10% were assigned to the Rhizaria. Finally, 66% of the holobiont reads did not find similarities to either host of the symbiont libraries. For the *Euchitonia furcata* holobiont (33,899,424 reads), 15% of the reads are categorized as host (11%), symbiont (2%) and shared (3%) (Fig. 1b). Overall, 85% of the holobiont reads remained unassigned. In the last holobiont *Raphidozoum* sp., the total proportion of assigned reads correspond to 9% while the remaining reads that could not be assigned consists of 91% of the total holobiont reads. Though few reads of the total 66,263,508 holobiont reads could be assigned to the host (2%), the symbiont (6%) and were shared between host and symbiont (1%) (Fig. 1c).

Transcriptomes assembly and analyses of disentangled reads subsets

For each read subset from each holobiont, *de novo* transcriptome assembly, protein coding domains prediction and functional annotations were performed independently (Tab. 2). In the case of

Collozoum sp., a total of 164,514 contigs were assembled of which 0.65% are host contigs, 2.74% are symbiont contigs, 0.18% correspond to shared contigs and the major part (96.43%) are contigs assembled from unassigned reads (Tab. 2). We found that 6.3% and 28.8% of host and symbiont assembled contigs, respectively, contained protein coding domains of which 43.3% and 60.3% has been functionally annotated. A total of 26.7% of the shared contigs contained protein coding domains and functional annotations were detected for 83.5% of them (Tab. 2). Of the unassigned contigs, protein coding domains were detected for 44.7% of them and 61.7% of these domains were functionally annotated (Tab. 2). *Euchitonia furcata* shows the lowest amount of assembled contigs with a total of 1,558 contigs. Most of the assembled contigs (90.82%) belongs to the unassigned category, while 3.92% belongs to the host, 0.19% belongs to the symbiont (3 contigs) and 5.07% belongs to the shared category (Tab. 2). For the 3 symbiont contigs we could not detect protein coding domains, and hence the functional annotation step was not performed. We detected protein coding domains for 6.6% of the host contigs and for 24.1% of the shared contigs of which 25% and 68.4% were functionally annotated. Among the unassigned contigs, 9.4% contained protein coding domains of which 44.5% found functional annotations. *Raphidozoum* sp. constitutes the largest dataset with a total of 204,944 assembled contigs. Protein coding domains were detected for 5.1% and 44.6% of host and symbiont contigs respectively. 55.4% and 57% of these protein coding domains were functionally annotated. Among shared contigs, a proportion of 28.1% had protein coding domains and 63.5% of them has been functionally annotated. Finally, we detected protein coding domains for 30.5% of the unassigned contigs and functional annotations were detected in 62.2% of these protein coding domains.

Seeking for putative symbiosis family genes

[future analyses]

Discussion

Holobiont reads sorting with SRC_c

Compared to our previous study [Meng et al. *in prep for Microbiome*], the host reference library used to analyse the radiolaria-dinoflagellate holobionts was significantly improved in the present study (*i.e.* involving 39,853,158,029 pb (from 7 datasets) instead of 4,561,294 pb (from 4 datasets)) and we used reads instead of assembled sequences. This allows to minimize the potential bias inherent to the use of contigs obtained from transcriptome assembly which could contain potential chimeras and truncated sequences [13] that would impact the reads assignation by SRC_c.

Considering results of the reads assignations for *Collozoum* sp. (Fig. 1 A) which was also analyzed in [Meng et al. *in prep for Microbiome*], we found nearly 3 times more host reads (9,437,876 reads vs. 3,188,944 reads) than in our previous study. Otherwise, no significant difference was detected for reads assigned to symbiont and shared categories. Consequently, the number of reads that could not be

assigned to either host, symbiont or the shared categories has been reduced (64,684,184 reads (66% of the total holobiont reads) vs. 71,003,016 reads (72% of the total holobiont reads)). We suggest that the addition of 6 non-symbiotic radiolarians and 1 cercozoa reads datasets (from this study and [9]) allows to significantly improve the reads assignation step to the host category by SRC_c, confirming our previous conclusions [Meng et al. *in prep for Microbiome*].

Euchitonia furcata and *Raphidozoum* sp. showed small proportions of assigned reads to the symbionts categories (2% and 6% respectively) (Fig. 1 B and C). Our previous study [Meng et al. *in prep for Microbiome*] shows that a lack of data from taxonomically close species to the partners in reference library result in poor read assignations. Here we suggest that no representative species for the symbionts of *E. furcata* and *Raphidozoum* sp. were contained in the reference libraries. In both holobionts, presumed symbionts are *Brandtodinium nutricula* or *Gymnoxanthella radiolariae* (Tab. 1). However, 3 RNA-seq datasets (strains: *B. nutricula* RCC3387, RCC3468 and *G. radiolariae* RCC3507) are included in symbiont reference library. Considering the read assignation results of SRC_c for the three holobionts (Fig. 1), and considering that it was successful for *Collozoum* sp. (i.e. 23% of the data), we suggest that these dinoflagellates strains are contained in *Collozoum* sp. but not in our specimens of *E. furcata* and *Raphidozoum* sp..

In order to improve these proportions of assignations, several strategies can be implemented. A SRC_c parameters tuning could help to increase the detection of similar k-mers between holobiont and reference libraries. This might result in more holobiont reads assigned to both host and symbiont categories. Moreover, we suggest that a sequential enrichment algorithm as previously described in [Meng et al. *in prep for Microbiome*] could also impact and improve the overall proportion of assigned reads to both host and symbiont categories. Finally, the addition of dinoflagellates and radiolarians/rhizarian data from additional strains that are not already contained in the current reference libraries could enhance the reads assignation step with SRC_c.

Seeking for putative symbiosis genes

[future analyses]

Materials & Methods

Sampling & sequencing protocol

Three radiolarian holobionts were sampled and sequenced. The first holobiont is a *Collozoum* sp. sampled in 2011 in the south Pacific Ocean (see Meng et al. 2017 *submitted* for more details). The other two (unpublished) holobionts correspond to *Euchitonia furcata* (strain SES43) and *Raphidozoum* sp. (strain LIPID2). *E. furcata* was sampled in 2011 at Sesoko island (Okinawa, Japan), and [X sampling details X]. A total of 68 cells were sorted manually from the plankton mix using a stereomicroscope by pipetting out of the samples, one by one, the cells of interest to be sequenced. For *Raphidozoum* sp. a colony of cell was collected from sea water in 2013 at Villefranche sur Mer

(France). [ADD: sampling details] (Sup. 1). Additionally, 5 rhizaria were sampled between 2011 and 2013, that include *Actinelius primordialis* (Radiolaria, Acantharia; strain SES20), *Protocystis ornithocephala* (Cercozoa, Phaeodaria; strain OSH121), *Amphisphaera tanzhiyuani* (Radiolaria, Spumellaria; strain OSH202), *Eucyrtidium acuminatum* (Radiolaria, Nasselaria; strain VIL377) and *Cycladophora davisiana* (Radiolaria, Nasselaria; strain OSH117) (Sup. 1). [ADD: sampling details for the 5 samples].

[ADD: Details sequencing protocol].

Host/symbiont library construction and the use of Short Read Connector to disentangle holobiont reads

RNA-seq raw paired reads data from rhizaria and dinoflagellate specimens were pooled in a host library and a symbiont library respectively. Details on reference libraries can found in Sup. 3. The host library contained data from 6 radiolarians (of which 4 are yet unpublished), and from 1 cercozoa (of which one is unpublished). Two radialoria RNA-seq raw reads (300 bp) datasets were retrieved from [9]: 11,590,658 reads of *Lithomelissa Setosa* (Polycystinea, Nasselaria) and 19,894,654 reads of *Sticholonche zanclea* (Sticholonchea, Sticholonchida) (https://www.ncbi.nlm.nih.gov/sra/?term=SRP115316). Four unpublished RNA-seq raw reads (150 bp) datasets of non-symbiotic radiolarians and of one unpublished cercozoa has been added to the host library, including 154,117,366 reads of *Actinelius Primordialis* (Acantharea, Acantharia), 175,624,508 reads of *Amphisphaera tanzhiyuani* (Polycystinea, Spumellaria), 67,893,922 reads of *Eucyrtidium acuminatum* (Polycystinea, Nasselaria), 40,834,536 reads of *Cycladophora Davisiana* (Polycystinea, Nasselaria) and 54,976,032 reads of *Protocystis omithocephala* (Filosa, Phaeodaria). The total host library content then included 524,931,676 paired-end reads. The symbiont library exploited in [Meng et al. 2017] has been reused in this study, it encompasses 123 RNA-seq raw paired reads (a total of 5,563,498,607 reads) of dinoflagellates species from both the MMETSP project [14] and 4 that have been recently published in [Meng et al. 2017].

Short Read Connector [11] in its latest counter version (SRC_c) (https://github.com/GATB/short_read_connector, commit 94aa6a65b5ddf61eba95108069fae29c41e51fb0) was used to compare holobiont read datasets to either the host or the symbiont read library described above. The 3 holobiont datasets (RNA-seq paired reads, 30-101 bp) included 97,857,794 reads of *Collozoum* sp. (Polycystinea, Collodaria), 61,263,508 reads of *Rhaphidozoum* sp. (Polycystinea, Collodaria) and 33,899,424 of *Euchitonia furcata* (Polycystinea, Spumellaria). Default parameters of SRC_c has been used except for two parameters: the k-mer size value was set to 25 (default is 30), to ensure that SRC_c could calculate k-mer from short reads (*i.e*. shorter than 30 bp) that occur in holobiont and library reads. The coverage parameter *s*

was set to 50, each reads of size *l* must have *l* x *s* positions covered by an indexed k-mer from library to be declare similar to this indexed k-mer. This protocol has been detailed in [Meng et al. *in prep. for Microbiome*].

Transcriptomes assembly & downstream analyses

A RNA-seq analysis pipeline written in Python and using the Snakemake workflow management system (http://snakemake.readthedocs.io/en/stable/) has been used to process the subsets of reads resulting from the Short Read Connector step. This pipeline is publicly available from on a GitHub repository: https://github.com/arnaudmeng/dntap. It includes 4 analysis steps involving published softwares for which versions are listed in the GitHub repository. The reads filtering is done with the Trimmomatic software [15] for which custom parameters has been used as follow: SLIDINGWINDOW:10:20 to remove reads with quality lower than 20 into each 10 bases windows of a read. The filtered reads are then assembled with the Trinity software [16] using default parameters. Trandescoder [17] performs then protein coding domains predictions of the assembled contigs. Finally, the functional annotations of predicted coding domains is done with InterProScan 5 [18].

Sequence Similarity Network (SSN) building

We retrieved 3,485,893 sequences that encompasses 57 proteomes of dinoflagellates (12 from symbiotic lineages and 45 from non-symbiotic lineages) corresponding to 47 species [Meng et al. 2017 *submitted*] and added them to the radiolarian protein coding domains generated and described in the previous sections. To build a SSN [19, 20], sequence alignments were computed with DIAMOND [21] between all protein coding domains from the 60 datasets (57 dinoflagellates, 6 radiolaria holobiontes, 6 heterotrophic radiolaria, 2 cercozoa) with e-value of 1e-25. Different similarity thresholds were tested in order to choose an optimal threshold according to the two following criteria: maximizing the number of large CCs (i.e. minimum of 30 vertices) and the number of CCs involving a single homogeneous functional annotation (i.e. a unique GOslim term at the Biological Process level) (as in [Meng et al. 2017 *submitted*]). We used the igraph R Cran package [22] to generate the SSN, to compute and sort the connected components (CCs).

Statistics and analyses based on the CCs

[future analyses]

Figures & Tables

Fig. 1: Results of the holobiont reads assignation with Short Read Connector. The table shows the number of reads assigned in each of the four categories defined from the holobiont (*i.e.* host, symbiont, shared and unassigned). The pie charts displayed the corresponding proportions of the reads (a) corresponds to *Collozoum* sp., (b) *Euchitonia furcata* and (c) *Raphidozoum* sp..

Tab. 1: Table summarizing taxonomic and functional trait information for rhizarian specimens used in this study.

Tab. 2: Results for *de novo* assembly, coding domains prediction and functional annotations steps. For the 3 holobionts datasets, the 4 reads subsets (host, symbiont, shared and unassigned) were assembled independently. Metrics including the number of assembled contigs, the size of the smallest and of the longest contig, the total number of bases, the N50 & mean length of contigs, and the GC content are displayed. The number of coding domains predicted and the number of functional annotated coding domains are shown. The proportions of the 4 categories of assembled contigs were displayed among the total assembled contigs per holobiont. The proportions of predicted coding domains and of functional annotated contigs detected in each subset are shown. Two mapping rates are displayed, the first corresponds to the remapping rates of contigs subsets on the corresponding assemble contigs while the second shows the proportion of the total holobiont reads mapped to each category of assembled contigs.

Supplementary materials

Sup. 1: Table showing sequencing and sampling information for the rhizarian data used in this study. Sources of published data are displayed.

Sup. 2: Phylogenetic positions of the rhizarian data used in this study (modified from Suzuki et Not 2015). Newly sequenced holobionts and non-symbiotic radiolarian datasets are indicated with their respective strains identifier. Green boxes correspond to symbiotic lineages (holobionts) and blue boxes for non-symbiotic lineages. Datasets from Balzano et al. 2015 are indicated with orange stars and a green star while datasets from Krabberod et al. 2017 are indicated with red stars.

Sup. 3: Table of read statistics and data sources.

Bibliography

1. Decelle J, Colin S, Foster RA. Photosymbiosis in Marine Planktonic Protists. In: Ohtsuka S, Suzaki T, Horiguchi T, Suzuki N, Not F, editors. Marine Protists. Springer Japan; 2015. p. 465–500. doi:10.1007/978-4-431-55130-0_19.

2. Haeckel E. Kunstformen der Natur. Verlag des Bibliographischen Instituts; 1904.

3. Yuasa T, Horiguchi T, Mayama S, Takahashi O. Gymnoxanthella radiolariae gen. et sp. nov. (Dinophyceae), a dinoflagellate symbiont from solitary polycystine radiolarians. J Phycol. 2016;52:89–104.

4. Probert I, Siano R, Poirier C, Decelle J, Biard T, Tuji A, et al. Brandtodinium gen. nov. and B. nutricula comb. Nov. (Dinophyceae), a dinoflagellate commonly found in symbiosis with polycystine radiolarians. J Phycol. 2014;50:388–99.

5. Decelle J, Siano R, Probert I, Poirier C, Not F. Multiple microalgal partners in symbiosis with the acantharian Acanthochiasma sp. (Radiolaria). Symbiosis. 2012;58:233–44.

6. Biard T, Stemmann L, Picheral M, Mayot N, Vandromme P, Hauss H, et al. In situ imaging reveals the biomass of giant protists in the global ocean. Nature. 2016;advance online publication. doi:10.1038/nature17652.

7. Guidi L, Chaffron S, Bittner L, Eveillard D, Larhlimi A, Roux S, et al. Plankton networks driving carbon export in the oligotrophic ocean. Nature. 2016;532:465–70.

8. Sibbald SJ, Archibald JM. More protist genomes needed. Nat Ecol Evol. 2017;1:0145.

9. Krabberød AK, Orr RJS, Bråte J, Kristensen T, Bjørklund KR, Shalchian-Tabrizi K. Single Cell Transcriptomics, Mega-Phylogeny, and the Genetic Basis of Morphological Innovations in Rhizaria. Mol Biol Evol. 2017;34:1557–73.

10. Burki F, Keeling PJ. Rhizaria. Curr Biol. 2014;24:R103–7.

11. Marchet C, Limasset A, Bittner L, Peterlongo P. A resource-frugal probabilistic dictionary and applications in (meta)genomics. ArXiv160508319 Cs Q-Bio. 2016. http://arxiv.org/abs/1605.08319. Accessed 7 Jun 2017.

12. Balzano S, Corre E, Decelle J, Sierra R, Wincker P, Da Silva C, et al. Transcriptome analyses to investigate symbiotic relationships between marine protists. Microb Physiol Metab. 2015;6:98.

13. Toseland A, Moxon S, Mock T, Moulton V. Metatranscriptomes from diverse microbial communities: assessment of data reduction techniques for rigorous annotation. BMC Genomics. 2014;15:901.

14. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, 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. 2014;12:e1001889.

15. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics. 2014;:btu170.

16. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome (Trinity). Nat Biotechnol. 2011;29:644–52.

17. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.

18. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinforma Oxf Engl. 2014;30:1236–40.

19. Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO. Gene similarity networks provide tools for understanding eukaryote origins and evolution. Proc Natl Acad Sci U S A. 2013;110:E1594-1603.

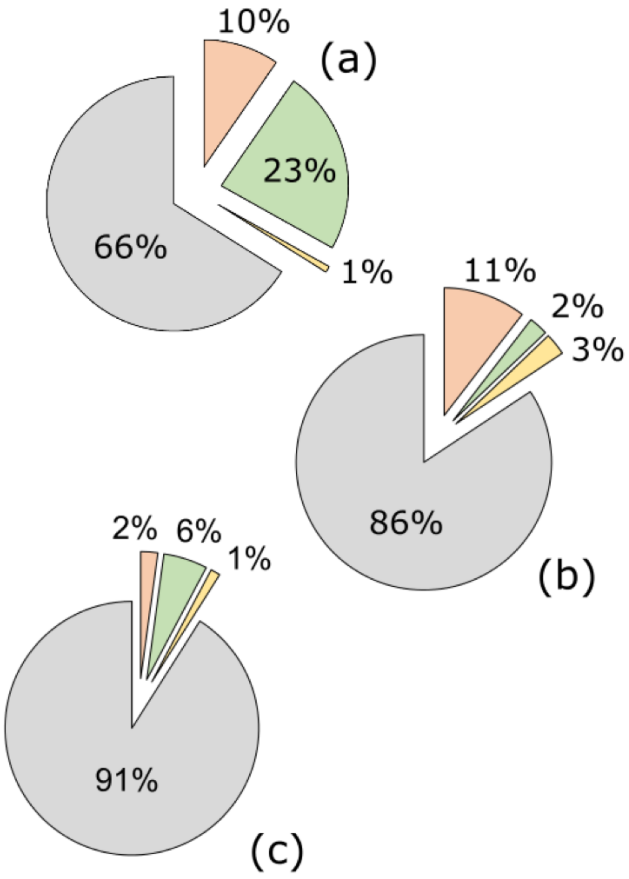
20. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. PLoS ONE. 2009;4. doi:10.1371/journal.pone.0004345.

21. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.

22. Csárdi G, Nepusz T. The igraph software package for complex network research. InterJournal Complex Syst. 2006. http://wbldb.lievers.net/10011687.html. Accessed 20 Feb 2017.

Fig. 1

		# reads
***Collozoum* sp.** (a)	**total**	**97 857 794**
	assigned to host	9 437 876
	assigned to symbiont	22 964 914
	assigned to shared	800 820
	unassigned	64 654 184
Euchitonia furcata (b)	**total**	**33 899 424**
	assigned to host	3 607 726
	assigned to symbiont	833 306
	assigned to shared	966 082
	unassigned	29 089 308
***Raphidozoum* sp.** (c)	**total**	**61 263 508**
	assigned to host	1 331 716
	assigned to symbiont	3 447 578
	assigned to shared	756 954
	unassigned	56 015 868



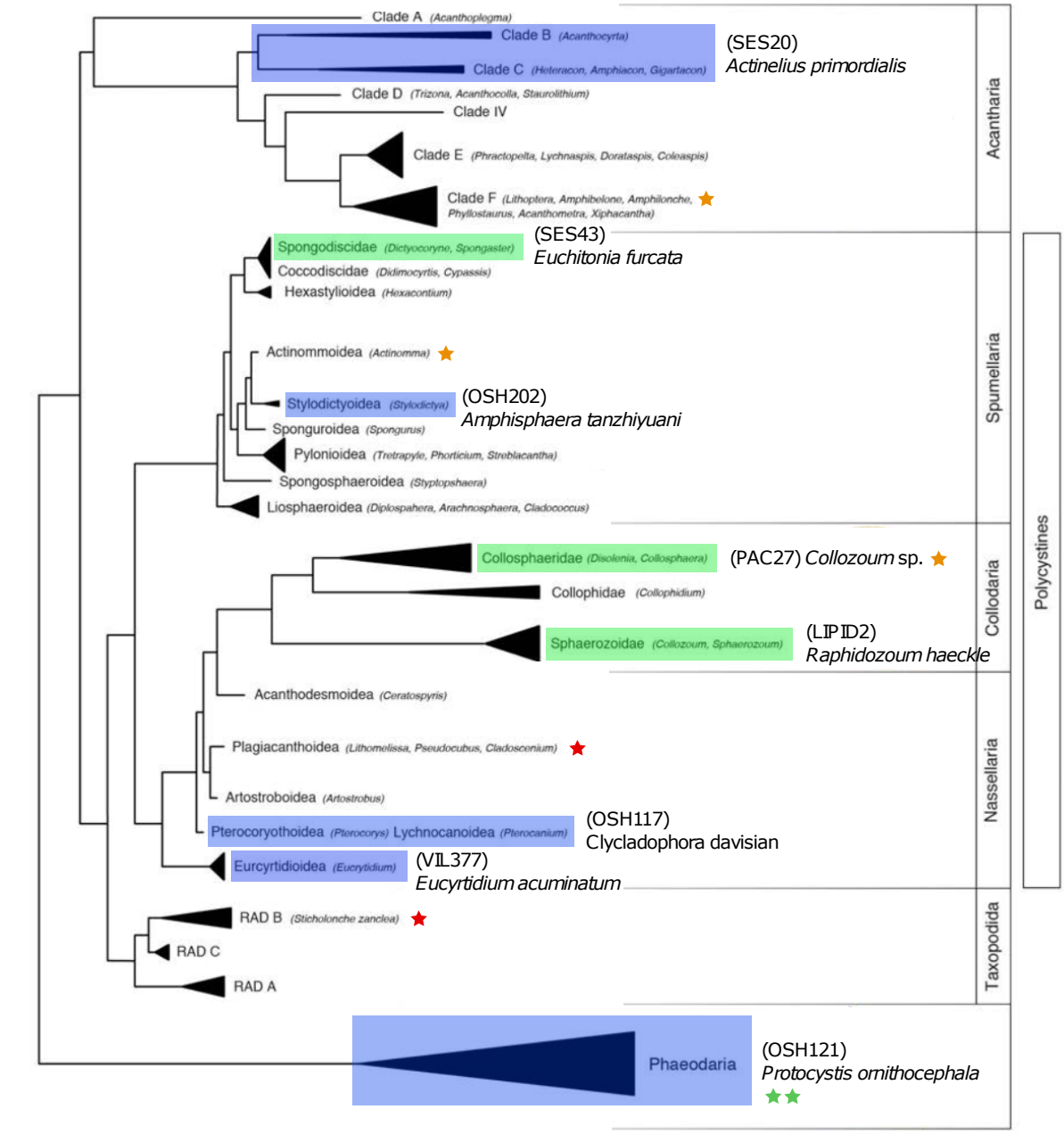
Tab. 1

Code		Taxonomy				Functional trait	
		Class	**Order**	**Genus**	**Species**	**Photo-Symbiotic**	**Symbionts**
BDQ	Radiolaria	Polycystinea	Collodaria	*Collozoum*	sp.	yes	dinoflagellates (B. nutricula, G. radiolaroae)
BFT	Radiolaria	Polycystinea	Spumellaria	*Euchitonia*	*furcata*	yes	dinoflagellates (B. nutricula, G. radiolaroae)
BRX	Radiolaria	Polycystinea	Collodaria	*Raphidozoum*	sp.	yes	dinoflagellates (B. nutricula, G. radiolaroae)
BRW	Radiolaria	Acantharea	Acantharia	*Actinelius*	*primordialis*	no	
BGU	Cercozoa	Filosa	Phaeodaria	*Protocystis*	*ornithocephala*	no	
BRY	Radiolaria	Polycystinea	Spumellaria	*Amphisphaera*	*tanzhiyuani*	no	
BGX	Radiolaria	Polycystinea	Nassellaria	*Eucyrtidium*	*acuminatum*	no	
BRR	Radiolaria	Polycystinea	Nassellaria	*Cycladophora*	*davisiana*	no	
ASRAAA	Cercozoa	Filosa	Phaeodaria	*Aulacantha*	*scolymantha*	no	
ASRAAB	Radiolaria	Polycystinea	Collodaria	*Collozoum*	sp.	yes	dinoflagellates (B. nutricula, G. radiolaroae)
ASRAAC	Radiolaria	Polycystinea	Spumellaria	*Spongosphaera*	*streptacantha*	yes	?
ASRAAD	Radiolaria	Acantharea	Acantharia	*Amphilonche*	*elongata*	yes	Haptophyta (Phaeocystis sp.)
SRR5929440	Radiolaria	Polycystinea	Nassellaria	*Lithomelissa*	*setosa*	no	
SRR5929439	Radiolaria	Sticholonchea	Sticholonchida	*Sticholonche*	*zanclea*	no	

Tab. 2

		# contigs	% contigs in holobiont	smallest	largest	# bases	N50	mean len	%GC	# with ORFs	# predicted cds	% contigs with predicted cds	# annotated cds	% cds with functional annotations
	host	1,070	0.65	201	3,337	327,475	267	306.05	0.4	4	67	6.3	29	43.3
	symbiont	4,511	2.74	201	1,874	1,455,020	317	332.55	0.55	518	1,299	28.8	783	60.3
***Collozoum* sp.**	**shared**	296	0.18	201	993	93,479	304	315.81	0.44	25	79	26.7	66	83.5
	unassigned	158,637	96.43	201	8,288	92,554,970	718	583.4	0.41	48,162	70,878	44.7	43,750	61.7
	total	**164,514**								**48,709**	**72,323**	44.0	**44,628**	61.7
	host	61	3.92	201	456	16,222	256	265.93	0.43	0	4	6.6	1	25.0
	symbiont	3	0.19	206	224	651	224	217	0.46	0	0	0	0	
Euchitinia furcata	**shared**	79	5.07	202	1,177	28,188	359	356.81	0.42	5	19	24.1	13	68.4
	unassigned	1,415	90.82	201	544	351,467	241	248.39	0.49	1	123	8.7	51	41.5
	total	**1,558**								**6**	**146**	9.4	**65**	44.5
	host	1,441	0.70	201	3,826	443,005	279	307.43	0.38	10	74	5.1	41	55.4
	symbiont	25,330	12.36	201	2,720	10,757,061	480	424.68	0.64	6,840	11,286	44.6	6,431	57.0
Raphidozoum haeckle	**shared**	540	0.26	201	1,005	173,895	315	322.03	0.51	68	152	28.1	71	46.7
	unassigned	17,7633	86.67	201	9,617	81,798,874	559	460.49	0.42	32,201	50,913	28.7	32,314	63.5
	total	**204,944**								**39,119**	**62,425**	30.5	**38,857**	62.2

Code	Sequencing information				Sampling information			
	ID (strain)	tech	Lib	info	Date	Localisation	expedition	source
BDQ	PAC37	Illumina	OVATION	colony	2011	south Pacific Ocean (station 112.01)	*Tara* Oceans	
BFT	SES43	Illumina	OVATION	68 cells	2011	Sesoko island, Okinawa, Japan		
BRX	LIPID2	Illumina		colony	2013	Villefranche sur Mer, France		
BRW	SES20	Illumina		Few cells	2011	Sesoko island, Okinawa, Japan		
BGU	OSH121	Illumina	SMARTER	160 cells	2013		Oshoro Maru	
BRY	OSH202	Illumina		33 cells	2013		Oshoro Maru	
BGX	VIL377	Illumina	SMARTER	49 cells	2012	Villefranche sur Mer, France		
BRR	OSH117	Illumina		37 cells	2013		Oshoro Maru	
ASRAAA		454	SMARTER	50 cells	NC	Villefranche sur Mer, France		Balzano et al. 2015
ASRAAB		454	SMARTER	colony	NC	Villefranche sur Mer, France		Balzano et al. 2015
ASRAAC		454	SMARTER	50 cells	NC	Villefranche sur Mer, France		Balzano et al. 2015
ASRAAD		454	SMARTER	150 cells	NC	Gulf of Eilat, red sea		Balzano et al. 2015
SRR5929440		single cell			2014	inner part of the Oslo fjord		Krabberød et al. 2017
SRR5929439		single cell			2014	inner part of the Oslo fjord		Krabberød et al. 2017


Clade A (Acanthoplegma)
Clade B (Acanthocyrta)
Clade C (Heteracon, Amphiacon, Gigartacon)
(SES20)
Actinelius primordialis
Clade D (Trizona, Acanthocolla, Staurolithium)
Clade IV
Clade E (Phractopelta, Lychnaspis, Dorataspis, Coleaspis)
Clade F (Lithoptera, Amphibelone, Amphilonche, Phyllostaurus, Acanthometra, Xiphacantha)
Acantharia
Spongodiscidae (Dictyocoryne, Spongaster)
(SES43)
Euchitonia furcata
Coccodiscidae (Didimocyrtis, Cypassis)
Hexastylioidea (Hexacontium)
Actinommoidea (Actinomma)
Stylodictyoidea (Stylodictya)
(OSH202)
Amphisphaera tanzhiyuani
Sponguroidea (Spongurus)
Pylonioidea (Tretrapyle, Phorticium, Streblacantha)
Spongosphaeroidea (Styptopshaera)
Liosphaeroidea (Diplospahera, Arachnosphaera, Cladococcus)
Spumellaria
Collosphaeridae (Disolenia, Collosphaera)
(PAC27) Collozoum sp.
Collophidae (Collophidium)
Sphaerozoidae (Collozoum, Sphaerozoum)
(LIPID2)
Raphidozoum haeckle
Collodaria
Polycystines
Acanthodesmoidea (Ceratospyris)
Plagiacanthoidea (Lithomelissa, Pseudocubus, Cladoscenium)
Artostroboidea (Artostrobus)
Pterocoryothoidea (Pterocorys) Lychnocanoidea (Pterocanium)
(OSH117)
Clycladophora davisian
Eurcyrtidioidea (Eucyrtidium)
(VIL377)
Eucyrtidium acuminatum
Nassellaria
RAD B (Sticholonche zanclea)
RAD C
RAD A
Taxopodida
Phaeodaria
(OSH121)
Protocystis ornithocephala

Sup. 2

Code			Data information		
	Genus	Species	# reads	read sizes (bp)	source
BRW	*Actinelius*	*primordialis*	154,117,366	30-101	this study
BGU	*Protocystis*	*ornithocephala*	54,976,032	30-101	this study
BRY	*Amphisphaera*	*tanzhiyuani*	175,624,508	30-101	this study
BGX	*Eucyrtidium*	*acuminatum*	67,893,922	30-101	this study
BRR	*Cycladophora*	*davisiana*	19,576,340	30-101	this study
ASRAAA	*Aulacantha*	*scolymantha*	195,070	200	Balzano et al. 2015
ASRAAB	*Collozoum*	sp.	214,475	200	Balzano et al. 2015
ASRAAC	*Spongosphaera*	*streptacantha*	220,239	200	Balzano et al. 2015
ASRAAD	*Amphilonche*	*elongata*	195,890	200	Balzano et al. 2015
SRR5929440	*Lithomelissa*	*setosa*	11,590,658	300	Krabberød et al. 2017
SRR5929439	*Sticholonche*	*zanclea*	19,894,654	300	Krabberød et al. 2017
		Total	504,499,154		

Le séquençage transcriptomique de 3 holobiontes et de 5 spécimens non-symbiotiques de Rhizaria (dont 7 radiolaires) représente un ensemble de données conséquent et inédit pour l'étude de cette lignée (Figures 4.4 et 4.5) [Caron 2016]. A l'heure actuelle, les 3 jeux d'holobiontes ont été traités avec `SRC_c`, puis chacun des 4 sous-jeux de reads obtenus a été assemblé *de novo*. Des domaines fonctionnels prédits ainsi que les annotations fonctionnelles ont été recherchés dans les contigs* résultants. Les 5 jeux de Rhizaria non-symbiotiques suivent actuellement les mêmes étapes d'analyse. De plus, deux transcriptomes de radiolaires non-symbiotiques (*Lithomelissa setosa* (Polycystinea, Nassellaria) et *Sticholonche zanclea* (Taxopodia)) publiés dans [Krabberød et al. 2017] sont également ré-assemblés à l'aide de cette même chaîne d'analyses, afin de conserver une homogénéité des méthodes d'assemblage puis d'annotation fonctionnelle. Au total, 10 transcriptomes/jeux de séquences protéiques de Rhizaria (dont 8 nouveaux) seront générés. A ces jeux, seront ajoutées les séquences protéiques issues de 4 transcriptomes de Rhizaria [Balzano et al. 2015] : 3 radiolaires symbiotiques (*Collozoum* sp. (Polycystinea, Collodaria), *Spongosphaera streptacantha* (Polycystinea, Spumellaria) et *Amphilonche elongata* (Acantharea)) et un cercozoaire non-symbiotique *Aulacantha scolymantha*. Ces données ont été obtenues à partir de l'assemblage *de novo* de séquences issues d'un séquençage 454 en utilisant un assembleur adapté à la longueur des lectures (*i.e.* Newbler implémentant une approche de *overlap-layout-consensus* (cf. section 1.4) [Margulies et al. 2005]). Ainsi, dans ces séquences assemblées, seront recherchés les domaines protéiques ainsi que les annotations fonctionnelles à l'aide de notre chaîne d'analyses (étape 4 de la chaîne d'analyses, cf. chapitre 3) afin de conserver l'homogénéité de l'annotation fonctionnelle vis à vis des 10 autres jeux de données.

L'ensemble des séquences protéiques des 14 jeux de données décrits ci-dessus ainsi que les séquences protéiques obtenues grâce à l'assemblage des transcriptomes de dinoflagellés en culture (cf. chapitre 3) Article Meng et al *submitted*, p.55] seront analysées via une analyse de réseaux de similarité de séquences (SSN*). Les annotations fonctionnelles et les informations de traits fonctionnels (caractère symbiotique ou non-symbiotique de chaque échantillon de radiolaire) seront ajoutés aux noeuds du réseau (Figure 4.6 A). Les composantes connexes (CCs) composées exclusivement

de domaines de radiolaires symbiotiques ainsi que de dinoflagellés symbiotiques seront recherchées (CCs taxonomiques mixtes, Figure 4.6 B). L'étude des annotations fonctionnelles présentes dans ces composantes permettra d'établir des hypothèses quant à l'importance des fonctions observées dans les mécanismes régissant la symbiose entre radiolaires et dinoflagellés. Il sera également possible de rechercher des marqueurs de symbiose déjà identifiés dans ces holobiontes : *e.g.* la présence de lectine dans les holobiontes radiolaire-dinoflagellés [Balzano et al. 2015], ou encore les séquences des 90 794 CCs « marqueurs » de symbiose (Article Meng et al *submitted*, p.55). Nous pourrons également identifier parmi ces CCs taxonomiques mixtes symbiotiques, celles dont aucun domaine n'est encore annoté. Des études plus approfondies de ces séquences seront alors nécessaires pour établir le caractère nouveau de ces domaines chez les radiolaires et chez les dinoflagellés symbiotiques.

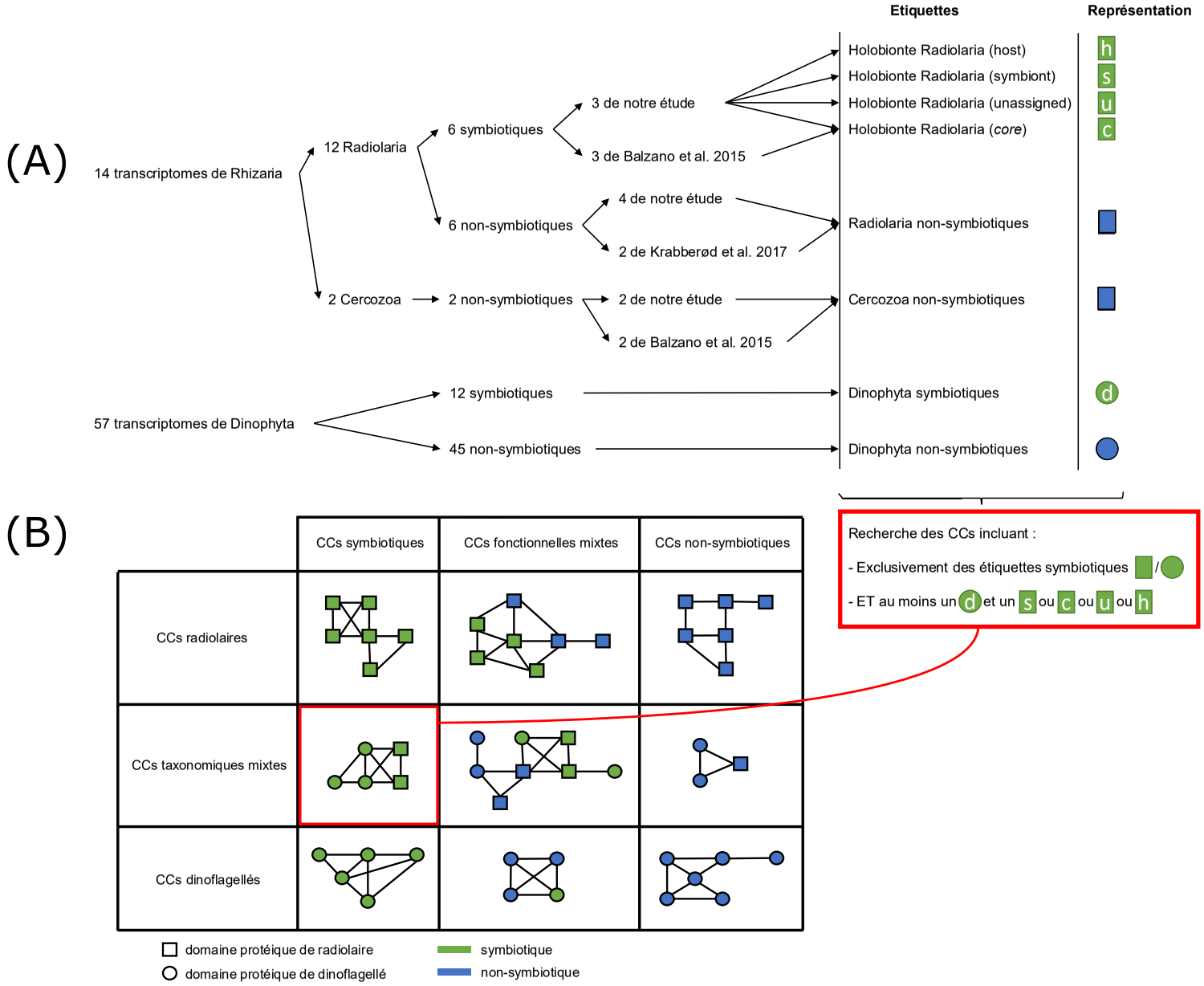


FIGURE 4.6 – Représentation schématique des composantes connexes étudiées dans l'article Meng et al *in prep.*, p.185. Les domaines protéiques de Rhizaria sont symbolisés pas des carrés et ceux des dinoflagellés par des ronds, et le trait fonctionnel symbiotique en vert et non-symbiotique en bleu. Dans cet exemple, les CCs composées de domaines protéiques de Rhizaria et de dinoflagellés sont appelés CCs taxonomiques mixtes. Les CCs incluant des domaines protéiques de lignées symbiotiques et non-symbiotiques sont appelées CCs fonctionnelles mixtes. Nos analyses se focaliseront principalement sur l'étude des CCs taxonomiques mixtes symbiotiques.

Chapitre 5

Conclusions et perspectives

Aujourd'hui, les données de génomique s'accumulent rapidement grâce aux projets de séquençage de grande envergure d'échantillons environnementaux. Toutefois, la majorité des lignées planctoniques eucaryotes ne possèdent toujours pas de références génomiques disponibles dans les bases de données publiques ce qui entrave sérieusement leur étude (Figures 1.3 et 1.7). Ce constat s'explique essentiellement par la difficulté de (1) collecter et isoler ces organismes depuis l'environnement, et (2) des les maintenir en culture. De plus, l'accès à leur matériel génétique est difficile et la faible quantité d'ADN ou d'ARN qu'il est possible d'extraire à partir de ces protistes est encore limitée. Les avancées dans les techniques d'extraction de matériel génomique et de séquençage sur cellule unique (*single-cell*) sont rapides et permettent d'envisager à relativement moyen terme de travailler à partir de quantités d'ADN ou d'ARN réduites provenant d'échantillon *in situ*, mais aussi d'observer les variations inter-organismes fines qui participent à l'hétérogénéité au sein de population cellulaires [Martinez-Garcia et al. 2012 ; Yoon et al. 2011]. A l'heure actuelle, l'étude de ces lignées planctoniques représentent un défi pour lequel il est nécessaire de développer des stratégies d'analyses afin d'accroître nos connaissances sur la biologie de ces organismes. Au cours de ce travail de thèse, je me suis intéressé à l'étude de la symbiose entre deux lignées de protistes, les radiolaires et les dinoflagellés. Les projets d'études génomiques de ces organismes planctoniques marins et de leur interactions sont encore rares car ils n'échappent pas aux contraintes évoquées ci-dessus et, de fait, nos connaissances sur l'évolution et l'écologie de ces lignées sont encore très limitées. Cependant au début de ma thèse un certain nombre de données avait été générées. Je me suis donc attaché à mettre en place des protocoles pour traiter ces données provenant du séquençage des ARNs d'organismes non-modèles* afin de les analyser dans le contexte de la caractérisation fonctionnelle des symbioses.

5.1 Approches bioinformatiques pour l'étude *de novo* de transcriptomes d'holobiontes

La première partie de mes travaux m'a permis d'appréhender les méthodes liées aux techniques d'assemblage *de novo* ainsi qu'à leurs difficultés intrinsèques. La construction de cette chaîne d'analyses m'a offert dans un premier temps la possibilité de reproduire facilement les expériences (ici une série d'applications d'outils pour le traitement et l'analyse de séquences) pour les jeux de données RNA-seq. La reproductibilité des analyses en bio-informatique est indispensable en recherche et permet une plus grande transparence vis à vis des analyses réalisées. Par ailleurs, l'utilisation d'une chaîne d'analyses assure un traitement homogène pendant toute la durée de vie d'un projet pour l'ensemble des données qui sont produites et analysées. Cette chaîne d'analyses m'a permis de réaliser un traitement identique des données RNA-seq de 123 échantillons de dinoflagellés (cf. chapitre 3) mais aussi des 4 ensembles de lectures de chaque transcriptome d'holobionte résultants de l'outil `SRC_c` ((cf. chapitre 4)) (un total de 12 ensembles de lectures pour 3 holobiontes). L'homogénéité des assemblages ainsi que des annotations fonctionnelles issues du traitement de ces transcriptomes a rendu possible la comparaison systématique des contenus en fonctions de chaque jeu de données.

Le début de la mise en place de cette chaîne d'analyses date de 2014, mais les outils bio-informatiques qui y sont intégrés sont toujours largement utilisés à l'heure actuelle. Par exemple, Trinity reste l'assembleur *de novo* de transcriptomes le plus cité en 2017 (Figure 1.10) et la dernière version de InterProScan date du 28 septembre 2017. De fait, elle peut être employée pour le traitement de données RNA-seq dans le cadre d'autres projets de transcriptomique, et sa récente intégration au serveur de calcul de la plateforme bioinformatique de Roscoff (`http://abims.sb-roscoff.fr/`) permet son utilisation par le plus grand nombre. Elle reste de plus disponible sur le dépôt de projet GitHub accompagnée d'une notice d'informations et d'utilisation détaillée : `https://github.com/arnaudmeng/dntap`. Grâce à sa structure modulaire, cette chaîne d'analyses pourra intégrer de nouvelles fonctionnalités. Par exemple, il serait envisageable d'ajouter une étape dédiée à la

détection des chimères* créées au cours de l'étape d'assemblage. D'autres chaînes d'analyses ont été développées pour le traitement et l'analyse *de novo* de transcriptome notamment pour l'assemblage des données issues de la MMETSP. Le projet de Lisa Cohen, étudiante en thèse dans le laboratoire de Titus Brown affilié à l'université de Californie à Davis (États-Unis), a pour but de ré-assembler la totalité des 678 transcriptomes de microbes marins issus du projet MMETSP. La chaîne d'analyses developpée par Lisa Cohen (https://github.com/ljcohen/MMETSP) et mise a disposition depuis février 2017, est constituée d'une série de scripts indépendants qui ne sont pas chaînés au travers d'un système de gestion tel que Snakemake ou Galaxy, et n'utilise pas de fichier de configuration. En comparaison, la chaîne d'analyses que je propose inclut ces fonctionnalités qui la rendent plus accessible aux utilisateurs débutants en programmation et permet une meilleure traçabilité des expériences réalisées au travers des fichiers *log*. Finalement, l'utilisation de la chaîne d'analyses que j'ai développé pourrait être facilitée au travers d'un système de déploiement tel que Docker (https://www.docker.com/). Ce système permet une installation rapide d'un programme et son utilisation au travers d'un environnement virtuel dans lequel le programme est déjà pré-installé. Les utilisateurs pourraient utiliser localement la chaîne d'analyses que je propose sans installer les différents programmes qui la compose, étape souvent difficile (pour des débutants) et chronophage.

L'étude des holobiontes m'a amenée à mettre en place un protocole d'analyse permettant, à partir d'un transcriptome d'holobionte, de séparer les séquences des partenaires symbiotiques. Une variété de stratégies d'analyses existent actuellement dans le but de pallier aux biais inhérents à l'étude de transcriptomes composés de lignées hétérospécifiques (*e.g.* la création de chimères* au cours de l'étape d'assemblage). Certaines de ces stratégies évitent par exemple l'étape d'assemblage *de novo* et se concentrent exclusivement sur le *mapping** des lectures pour identifier les lignées taxonomiques et les fonctions les plus abondantes au sein d'un transcriptome d'holobionte ou d'un méta-transcriptome [Martinez et al. 2016 ; Westreich et al. 2016]. Des études d'analyses différentielles sont alors possibles sous certaines conditions, mais ces stratégies ne permettent pas de générer des références génomiques et de comparer ces nouveaux transcrits assemblés aux bases de données publiques

(*e.g.* nr NCBI, SwissProt). D'autres stratégies utilisent des approches statistiques ou des algorithmes de *machine-learning** pour améliorer la qualité des assemblages *de novo* de méta-transcriptomes ou de transcriptomes d'holobiontes en s'appuyant sur les données d'abondance de séquences [Mohsen, Tang et Ye 2017], mais elles ne permettent cependant pas d'identifier l'appartenance des séquences aux différents acteurs d'un méta-transcriptome ou d'un holobionte.

L'utilisation de `Short Read Connector` et notre collaboration avec l'équipe GenScale de Rennes nous a permis de mettre en place un protocole qui est valorisé dans un article soumis et plusieurs communications (voir cf. chapitre 4, article Meng et al *in prep.*, p.141). Notre protocole facilite l'étude de jeux de données massifs issus de systèmes composés d'organismes non-modèles* et ouvre donc des perspectives d'utilisation dans le cadre d'études plus large de méta-transcriptomes environnementaux (*e.g.* sols, microbiome intestinal), dans la mesure où les ressources informatiques et les bases de données de références adéquates seraient disponibles. En effet, la difficulté à identifier les multiples éléments d'un méta-génome ou méta-transcriptome, dont certains sont phylogénétiquement proches reste un défi actuel majeur [Tu, Z. He et Zhou 2014]. La capacité de `SRC_c` à comparer, dans un temps raisonnable et avec des ressources informatiques relativement limitées, plusieurs banques dont le niveau de spécificité taxonomique est un paramètre défini par l'utilisateur, pourrait permettre d'assigner les séquences de manière précise au sein de ces jeux de données complexes.

Des perspectives pour l'amélioration du protocole d'analyse `SRC_c` incluent plusieurs points soulevés dans l'article « *A de novo approach to disentangle/decouple partner identity and function in holobiont systems* » (article Meng et al *in prep.*, p.141). Par exemple, l'évaluation de l'impact du seuil de similarité donné en paramètre à `SRC_c` pourrait permettre de diminuer ce seuil, et ainsi potentiellement d'augmenter le nombre de séquences assignées aux 3 catégories : hôte, symbiontes, et core, et donc de diminuer la proportion de séquences non-assignées. Pour cela, il s'agirait de tester plusieurs valeurs de seuil et d'en observer l'impact sur les quantités de lectures assignées aux différentes catégories. Nous suggérons également l'implémentation d'une fonction additionnelle dans l'algorithme de `SRC_c` afin d'enrichir

itérativement les banques « hôte » et « symbiontes » avec les séquences assignées à l'expérience précédente. Suite à une première analyse, les séquences de l'holobionte assignées en sortie à l'hôte et aux symbiontes, seraient ajoutées aux banques respectives, et une analyse `SRC_c` serait relancée. Cette boucle d'actions itératives serait répétée jusqu'à ce que l'enrichissement n'est plus d'impact sur l'assignation des séquences de l'holobionte. Il serait également intéressant de réaliser un suivi des assignations des lectures ayant permis l'assemblage des contigs* des différentes catégories « hôte » et « symbiontes ». Il s'agirait d'aligner les lectures assignées par `SRC_c` sur les séquences de contigs* assemblées afin de vérifier l'homogénéité taxonomique des lectures sur chacun des contigs*. Par exemple, les potentiels contigs chimériques* « hôte » assemblés à partir de lectures appartenant à deux ou plusieurs espèces présentes dans une banque de référence « hôte » pourraient êtres détectées.

La comparaison de transcriptomes, d'abord dans le cadre de l'étude de la diversité fonctionnelle des dinoflagellés « *Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome based sequence similarity network* » (article Meng et al *submitted*, p.55), puis dans celle de l'étude des transcriptomes d'hôte et symbiontes isolés à partir d'un transcriptome d'holobionte « *Key functions involved in the establishment and the maintenance of marine plankton symbiosis revealed by a meta-transcriptome approach* » (article Meng et al *in prep.*, p.185), ont été réalisées en utilisant des réseaux de similarité de séquences (SSN*). L'approche SSN* a permis non seulement d'identifier les Composantes Connexes (CCs, *i.e* des domaines fonctionnels conservés chez plusieurs transcriptomes analysés) d'intérêts pour des fonctions liées à la toxicité et la symbiose chez les dinoflagellés mais a également permis d'identifier des CCs uniquement composées de séquences sans annotation fonctionnelle. Ces CCs pourraient correspondre à des gènes encore non identifiés chez les dinoflagellés. De fait, les SSN* constituent une stratégie d'analyse adaptée à l'exploration de la « matière noire biologique ». La matière noire biologique désigne l'ensemble des séquences (au sein des génomes, méta-génomes, transcriptomes et méta-transcriptomes) qui demeurent encore non identifiée (taxonomiquement et/ou fonctionnellement) [Lopez, Halary et Bapteste 2015 ; Marcy et al.

2007 ; Rinke et al. 2013]. Dans les grands projets de séquençage, la matière noire génomique constitue une part conséquente des jeux de données, qui sont par conséquent sous-exploités [Afshinnekoo et al. 2015 ; Hug et al. 2016 ; Lopez, Halary et Bapteste 2015 ; Sunagawa et al. 2015, Carradec et *al. in press*].

L'approche SSN* telle qu'utilisée dans « *Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome based sequence similarity network* » (article Meng et al *submitted*, p.55) peut néanmoins être encore optimisée. Dans un premier temps, il serait intéressant de tester de nouvelles stratégies pour le découpage du SSN*. Dans notre analyse, différents seuils d'identité de séquences ont été testés pour finalement sélectionner celui maximisant l'obtention de CCs de grandes tailles et dont l'homogénéité fonctionnelle des séquences étaient conservées. J'ai choisi cette stratégie car elle se base principalement sur un critère de pertinence biologique des CCs qui me semblait essentiel pour les analyses envisagées. Néanmoins, le choix d'un seuil unique pour l'ensemble des alignements reste subjectif et ne peut être validé pour l'ensemble des CCs d'un SSN*. Ainsi dans le cadre de mes travaux, où je me focalise principalement sur les informations de traits fonctionnels liées aux séquences, il serait intéressant d'utiliser des algorithmes de recherche de motifs particuliers afin de détecter de façon automatique des patrons récurrents tout en testant différents seuils possibles. Par exemple, une recherche de motif peut correspondre à la détection des CCs dont les séquences sont toutes connectées les unes aux autres (*i.e* recherche de cliques*). Autre exemple, un motif peut correspondre à une chaîne : *i.e.* une CC dans laquelle chaque séquence n'est alignée qu'à deux autres séquences. Les motifs structuraux au sein de SSN* peuvent également être utilisés pour la détection de « gènes composites* » dans les génomes [Jachiet et al. 2013 ; Méheust et al. 2016, Pathmanathan et al *in press*]. La recherche de tels motifs génomiques, apparaissant souvent dans le cadre d'associations symbiotiques [Corel et al. 2016], serait donc particulièrement intéressante dans nos jeux de données d'holobiontes.

5.2 Vers la caractérisation fonctionnelle des symbioses planctoniques à partir d'approches génomiques

L'utilisation de notre chaîne d'analyses pour l'assemblage *de novo* de transcriptomes a permis de générer un jeu de données inédit et ainsi d'aborder le second volet de ma thèse qui a porté sur l'étude des dinoflagellés. Les dinoflagellés sont un des groupes proportionnellement les plus représentés dans les océans selon l'étude des lectures générées à partir des données eucaryotes échantillonnées pour 46 stations de l'expédition *Tara* Océans [Le Bescot et al. 2016 ; Vargas, Audic et al. 2015]. Ils possèdent une grande diversité de traits fonctionnels, par exemple la capacité qu'ont certaines espèces de dinoflagellés à produire des molécules toxiques leur attribue un rôle sanitaire et économique important [Murray et al. 2016]. Leur implication récurrente en tant que photosymbiontes avec une large diversité d'organismes marins, et en particuliers les radiolaires, souligne également l'importance de ces protistes au sein de la structure des écosystèmes marins. Grâce aux données générées par le projet de la MMETSP ainsi qu'aux cultures de la RCC, j'ai pu réaliser l'analyse intégrée de transcriptomes pour 47 espèces différentes de dinoflagellés. L'ensemble des échantillons séquencés provient de souches en culture. Par conséquent, dans le cadre de mon travail de thèse, il est important de souligner que l'expression de ces gènes ne correspond pas à des conditions dans l'état de symbiose. Ainsi, l'observation des fonctions dont l'expression est spécifique aux conditions de symbiose chez les dinoflagellés n'est pas possible au travers de mon jeu de données actuel. En revanche, les fonctions portées par les gènes dont l'expression est constitutive et potentiellement impliquées dans les mécanismes régissant la symbiose chez ces organismes (*i.e.* présent uniquement chez les espèces symbiotiques) ont pu être observées. On peut aujourd'hui raisonnablement penser que dans un futur relativement proche, l'évolution des protocoles de biologie moléculaire vers l'utilisation de quantité d'ARN toujours plus bas et les progrès des méthodes d'échantillonnages et de culture permettront l'isolement de cellules symbiotiques directement à partir

d'échantillons d'holobiontes, suivi de leur séquençage à partir d'une cellule unique. Les fonctions différentiellement exprimées en condition de symbiose pourraient alors êtres identifiées précisément avec les méthodes d'analyses proposée dans mon travail de thèse. Par ailleurs, les jeux de données utilisés pour cette étude ne représentent pas l'entière diversité phylogénétique des dinoflagellés (11 ordres représentés sur 21 reconnus, cf. chapitre 3). Ainsi, il n'est pas possible d'affirmer que les composantes connexes (et donc familles de protéines) faisant ici l'objet d'hypothèses pour expliquer la symbiose chez les dinoflagellés ont été toutes identifiées. De nouvelles campagnes d'échantillonnage et de séquençage permettront donc de venir compléter le jeux de données utilisés et les hypothèses formulées dans le cadre de cette thèse.

Les fonctions et familles de protéines (CCs) détectées dans mon étude représentent des marqueurs potentiellement liés à différents traits fonctionnels d'intérêts (*e.g.* à la symbiose, à la toxicité). Il sera important de valider ces marqueurs au travers d'expériences *in vitro* : par exemple par des PCR* simples voire quantitatives pour mettre en évidence la présence de ces marqueurs au sein d'échantillons impliquant des dinoflagellés symbiotiques ou d'échantillons de dinoflagellés toxiques. A ce propos, les marqueurs de toxicité que j'ai pu identifié représentent un intérêt écologique et économique important et seront testés par notre collaborateur Raffaele Siano (Ifremer, Brest) au cours de l'année 2018. D'autre part, les données de métagénomique* et méta-transcriptomique* déposés dans les bases de données publiques (ou dans le cadre de collaborations), offrent l'opportunité d'identifier et de localiser ces marqueurs dans l'environnement. Ainsi, il est désormais possible d'étudier la corrélation de la présence des domaines protéiques liés aux dinoflagellés symbiotiques aux différents échantillons et donc aux paramètres environnementaux associés. Des premiers tests ont été effectués sur les données de l'expédition *Tara* Océans dans le cadre du stage de Master 1 (Biologie Informatique de l'Université d'Orsay) de Ophélie Da Silva pendant 2 mois de juin à juillet 2017. Les premiers résultats sont prometteurs et des analyses plus approfondies permettront d'accroître nos connaissances sur les conditions environnementales de mise en place et de maintien des associations symbiotiques impliquant les dinoflagellés.

L'étude des holobiontes radiolaire-dinoflagellés représente le dernier volet de cette

thèse. L'ensemble des outils et des approches établis au cours des deux premiers volets de ma thèse ont été et seront employés pour cette étude. Les premiers résultats ont permis l'assemblage de transcriptomes distincts pour les deux partenaires, représentant ainsi des jeux de données inédits pour l'étude de ces organismes dans un contexte de symbiose. L'étude de ces transcriptomes au travers des réseaux de similarité de séquences pourra, dans les mois à venir, nous permettre d'identifier précisément des fonctions clefs liées aux processus de symbiose entre radiolaires et dinoflagellés. À l'instar de l'étude de la diversité fonctionnelle des dinoflagellés (cf. chapitre 3), des séquences protéiques cibles seront listées et il sera nécessaire de les valider aux travers d'expériences *in vitro*. Enfin, les résultats que j'ai obtenus au cours de ce travail pourront contribuer, par exemple, à étudier les bases génomiques des mécanismes de communication entre organismes (eucaryotes unicellulaires) de lignées non apparentées. Ces travaux pourront également servir de base à des études plus fines sur les impacts écologiques, biogéochimiques, voire évolutifs, de ces phénomènes symbiotiques ainsi que sur les processus biologiques impliqués.

Bibliographie

Abernathy, Jason et Ken Overturf (2016). « Comparison of Ribosomal RNA Removal Methods for Transcriptome Sequencing Workflows in Teleost Fish ». In : *Animal Biotechnology* 27.1, p. 60–65. ISSN : 1049-5398. DOI : 10.1080/10495398.2015.1086365. URL : http://dx.doi.org/10.1080/10495398.2015.1086365 (cf. p. 18).

Afshinnekoo, Ebrahim et al. (2015). « Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics ». In : *Cell Systems* 1.1, p. 72–87. ISSN : 2405-4712. DOI : 10.1016/j.cels.2015.01.001. URL : http://www.cell.com/cell-systems/abstract/S2405-4712(15)00002-2 (visité le 31/10/2017) (cf. p. 211).

Aguiar-Pulido, Vanessa et al. (2016). « Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis ». In : *Evolutionary Bioinformatics Online* 12 (Suppl 1), p. 5–16. ISSN : 1176-9343. DOI : 10.4137/EBO.S36436. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4869604/ (cf. p. 15, 18).

Alamancos, Gael P., Eneritz Agirre et Eduardo Eyras (2013). « Methods to study splicing from high-throughput RNA Sequencing data ». In : *arXiv :1304.5952 [q-bio]*. arXiv : 1304.5952. URL : http://arxiv.org/abs/1304.5952 (visité le 24/06/2014) (cf. p. 15).

Altschul, Stephen F. et al. (1990). « Basic local alignment search tool ». In : *Journal of Molecular Biology* 215.3, p. 403–410. ISSN : 0022-2836. DOI : 10.1016/S0022-2836(05)80360-2. URL : http://www.sciencedirect.com/science/article/pii/S0022283605803602 (cf. p. 136).

Alvarez-Ponce, David et al. (2013). « Gene similarity networks provide tools for understanding eukaryote origins and evolution ». In : *Proceedings of the National Academy of Sciences of the United States of America* 110.17, E1594–1603. ISSN : 1091-6490. DOI : 10.1073/pnas.1211371110 (cf. p. 22).

Anantharaman, Vivek, Lakshminarayan M. Iyer et L. Aravind (2007). « Comparative Genomics of Protists : New Insights into the Evolution of Eukaryotic Signal Transduction and Gene Regulation ». In : *Annual Review of Microbiology* 61.1, p. 453–475. DOI : 10.1146/annurev.micro.61.080706.093309. URL : https://doi.org/10.1146/annurev.micro.61.080706.093309 (cf. p. 12).

Anderson, D. M., S. W. Chisholm et C. J. Watras (1983). « Importance of life cycle events in the population dynamics of Gonyaulax tamarensis ». In : *Marine Biology* 76.2, p. 179–189.

ISSN : 0025-3162, 1432-1793. DOI : 10.1007/BF00392734. URL : https://link.springer.com/article/10.1007/BF00392734 (visité le 01/10/2017) (cf. p. 30).

ANDERSON, O. Roger (2012). *Radiolaria*. Google-Books-ID : kSnUBwAAQBAJ. Springer Science & Business Media. 363 p. ISBN : 978-1-4612-5536-9 (cf. p. 30).

ARANDA, M. et al. (2016). « Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle ». In : *Scientific Reports* 6. ISSN : 2045-2322. DOI : 10.1038/srep39734. URL : http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5177918/ (visité le 30/01/2017) (cf. p. 39).

ARCHIBALD, John M. (2015). « Endosymbiosis and Eukaryotic Cell Evolution ». In : *Current Biology* 25.19, R911–R921. ISSN : 0960-9822. DOI : 10.1016/j.cub.2015.07.055. URL : http://www.cell.com/current-biology/abstract/S0960-9822(15)00889-1 (visité le 30/05/2017) (cf. p. 27).

ATKINSON, Holly J. et al. (2009). « Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies ». In : *PLoS ONE* 4.2. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0004345. URL : http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2631154/ (visité le 20/07/2016) (cf. p. 22).

BACHVAROFF, Tsvetan R. et al. (2014). « Dinoflagellate phylogeny revisited : Using ribosomal proteins to resolve deep branching dinoflagellate clades ». In : *Molecular Phylogenetics and Evolution* 70, p. 314–322. ISSN : 1055-7903. DOI : 10.1016/j.ympev.2013.10.007. URL : http://www.sciencedirect.com/science/article/pii/S105579031300393X (visité le 22/04/2016) (cf. p. 34).

BAKER, Henry (1753). *Employment for the microscope. In two parts. Likewise a description of the microscope used in these experiments ...* London, Printed for R. Dodsley, 512 p. URL : https://www.biodiversitylibrary.org/bibliography/51442 (cf. p. 34).

BALZANO, Sergio et al. (2015). « Transcriptome analyses to investigate symbiotic relationships between marine protists ». In : *Microbial Physiology and Metabolism* 6, p. 98. DOI : 10.3389/fmicb.2015.00098. URL : http://journal.frontiersin.org/article/10.3389/fmicb.2015.00098/full (visité le 13/09/2016) (cf. p. 34, 134, 183, 200, 201).

BENOISTON, Anne-Sophie et al. (2017). « The evolution of diatoms and their biogeochemical functions ». In : *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 372.1728. ISSN : 1471-2970. DOI : 10.1098/rstb.2016.0397 (cf. p. 5).

BERGLUND, Johnny et al. (2007). « Efficiency of a phytoplankton-based and a bacterial-based food web in a pelagic marine system ». In : *Limnology and Oceanography* 52.1, p. 121–131. ISSN : 1939-5590. DOI : 10.4319/lo.2007.52.1.0121. URL : http://onlinelibrary.wiley.com/doi/10.4319/lo.2007.52.1.0121/abstract (cf. p. 4).

BIARD, Tristan, Estelle BIGEARD et al. (2017). « Biogeography and diversity of Collodaria (Radiolaria) in the global ocean ». In : *The ISME Journal* 11.6, p. 1331–1344. ISSN : 1751-7362.

DOI : 10.1038/ismej.2017.12. URL : http://www.nature.com.insb.bib.cnrs.fr/ismej/journal/v11/n6/full/ismej201712a.html?foxtrotcallback=true (visité le 15/10/2017) (cf. p. 30).

Biard, Tristan, Lars Stemmann et al. (2016). « In situ imaging reveals the biomass of giant protists in the global ocean ». In : *Nature* advance online publication. ISSN : 0028-0836. DOI : 10.1038/nature17652. URL : http://www.nature.com/nature/journal/vaop/ncurrent/full/nature17652.html (visité le 22/04/2016) (cf. p. 5, 29, 33).

Biard, Tristand (2016). « Diversité, biogéographie et écologie des Collodaires (Radiolaires) dans l'océan mondial » (cf. p. 32).

Bittner, Lucie et al. (2013). « Diversity patterns of uncultured Haptophytes unravelled by pyrosequencing in Naples Bay ». In : *Molecular Ecology* 22.1, p. 87–101. ISSN : 1365-294X. DOI : 10.1111/mec.12108. URL : http://onlinelibrary.wiley.com/doi/10.1111/mec.12108/abstract (cf. p. 5, 6).

Bogitsh, Burton J., Clint E. Carter et Thomas N. Oeltmann (2013). « Chapter 1 - Symbiosis and Parasitism ». In : *Human Parasitology (Fourth Edition)*. DOI : 10.1016/B978-0-12-415915-0.00001-7. Boston : Academic Press, p. 1–13. ISBN : 978-0-12-415915-0. URL : http://www.sciencedirect.com/science/article/pii/B9780124159150000017 (cf. p. 26).

Bolger, Anthony M., Marc Lohse et Bjoern Usadel (2014). « Trimmomatic : A flexible trimmer for Illumina Sequence Data ». In : *Bioinformatics*, btu170. ISSN : 1367-4803, 1460-2059. DOI : 10.1093/bioinformatics/btu170. URL : http://bioinformatics.oxfordjournals.org/content/early/2014/04/01/bioinformatics.btu170 (visité le 24/06/2014) (cf. p. 44).

Bone, Q. et L. Noble (2016). « Christian Andreas Viktor Hensen and his studies of plankton ». In : *Archives of Natural History* 43.1, p. 109–118. ISSN : 0260-9541. DOI : 10.3366/anh.2016.0350. URL : http://www.euppublishing.com/doi/abs/10.3366/anh.2016.0350 (cf. p. 2).

Bordenstein, Seth R. et Kevin R. Theis (2015). « Host Biology in Light of the Microbiome : Ten Principles of Holobionts and Hologenomes ». In : *PLoS Biology* 13.8. ISSN : 1544-9173. DOI : 10.1371/journal.pbio.1002226. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4540581/ (cf. p. 16).

Bork, P. et al. (2015). « Tara Oceans studies plankton at planetary scale ». In : *Science* 348.6237, p. 873–873. ISSN : 0036-8075, 1095-9203. DOI : 10.1126/science.aac5605. URL : http://science.sciencemag.org/content/348/6237/873 (visité le 20/09/2017) (cf. p. 11).

Botebol, Hugo et al. (2017). « Acclimation of a low iron adapted Ostreococcus strain to iron limitation through cell biomass lowering ». In : *Scientific Reports* 7.1, p. 327. ISSN : 2045-2322. DOI : 10.1038/s41598-017-00216-6. URL : http://www.nature.com/articles/s41598-017-00216-6 (visité le 06/04/2017) (cf. p. 48).

Brandt, Karl (1881). *Monatsberichte der Koniglish preussischen Akademie der Wissenschaften zu Berlin*. T. 6 (cf. p. 29).

Bruijn, F.A de (1946). *A combinatorial problem*. URL : http://repository.tue.nl/415282b7-6c10-4b9f-9624-4437629cc621 (visité le 04/10/2017) (cf. p. 19).

Buesseler, Ken O. et Philip W. Boyd (2009). « Shedding light on processes that control particle export and flux attenuation in the twilight zone of the open ocean ». In : *Limnology and Oceanography* 54.4, p. 1210–1232. ISSN : 1939-5590. DOI : 10.4319/lo.2009.54.4.1210. URL : http://onlinelibrary.wiley.com/doi/10.4319/lo.2009.54.4.1210/abstract (cf. p. 5).

Burki, Fabien, Maia Kaplan et al. (2016). « Untangling the early diversification of eukaryotes : a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista ». In : *Proc. R. Soc. B* 283.1823, p. 20152802. ISSN : 0962-8452, 1471-2954. DOI : 10.1098/rspb.2015.2802. URL : http://rspb.royalsocietypublishing.org/content/283/1823/20152802 (visité le 28/10/2017) (cf. p. 6).

Burki, Fabien et Patrick J. Keeling (2014). « Rhizaria ». In : *Current Biology* 24.3, R103–R107. ISSN : 0960-9822. DOI : 10.1016/j.cub.2013.12.025. URL : https://www.sciencedirect.com/science/article/pii/S0960982213015844 (visité le 06/02/2017) (cf. p. 30, 34).

Burki, Fabien, Alexander Kudryavtsev et al. (2010). « Evolution of Rhizaria : new insights from phylogenomic analysis of uncultivated protists ». In : *BMC Evolutionary Biology* 10, p. 377. ISSN : 1471-2148. DOI : 10.1186/1471-2148-10-377. URL : http://dx.doi.org/10.1186/1471-2148-10-377 (visité le 26/01/2017) (cf. p. 6, 30).

Carlos, Alvin A. et al. (1999). « Phylogenetic Position of Symbiodinium (dinophyceae) Isolates from Tridacnids (bivalvia), Cardiids (bivalvia), a Sponge (porifera), a Soft Coral (anthozoa), and a Free-Living Strain ». In : *Journal of Phycology* 35.5, p. 1054–1062. ISSN : 1529-8817. DOI : 10.1046/j.1529-8817.1999.3551054.x. URL : http://onlinelibrary.wiley.com.insb.bib.cnrs.fr/doi/10.1046/j.1529-8817.1999.3551054.x/abstract (cf. p. 37).

Caron, David A. (2016). « Ocean science : The rise of Rhizaria ». In : *Nature* 532.7600, p. 444–445. ISSN : 0028-0836. DOI : 10.1038/nature17892. URL : http://www.nature.com/nature/journal/v532/n7600/full/nature17892.html (visité le 30/10/2017) (cf. p. 200).

Caron, David A., Harriet Alexander et al. (2016). « Probing the evolution, ecology and physiology of marine protists using transcriptomics ». In : *Nature Reviews Microbiology* advance online publication. ISSN : 1740-1526. DOI : 10.1038/nrmicro.2016.160. URL : http://www.nature.com/nrmicro/journal/vaop/ncurrent/full/nrmicro.2016.160.html (visité le 21/11/2016) (cf. p. 13).

Caron, David A., Alexandra Z. Worden et al. (2008). « Protists are microbes too : a perspective ». In : *The ISME Journal* 3.1, p. 4–12. ISSN : 1751-7362. DOI : 10.1038/ismej.2008.101. URL : https://www.nature.com/ismej/journal/v3/n1/full/ismej2008101a.html (visité le 03/10/2017) (cf. p. 12).

Chang, Zheng et al. (2015). « Bridger : a new framework for de novo transcriptome assembly using RNA-seq data ». In : *Genome Biology* 16, p. 30. ISSN : 1465-6906. DOI : 10.1186/s13059-015-0596-2. URL : https://doi.org/10.1186/s13059-015-0596-2 (cf. p. 19).

Chebreux, Benoist (2017). *Genome Sequence Assembly Using Trace Signals and Additional Sequence Information*. URL : http://www.bioinfo.de/isb/gcb99/talks/chevreux/ (visité le 18/10/2017) (cf. p. 19).

Chistoserdova, Ludmila (2009). « Functional Metagenomics : Recent Advances and Future Challenges ». In : *Biotechnology and Genetic Engineering Reviews* 26.1, p. 335–352. ISSN : 0264-8725. DOI : 10.5661/bger-26-335. URL : http://dx.doi.org/10.5661/bger-26-335 (cf. p. 15).

Consortium, The 1000 Genomes Project (2010). « A map of human genome variation from population-scale sequencing ». In : *Nature* 467.7319, p. 1061–1073. ISSN : 0028-0836. DOI : 10.1038/nature09534. URL : http://www.nature.com/nature/journal/v467/n7319/full/nature09534.html (visité le 28/10/2017) (cf. p. 10).

Corel, Eduardo et al. (2016). « Network-Thinking : Graphs to Analyze Microbial Complexity and Evolution ». In : *Trends in Microbiology* 24.3, p. 224–237. ISSN : 0966-842X. DOI : 10.1016/j.tim.2015.12.003. URL : http://www.sciencedirect.com/science/article/pii/S0966842X15002796 (visité le 13/07/2016) (cf. p. 22, 211).

Davy, Simon K., Denis Allemand et Virginia M. Weis (2012). « Cell Biology of Cnidarian-Dinoflagellate Symbiosis ». In : *Microbiology and Molecular Biology Reviews : MMBR* 76.2, p. 229–261. ISSN : 1092-2172. DOI : 10.1128/MMBR.05014-11. URL : http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3372257/ (visité le 05/10/2016) (cf. p. 28).

Decelle, Johan, SÃ©bastien Colin et Rachel A. Foster (2015). « Photosymbiosis in Marine Planktonic Protists ». In : *Marine Protists*. DOI : 10.1007/978-4-431-55130-0_19. Springer, Tokyo, p. 465–500. ISBN : 978-4-431-55129-4 978-4-431-55130-0. URL : https://link.springer.com/chapter/10.1007/978-4-431-55130-0_19 (visité le 29/10/2017) (cf. p. 27, 29, 40).

Decelle, Johan, Ian Probert et al. (2012). « An original mode of symbiosis in open ocean plankton ». In : *Proceedings of the National Academy of Sciences* 109.44, p. 18000–18005. ISSN : 0027-8424, 1091-6490. DOI : 10.1073/pnas.1212303109. URL : http://www.pnas.org/content/109/44/18000 (visité le 01/08/2014) (cf. p. 29, 134).

Decelle, Johan, Noritoshi Suzuki et al. (2012). « Molecular Phylogeny and Morphological Evolution of the Acantharia (Radiolaria) ». In : *Protist* 163.3, p. 435–450. ISSN : 1434-4610. DOI : 10.1016/j.protis.2011.10.002. URL : http://www.sciencedirect.com/science/article/pii/S1434461011000988 (cf. p. 30, 134).

Falkowski, Paul (2012). « Ocean Science : The power of plankton ». In : *Nature* 483.7387, S17–S20. ISSN : 0028-0836. DOI : 10.1038/483S17a. URL : http://www.nature.com/nature/journal/v483/n7387_supp/full/483S17a.html (visité le 28/10/2017) (cf. p. 5).

Faust, Karoline et Jeroen Raes (2012). « Microbial interactions : from networks to models ». In : *Nature Reviews Microbiology* 10.8, p. 538–550. ISSN : 1740-1526. DOI : 10.1038/nrmicro2832. URL : http://www.nature.com/nrmicro/journal/v10/n8/full/nrmicro2832.html (visité le 01/10/2017) (cf. p. 27).

Field, Christopher B. et al. (1998). « Primary Production of the Biosphere : Integrating Terrestrial and Oceanic Components ». In : *Science* 281.5374, p. 237–240. ISSN : 0036-8075, 1095-9203. DOI : 10.1126/science.281.5374.237. URL : http://science.sciencemag.org.insb.bib.cnrs.fr/content/281/5374/237 (visité le 11/10/2017) (cf. p. 4).

Fiore, Cara L. et al. (2015). « Transcriptional activity of the giant barrel sponge, Xestospongia muta Holobiont : molecular evidence for metabolic interchange ». In : *Frontiers in Microbiology* 6. ISSN : 1664-302X. DOI : 10.3389/fmicb.2015.00364. URL : https://www.frontiersin.org/articles/10.3389/fmicb.2015.00364/full (visité le 16/10/2017) (cf. p. 140).

Flynn, Kevin J. et al. (2013). « Misuse of the phytoplankton–zooplankton dichotomy : the need to assign organisms as mixotrophs within plankton functional types ». In : *Journal of Plankton Research* 35.1, p. 3–11. ISSN : 0142-7873. DOI : 10.1093/plankt/fbs062. URL : https://academic.oup.com/plankt/article/35/1/3/1515294 (visité le 28/10/2017) (cf. p. 4).

Forster, Dominik et al. (2015). « Testing ecological theories with sequence similarity networks : marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms ». In : *BMC Biology* 13, p. 16. ISSN : 1741-7007. DOI : 10.1186/s12915-015-0125-5. URL : https://doi.org/10.1186/s12915-015-0125-5 (cf. p. 5, 6).

Gallego Romero, Irene et al. (2014). « RNA-seq : impact of RNA degradation on transcript quantification ». In : *BMC Biology* 12, p. 42. ISSN : 1741-7007. DOI : 10.1186/1741-7007-12-42. URL : https://doi.org/10.1186/1741-7007-12-42 (cf. p. 18).

Gallo, Juan Esteban et al. (2014). « The complex task of choosing a de novo assembly : Lessons from fungal genomes ». In : *Computational Biology and Chemistry*. Complexity in Genomes 53 (Part A), p. 97–107. ISSN : 1476-9271. DOI : 10.1016/j.compbiolchem.2014.08.014. URL : http://www.sciencedirect.com/science/article/pii/S1476927114000930 (cf. p. 21, 240).

Gasol, Josep M., Paul A. del Giorgio et Carlos M. Duarte (1997). « Biomass distribution in marine planktonic communities ». In : *Limnology and Oceanography* 42.6, p. 1353–1363. ISSN : 1939-5590. DOI : 10.4319/lo.1997.42.6.1353. URL : http://onlinelibrary.wiley.com/doi/10.4319/lo.1997.42.6.1353/abstract (cf. p. 29).

Gast, Rebecca J. et David A. Caron (2001). « Photosymbiotic associations in planktonic foraminifera and radiolaria ». In : *Hydrobiologia* 461.1, p. 1–7. ISSN : 0018-8158, 1573-5117. DOI : 10.1023/A:1012710909023. URL : https://link-springer-com.insb.bib.cnrs.fr/article/10.1023/A:1012710909023 (visité le 01/10/2017) (cf. p. 33).

Gilbert, Jack A., Janet K. Jansson et Rob Knight (2014). « The Earth Microbiome project : successes and aspirations ». In : *BMC biology* 12, p. 69. issn : 1741-7007. doi : 10.1186/s12915-014-0069-1 (cf. p. 11).

Glöckner, Gernot et al. (2014). « The Genome of the Foraminiferan Reticulomyxa filosa ». In : *Current Biology* 24.1, p. 11–18. issn : 0960-9822. doi : 10.1016/j.cub.2013.11.027. url : http://www.sciencedirect.com/science/article/pii/S0960982213014462 (visité le 07/03/2017) (cf. p. 34).

Gómez, Fernando (2014). « Problematic Biases in the Availability of Molecular Markers in Protists : The Example of the Dinoflagellates ». In : *Acta Protozoologica* 2014 (Volume 53, Issue 1, Special topic issue : "Marine Heterotrophic Protists"), p. 6375. issn : 1689-0027. doi : 10.4467/16890027AP.13.0021.1118. url : http://www.ejournals.eu/Acta-Protozoologica/Tom-53(2014)/Numer-1/art/2038/ (visité le 30/09/2017) (cf. p. 35).

Grabherr, Manfred G. et al. (2011). « Full-length transcriptome assembly from RNA-Seq data without a reference genome (Trinity) ». In : *Nature Biotechnology* 29.7, p. 644–652. issn : 1087-0156. doi : 10.1038/nbt.1883. url : http://www.nature.com/nbt/journal/v29/n7/full/nbt.1883.html (visité le 24/06/2014) (cf. p. 19, 45, 46).

Guidi, Lionel et al. (2016). « Plankton networks driving carbon export in the oligotrophic ocean ». In : *Nature* 532.7600, p. 465–470. issn : 0028-0836. doi : 10.1038/nature16942. url : http://www.nature.com/nature/journal/v532/n7600/full/nature16942.html (visité le 30/11/2016) (cf. p. 4, 5, 29).

Gutierrez-Rodriguez, Andres et al. (2017). « Dimethylated sulfur compounds in symbiotic protists : A potentially significant source for marine DMS(P) ». In : *Limnology and Oceanography* 62.3, p. 1139–1154. issn : 1939-5590. doi : 10.1002/lno.10491. url : http://onlinelibrary.wiley.com/doi/10.1002/lno.10491/abstract (cf. p. 39).

Haas, Brian J. et al. (2013). « De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis ». In : *Nature Protocols* 8.8, p. 1494–1512. issn : 1754-2189. doi : 10.1038/nprot.2013.084. url : http://www.nature.com/nprot/journal/v8/n8/full/nprot.2013.084.html (visité le 24/06/2014) (cf. p. 22, 45).

Hackett, Jeremiah D. et al. (2013). « Evolution of Saxitoxin Synthesis in Cyanobacteria and Dinoflagellates ». In : *Molecular Biology and Evolution* 30.1, p. 70–78. issn : 0737-4038. doi : 10.1093/molbev/mss142. url : http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3525144/ (visité le 02/11/2016) (cf. p. 39).

Haeckel, Ernst (1904). *Kunstformen der Natur* (cf. p. 31, 36).

Hallegraeff, G. M. (1993). « A review of harmful algal blooms and their apparent global increase ». In : *Phycologia* 32.2, p. 79–99. issn : 0031-8884. doi : 10.2216/i0031-8884-32-2-79.1. url : http://www.phycologia.org/doi/abs/10.2216/i0031-8884-32-2-79.1 (cf. p. 3).

HANDELSMAN, J. et al. (1998). « Molecular biological access to the chemistry of unknown soil microbes : a new frontier for natural products ». In : *Chemistry & Biology* 5.10, R245–249. ISSN : 1074-5521 (cf. p. 16).

HANSEN, Gert et Niels DAUGBJERG (2009). « Symbiodinium Natans Sp. Nov. : A "Free-Living" Dinoflagellate from Tenerife (northeast-Atlantic Ocean)1 ». In : *Journal of Phycology* 45.1, p. 251–263. ISSN : 1529-8817. DOI : 10.1111/j.1529-8817.2008.00621.x. URL : http://onlinelibrary.wiley.com/doi/10.1111/j.1529-8817.2008.00621.x/abstract (cf. p. 36).

HE, Shaomei et al. (2010). « Validation of two ribosomal RNA removal methods for microbial metatranscriptomics ». In : *Nature Methods* 7.10, p. 807–812. ISSN : 1548-7091. DOI : 10.1038/nmeth.1507. URL : http://www.nature.com/nmeth/journal/v7/n10/full/nmeth.1507.html?foxtrotcallback=true (visité le 12/10/2017) (cf. p. 17).

HERNDL, Gerhard J. et Thomas REINTHALER (2013). « Microbial control of the dark end of the biological pump ». In : *Nature Geoscience* 6.9, p. 718–724. ISSN : 1752-0894. DOI : 10.1038/ngeo1921 (cf. p. 5).

HOU, Yubo et Senjie LIN (2009). « Distinct Gene Number-Genome Size Relationships for Eukaryotes and Non-Eukaryotes : Gene Content Estimation for Dinoflagellate Genomes ». In : *PLOS ONE* 4.9, e6978. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0006978. URL : http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0006978 (visité le 31/10/2017) (cf. p. 14).

HUG, Laura A. et al. (2016). « A new view of the tree of life ». In : *Nature Microbiology* 1.5, nmicrobiol201648. ISSN : 2058-5276. DOI : 10.1038/nmicrobiol.2016.48. URL : https://www.nature.com/articles/nmicrobiol201648 (visité le 31/10/2017) (cf. p. 211).

JACHIET, Pierre-Alain et al. (2013). « MosaicFinder : Identification of fused gene families in sequence similarity networks ». In : *Bioinformatics (Oxford, England)* 29. DOI : 10.1093/bioinformatics/btt049 (cf. p. 211).

JANOUŠKOVEC, Jan et al. (2016). « Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics ». In : *Proceedings of the National Academy of Sciences*, p. 201614842. ISSN : 0027-8424, 1091-6490. DOI : 10.1073/pnas.1614842114. URL : http://www.pnas.org/content/early/2016/12/23/1614842114 (visité le 04/01/2017) (cf. p. 6, 35, 38, 40, 52, 53).

JIANG, Yue et al. (2016). « Metatranscriptomic analysis of diverse microbial communities reveals core metabolic pathways and microbiome-specific functionality ». In : *Microbiome* 4. ISSN : 2049-2618. DOI : 10.1186/s40168-015-0146-x. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4710996/ (cf. p. 15).

JOHNSON, William S. et Dennis M. ALLEN (2012). *Zooplankton of the Atlantic and Gulf Coasts : A Guide to Their Identification and Ecology*. Google-Books-ID : xgCVYfyj5MgC. JHU Press. 471 p. ISBN : 978-1-4214-0618-3 (cf. p. 33).

Jones, Philip et al. (2014). « InterProScan 5 : genome-scale protein function classification ». In : *Bioinformatics (Oxford, England)* 30.9, p. 1236–1240. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/btu031 (cf. p. 45).

Karsenti, Eric et al. (2011). « A Holistic Approach to Marine Eco-Systems Biology ». In : *PLOS Biology* 9.10, e1001177. ISSN : 1545-7885. DOI : 10.1371/journal.pbio.1001177. URL : http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001177 (visité le 29/10/2017) (cf. p. 6, 11).

Keeling, Patrick J., Fabien Burki et al. (2014). « The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) : Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing ». In : *PLOS Biol* 12.6, e1001889. ISSN : 1545-7885. DOI : 10.1371/journal.pbio.1001889. URL : http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001889 (visité le 08/06/2016) (cf. p. 6, 7, 40, 52).

Keeling, Patrick J. et Javier del Campo (2017). « Marine Protists Are Not Just Big Bacteria ». In : *Current Biology* 27.11, R541–R549. ISSN : 0960-9822. DOI : 10.1016/j.cub.2017.03.075. URL : http://www.cell.com/current-biology/abstract/S0960-9822(17)30405-0 (visité le 28/10/2017) (cf. p. 13).

Knight, Rob et al. (2012). « Unlocking the potential of metagenomics through replicated experimental design ». In : *Nature biotechnology* 30.6, p. 513–520. ISSN : 1087-0156. DOI : 10.1038/nbt.2235. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4902277/ (cf. p. 11).

Kodama, Yuuki et al. (2014). « Comparison of gene expression of Paramecium bursaria with and without Chlorella variabilissymbionts ». In : *BMC Genomics* 15, p. 183. ISSN : 1471-2164. DOI : 10.1186/1471-2164-15-183. URL : http://dx.doi.org/10.1186/1471-2164-15-183 (visité le 03/01/2017) (cf. p. 45).

Kohli, Gurjeet S. et al. (2015). « Polyketide synthesis genes associated with toxin production in two species of Gambierdiscus (Dinophyceae) ». In : *BMC Genomics* 16, p. 410. ISSN : 1471-2164. DOI : 10.1186/s12864-015-1625-y. URL : https://doi.org/10.1186/s12864-015-1625-y (cf. p. 14, 40).

Kopylova, Evguenia, Laurent Noé et Hélène Touzet (2012). « SortMeRNA : fast and accurate filtering of ribosomal RNAs in metatranscriptomic data ». In : *Bioinformatics* 28.24, p. 3211–3217. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/bts611. URL : https://academic.oup.com/bioinformatics/article/28/24/3211/246053/SortMeRNA-fast-and-accurate-filtering-of-ribosomal (visité le 13/10/2017) (cf. p. 18, 46).

Krabberød, Anders K., Jon Bråte et al. (2011). « Radiolaria Divided into Polycystina and Spasmaria in Combined 18S and 28S rDNA Phylogeny ». In : *PLoS ONE* 6.8. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0023526. URL : http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3154480/ (cf. p. 30).

Krabberød, Anders K., Russell J. S. Orr et al. (2017). « Single Cell Transcriptomics, Mega-Phylogeny, and the Genetic Basis of Morphological Innovations in Rhizaria ». In : *Molecular Biology and Evolution* 34.7, p. 1557–1573. ISSN : 0737-4038. DOI : 10.1093/molbev/msx075. URL : https://academic.oup.com/mbe/article/34/7/1557/3058782/Single-Cell-Transcriptomics-Mega-Phylogeny-and-the (visité le 19/06/2017) (cf. p. 34, 182, 183).

LaJeunesse, T. (2002). « Diversity and community structure of symbiotic dinoflagellates from Caribbean coral reefs ». In : *Marine Biology* 141.2, p. 387–400. ISSN : 0025-3162, 1432-1793. DOI : 10.1007/s00227-002-0829-2. URL : https://link-springer-com.insb.bib.cnrs.fr/article/10.1007/s00227-002-0829-2 (visité le 30/09/2017) (cf. p. 14, 37).

LaJeunesse, Todd C. et al. (2005). « Symbiodinium (pyrrhophyta) Genome Sizes (dna Content) Are Smallest Among Dinoflagellates1 ». In : *Journal of Phycology* 41.4, p. 880–886. ISSN : 1529-8817. DOI : 10.1111/j.0022-3646.2005.04231.x. URL : http://onlinelibrary.wiley.com/doi/10.1111/j.0022-3646.2005.04231.x/abstract (cf. p. 14).

Langmead, Ben et Steven L Salzberg (2012). « Fast gapped-read alignment with Bowtie 2 ». In : *Nature methods* 9.4, p. 357–359. ISSN : 1548-7091. DOI : 10.1038/nmeth.1923. URL : http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322381/ (visité le 03/01/2017) (cf. p. 45, 136).

Le Bescot, Noan et al. (2016). « Global patterns of pelagic dinoflagellate diversity across protist size classes unveiled by metabarcoding ». In : *Environmental Microbiology* 18.2, p. 609–626. ISSN : 1462-2920. DOI : 10.1111/1462-2920.13039. URL : http://onlinelibrary.wiley.com/doi/10.1111/1462-2920.13039/abstract (cf. p. 212).

Levine, Suzanne N. et al. (1999). « The Impact of Zooplankton Grazing on Phytoplankton Species Composition and Biomass in Lake Champlain (USA-Canada) ». In : *Journal of Great Lakes Research* 25.1, p. 61–77. ISSN : 0380-1330. DOI : 10.1016/S0380-1330(99)70717-3. URL : http://www.sciencedirect.com/science/article/pii/S0380133099707173 (cf. p. 4).

Li, Bo et al. (2014). « Evaluation of de novo transcriptome assemblies from RNA-Seq data ». In : *Genome Biology* 15, p. 553. ISSN : 1474-760X. DOI : 10.1186/s13059-014-0553-5. URL : https://doi.org/10.1186/s13059-014-0553-5 (cf. p. 22).

Li, Heng et Richard Durbin (2010). « Fast and accurate long-read alignment with Burrows–Wheeler transform ». In : *Bioinformatics* 26.5, p. 589–595. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btp698. URL : https://academic.oup.com/bioinformatics/article/26/5/589/211735 (visité le 26/10/2017) (cf. p. 136).

Lin, Senjie (2011). « Genomic understanding of dinoflagellates ». In : *Research in Microbiology*. The genome organisation of eukaryotic microbes 162.6, p. 551–569. ISSN : 0923-2508. DOI : 10.1016/j.resmic.2011.04.006. URL : http://www.sciencedirect.com/science/article/pii/S0923250811000684 (visité le 24/05/2016) (cf. p. 35).

Lin, Senjie et al. (2015). « The Symbiodinium kawagutii genome illuminates dinoflagellate gene expression and coral symbiosis ». In : *Science* 350.6261, p. 691–694. ISSN : 0036-8075, 1095-9203. DOI : 10.1126/science.aad0408. URL : http://science.sciencemag.org.insb.bib.cnrs.fr/content/350/6261/691 (visité le 13/12/2016) (cf. p. 39).

Lopez, Philippe, Sébastien Halary et Eric Bapteste (2015). « Highly divergent ancient gene families in metagenomic samples are compatible with additional divisions of life ». In : *Biology Direct* 10, p. 64. ISSN : 1745-6150. DOI : 10.1186/s13062-015-0092-3. URL : https://doi.org/10.1186/s13062-015-0092-3 (cf. p. 22, 210, 211).

Luo, Ruibang et al. (2012). « SOAPdenovo2 : an empirically improved memory-efficient short-read de novo assembler ». In : *GigaScience* 1.1, p. 18. ISSN : 2047-217X. DOI : 10.1186/2047-217X-1-18. URL : http://www.gigasciencejournal.com/content/1/1/18/abstract (visité le 17/07/2014) (cf. p. 19).

Marchet, Camille et al. (2016). « A resource-frugal probabilistic dictionary and applications in (meta)genomics ». In : *arXiv :1605.08319 [cs, q-bio]*. arXiv : 1605.08319. URL : http://arxiv.org/abs/1605.08319 (cf. p. 135, 136).

Marchler-Bauer, Aron et al. (2011). « CDD : a Conserved Domain Database for the functional annotation of proteins ». In : *Nucleic Acids Research* 39 (Database issue), p. D225–229. ISSN : 1362-4962. DOI : 10.1093/nar/gkq1189 (cf. p. 23).

Marcy, Yann et al. (2007). « Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth ». In : *Proceedings of the National Academy of Sciences* 104.29, p. 11889–11894. ISSN : 0027-8424, 1091-6490. DOI : 10.1073/pnas.0704662104. URL : http://www.pnas.org/content/104/29/11889 (visité le 19/10/2017) (cf. p. 210).

Margulies, Marcel et al. (2005). « Genome sequencing in microfabricated high-density picolitre reactors ». In : *Nature* 437.7057, p. 376–380. ISSN : 1476-4687. DOI : 10.1038/nature03959 (cf. p. 200).

Martin, Jeffrey A. et Zhong Wang (2011). « Next-generation transcriptome assembly ». In : *Nature Reviews. Genetics* 12.10, p. 671–682. ISSN : 1471-0064. DOI : 10.1038/nrg3068 (cf. p. 14, 19).

Martinez, Xavier et al. (2016). « MetaTrans : an open-source pipeline for metatranscriptomics ». In : *Scientific Reports* 6, srep26447. ISSN : 2045-2322. DOI : 10.1038/srep26447. URL : https://www.nature.com/articles/srep26447 (visité le 13/10/2017) (cf. p. 208).

Martinez-Garcia, Manuel et al. (2012). « Unveiling in situ interactions between marine protists and bacteria through single cell sequencing ». In : *The ISME Journal* 6.3, p. 703–707. ISSN : 1751-7362. DOI : 10.1038/ismej.2011.126. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3280149/ (cf. p. 206).

MATASCI, Naim et al. (2014). « Data access for the 1,000 Plants (1KP) project ». In : *GigaScience* 3, p. 17. ISSN : 2047-217X. DOI : 10.1186/2047-217X-3-17 (cf. p. 11).

MATSUOKA, Atsushi (2007). « Living radiolarian feeding mechanisms : new light on past marine ecosystems ». In : *Swiss Journal of Geosciences* 100.2, p. 273–279. ISSN : 1661-8726, 1661-8734. DOI : 10.1007/s00015-007-1228-y. URL : https://link-springer-com.insb.bib.cnrs.fr/article/10.1007/s00015-007-1228-y (visité le 28/09/2017) (cf. p. 33).

MÉHEUST, Raphaël et al. (2016). « Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis ». In : *Proceedings of the National Academy of Sciences* 113.13, p. 3579–3584. ISSN : 0027-8424, 1091-6490. DOI : 10.1073/pnas.1517551113. URL : http://www.pnas.org/content/113/13/3579 (visité le 19/01/2017) (cf. p. 22, 211).

MILLER, Jason R., Sergey KOREN et Granger SUTTON (2010). « Assembly algorithms for next-generation sequencing data ». In : *Genomics* 95.6, p. 315–327. ISSN : 0888-7543. DOI : 10.1016/j.ygeno.2010.03.001. URL : http://www.sciencedirect.com/science/article/pii/S0888754310000492 (cf. p. 19).

MITRA, Aditee et al. (2016). « Defining Planktonic Protist Functional Groups on Mechanisms for Energy and Nutrient Acquisition : Incorporation of Diverse Mixotrophic Strategies ». In : *Protist* 167.2, p. 106–120. ISSN : 1434-4610. DOI : 10.1016/j.protis.2016.01.003. URL : http://www.sciencedirect.com/science/article/pii/S1434461016000043 (cf. p. 37).

MOHSEN, Hussein, Haixu TANG et Yuzhen YE (2017). « Improving de novo metatranscriptome assembly via machine learning algorithms ». In : *International Journal of Computational Biology and Drug Design* 10.2, p. 91–107. ISSN : 1756-0756. DOI : 10.1504/IJCBDD.2017.083877. URL : http://www.inderscienceonline.com/doi/abs/10.1504/IJCBDD.2017.083877 (cf. p. 209).

MUKHERJEE, Supratim et al. (2017). « Genomes OnLine Database (GOLD) v.6 : data updates and feature enhancements ». In : *Nucleic Acids Research* 45 (Database issue), p. D446–D456. ISSN : 0305-1048. DOI : 10.1093/nar/gkw992. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210664/ (cf. p. 12).

MURRAY, Shauna A. et al. (2016). « Unravelling the functional genetics of dinoflagellates : a review of approaches and opportunities ». In : *Perspectives in Phycology*, p. 37–52. ISSN : , DOI : 10.1127/pip/2016/0039. URL : https://www.schweizerbart.de/papers/pip/detail/3/85474/Unravelling_the_functional_genetics_of_dinoflagellates_a_review_of_approaches_and_opportunities (visité le 18/10/2017) (cf. p. 14, 36, 39, 212).

NAGARAJAN, Niranjan et Mihai POP (2013). « Sequence assembly demystified ». In : *Nature Reviews Genetics* 14.3, p. 157–167. ISSN : 1471-0056. DOI : 10.1038/nrg3367. URL : http://www.nature.com/nrg/journal/v14/n3/full/nrg3367.html (visité le 24/06/2014) (cf. p. 14, 19).

OGANE, Kaoru et al. (2010). « Direct observation of the skeletal growth patterns of polycystine radiolarians using a fluorescent marker ». In : *Marine Micropaleontology* 77.3, p. 137–144. ISSN : 0377-8398. DOI : 10.1016/j.marmicro.2010.08.005. URL : http://www.sciencedirect.com/science/article/pii/S037783981000085X (cf. p. 30).

OSHLACK, Alicia, Mark D. ROBINSON et Matthew D. YOUNG (2010). « From RNA-seq reads to differential expression results ». In : *Genome Biology* 11.12, p. 220. ISSN : 1465-6906. DOI : 10.1186/gb-2010-11-12-220. URL : http://genomebiology.com/2010/11/12/220/abstract (visité le 17/07/2014) (cf. p. 15).

PANDE, Kalyan, Changbin CHEN et Suzanne M. NOBLE (2013). « Passage through the mammalian gut triggers a phenotypic switch that promotes *Candida albicans* commensalism ». In : *Nature Genetics* 45.9, ng.2710. ISSN : 1546-1718. DOI : 10.1038/ng.2710. URL : https://www.nature.com/articles/ng.2710 (visité le 30/10/2017) (cf. p. 26).

PASCAULT, Noémie et al. (2015). « Technical challenges in metatranscriptomic studies applied to the bacterial communities of freshwater ecosystems ». In : *Genetica* 143.2, p. 157–167. ISSN : 0016-6707, 1573-6857. DOI : 10.1007/s10709-014-9783-4. URL : https://link-springer-com.insb.bib.cnrs.fr/article/10.1007/s10709-014-9783-4 (visité le 12/10/2017) (cf. p. 17).

PEARSON, William R. (2013). « An Introduction to Sequence Similarity ("Homology") Searching ». In : *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* 0 3. ISSN : 1934-3396. DOI : 10.1002/0471250953.bi0301s42. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3820096/ (cf. p. 23).

PERRU, Olivier (2006). « Aux origines des recherches sur la symbiose vers 1868-1883 | The origins of research on symbiosis (1868-1883) ». In : *Revue d'histoire des sciences* 59.1, p. 5–27. DOI : 10.3406/rhs.2006.2264. URL : http://www.persee.fr/doc/rhs_0151-4105_2006_num_59_1_2264 (visité le 22/09/2017) (cf. p. 26).

PESANT, StÃ©phane et al. (2015). « Open science resources for the discovery and analysis of *Tara* Oceans data ». In : *Scientific Data* 2, sdata201523. ISSN : 2052-4463. DOI : 10.1038/sdata.2015.23. URL : https://www.nature.com/articles/sdata201523 (visité le 29/10/2017) (cf. p. 6).

PETERSEN, Jillian M. et al. (2011). « Hydrogen is an energy source for hydrothermal vent symbioses ». In : *Nature* 476.7359, p. 176–180. ISSN : 0028-0836. DOI : 10.1038/nature10325. URL : http://www.nature.com/nature/journal/v476/n7359/full/nature10325.html?foxtrotcallback=true (visité le 27/09/2017) (cf. p. 28).

PETTERSSON, Erik, Joakim LUNDEBERG et Afshin AHMADIAN (2009). « Generations of sequencing technologies ». In : *Genomics* 93.2, p. 105–111. ISSN : 0888-7543. DOI : 10.1016/j.ygeno.2008.10.003. URL : http://www.sciencedirect.com/science/article/pii/S0888754308002498 (cf. p. 10).

Piganeau, Gwenael et al. (2011). « How and Why DNA Barcodes Underestimate the Diversity of Microbial Eukaryotes ». In : *PLOS ONE* 6.2, e16342. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0016342. URL : http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0016342 (visité le 29/10/2017) (cf. p. 5).

Pinzón, Jorge H. et al. (2015). « Whole transcriptome analysis reveals changes in expression of immune-related genes during and after bleaching in a reef-building coral ». In : *Royal Society Open Science* 2.4, p. 140214. ISSN : 2054-5703. DOI : 10.1098/rsos.140214. URL : http://rsos.royalsocietypublishing.org/content/2/4/140214 (visité le 16/10/2017) (cf. p. 140).

Piskol, Robert, Gokul Ramaswami et Jin Billy Li (2013). « Reliable Identification of Genomic Variants from RNA-Seq Data ». In : *The American Journal of Human Genetics* 93.4, p. 641–651. ISSN : 0002-9297. DOI : 10.1016/j.ajhg.2013.08.008. URL : http://www.sciencedirect.com/science/article/pii/S0002929713003832 (visité le 29/07/2014) (cf. p. 15).

Polet, Stephane et al. (2004). « Small-Subunit Ribosomal RNA Gene Sequences of Phaeodarea Challenge the Monophyly of Haeckel's Radiolaria ». In : *Protist* 155.1, p. 53–63. ISSN : 1434-4610. DOI : 10.1078/1434461000164. URL : http://www.sciencedirect.com/science/article/pii/S1434461004701650 (cf. p. 30).

Poretsky, Rachel S. et al. (2005). « Analysis of Microbial Gene Transcripts in Environmental Samples ». In : *Applied and Environmental Microbiology* 71.7, p. 4121–4126. ISSN : 0099-2240, 1098-5336. DOI : 10.1128/AEM.71.7.4121-4126.2005. URL : http://aem.asm.org/content/71/7/4121 (visité le 11/10/2017) (cf. p. 15).

Probert, Ian et al. (2014). « Brandtodinium gen. nov. and B. nutricula comb. Nov. (Dinophyceae), a dinoflagellate commonly found in symbiosis with polycystine radiolarians ». In : *Journal of Phycology* 50.2, p. 388–399. ISSN : 1529-8817. DOI : 10.1111/jpy.12174. URL : http://onlinelibrary.wiley.com/doi/10.1111/jpy.12174/abstract (visité le 22/10/2014) (cf. p. 8, 29, 36, 37, 40, 134).

Qin, Junjie et al. (2010). « A human gut microbial gene catalog established by metagenomic sequencing ». In : *Nature* 464.7285, p. 59–65. ISSN : 0028-0836. DOI : 10.1038/nature08821. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3779803/ (cf. p. 11).

Raven, J. A. (1998). « Extrapolating feedback processes from the present to the past ». In : *Philosophical Transactions of the Royal Society of London B : Biological Sciences* 353.1365, p. 19–28. ISSN : 0962-8436, 1471-2970. DOI : 10.1098/rstb.1998.0187. URL : http://rstb.royalsocietypublishing.org/content/353/1365/19 (visité le 29/09/2017) (cf. p. 34).

Reuter, Jason A., Damek Spacek et Michael P. Snyder (2015). « High-Throughput Sequencing Technologies ». In : *Molecular cell* 58.4, p. 586–597. ISSN : 1097-2765. DOI : 10.1016/j.molcel.2015.05.004. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4494749/ (cf. p. 10).

Richardson, Anthony J. et David S. Schoeman (2004). « Climate impact on plankton ecosystems in the Northeast Atlantic ». In : *Science (New York, N.Y.)* 305.5690, p. 1609–1612. issn : 1095-9203. doi : 10.1126/science.1100958 (cf. p. 2).

Rinke, Christian et al. (2013). « Insights into the phylogeny and coding potential of microbial dark matter ». In : doi : 10.1038/nature12352. url : http://darchive.mblwhoilibrary.org/handle/1912/6194 (visité le 19/10/2017) (cf. p. 25, 211).

Rizzo, Peter J. (2003). « Those amazing dinoflagellate chromosomes ». In : *Cell Research* 13.4, p. 215–217. issn : 1001-0602. doi : 10.1038/sj.cr.7290166. url : https://www.nature.com/cr/journal/v13/n4/full/7290166a.html (visité le 30/09/2017) (cf. p. 35).

Robasky, Kimberly, Nathan E. Lewis et George M. Church (2014). « The Role of Replicates for Error Mitigation in Next-Generation Sequencing ». In : *Nature reviews. Genetics* 15.1, p. 56–62. issn : 1471-0056. doi : 10.1038/nrg3655. url : http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103745/ (visité le 22/07/2014) (cf. p. 19).

Robertson, Gordon et al. (2010). « De novo assembly and analysis of RNA-seq data ». In : *Nature Methods* 7.11, p. 909–912. issn : 1548-7091. doi : 10.1038/nmeth.1517. url : http://www.nature.com/nmeth/journal/v7/n11/full/nmeth.1517.html (visité le 17/07/2014) (cf. p. 19).

Rosenberg, Eugene et Ilana Zilber-Rosenberg (2013). « The Hologenome Concept : Human, Animal and Plant Microbiota ». In : *The Hologenome Concept : Human, Animal and Plant Microbiota*, p. 1–178. doi : 10.1007/978-3-319-04241-1 (cf. p. 16).

Roth, Melissa S. (2014). « The engine of the reef : photobiology of the coral–algal symbiosis ». In : *Frontiers in Microbiology* 5. issn : 1664-302X. doi : 10.3389/fmicb.2014.00422. url : http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4141621/ (cf. p. 28).

Russell, James B. et David B. Wilson (1996). « Why Are Ruminal Cellulolytic Bacteria Unable to Digest Cellulose at Low pH ? » In : *Journal of Dairy Science* 79.8, p. 1503–1509. issn : 0022-0302. doi : 10.3168/jds.S0022-0302(96)76510-4. url : http://www.sciencedirect.com/science/article/pii/S0022030296765104 (cf. p. 27).

Sabourault, Cécile et al. (2009). « Comprehensive EST analysis of the symbiotic sea anemone, Anemonia viridis ». In : *BMC Genomics* 10, p. 333. issn : 1471-2164. doi : 10.1186/1471-2164-10-333. url : https://doi.org/10.1186/1471-2164-10-333 (cf. p. 40).

Sanger, F., S. Nicklen et A. R. Coulson (1977). « DNA sequencing with chain-terminating inhibitors ». In : *Proceedings of the National Academy of Sciences of the United States of America* 74.12, p. 5463–5467. issn : 0027-8424 (cf. p. 10).

Schulz, Marcel H. et al. (2012). « Oases : robust de novo RNA-seq assembly across the dynamic range of expression levels ». In : *Bioinformatics (Oxford, England)* 28.8, p. 1086–1092. issn : 1367-4811. doi : 10.1093/bioinformatics/bts094 (cf. p. 19, 46).

Selosse, Marc-AndrÃ©, Marie Charpin et Fabrice Not (2017). « Mixotrophy everywhere on land and in water : the grand Ã©cart hypothesis ». In : *Ecology Letters* 20.2, p. 246–263. ISSN : 1461-0248. DOI : 10.1111/ele.12714. URL : http://onlinelibrary.wiley.com/doi/10.1111/ele.12714/abstract (cf. p. 27).

Shi, Xiao Li et al. (2009). « Groups without Cultured Representatives Dominate Eukaryotic Pico-phytoplankton in the Oligotrophic South East Pacific Ocean ». In : *PLOS ONE* 4.10, e7657. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0007657. URL : http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0007657 (visité le 24/10/2017) (cf. p. 6).

Shoguchi, Eiichi et al. (2013). « Draft Assembly of the Symbiodinium minutum Nuclear Genome Reveals Dinoflagellate Gene Structure ». In : *Current Biology* 23.15, p. 1399–1408. ISSN : 0960-9822. DOI : 10.1016/j.cub.2013.05.062. URL : http://www.sciencedirect.com/science/article/pii/S0960982213006878 (visité le 13/07/2016) (cf. p. 14, 39).

Siano, Raffaele et al. (2010). « Pelagodinium gen. nov. and P. béii comb. nov., a dinoflagellate symbiont of planktonic foraminifera ». In : *Protist* 161.3, p. 385–399. ISSN : 1618-0941. DOI : 10.1016/j.protis.2010.01.002 (cf. p. 8).

Sibbald, Shannon J. et John M. Archibald (2017). « More protist genomes needed ». In : *Nature Ecology & Evolution* 1, p. 0145. ISSN : 2397-334X. DOI : 10.1038/s41559-017-0145. URL : http://www.nature.com/articles/s41559-017-0145 (visité le 21/04/2017) (cf. p. 12, 13, 34, 40).

Sierra, Roberto et al. (2013). « Deep relationships of Rhizaria revealed by phylogenomics : A farewell to Haeckel's Radiolaria ». In : *Molecular Phylogenetics and Evolution* 67.1, p. 53–59. ISSN : 1055-7903. DOI : 10.1016/j.ympev.2012.12.011. URL : http://www.sciencedirect.com/science/article/pii/S1055790312004897 (visité le 03/12/2014) (cf. p. 30).

Siva, Nayanah (2008). « 1000 Genomes project ». In : *Nature Biotechnology* 26.3, p. 256–256. ISSN : 1087-0156. DOI : 10.1038/nbt0308-256b. URL : https://www.nature.com/nbt/journal/v26/n3/full/nbt0308-256b.html (visité le 21/09/2017) (cf. p. 10).

Smith-Unna, Richard et al. (2016). « TransRate : reference free quality assessment of de novo transcriptome assemblies ». In : *Genome Research*, gr.196469.115. ISSN : 1088-9051, 1549-5469. DOI : 10.1101/gr.196469.115. URL : http://genome.cshlp.org/content/early/2016/06/01/gr.196469.115 (visité le 23/06/2016) (cf. p. 45).

Stal, Lucas J. et Mariana Silvia Cretoiu (2016). *The Marine Microbiome : An Untapped Source of Biodiversity and Biotechnological Potential*. Google-Books-ID : MJ1PDAAAQBAJ. Springer. 501 p. ISBN : 978-3-319-33000-6 (cf. p. 37).

Stephens, Zachary D. et al. (2015). « Big Data : Astronomical or Genomical ? » In : *PLOS Biology* 13.7, e1002195. ISSN : 1545-7885. DOI : 10.1371/journal.pbio.1002195. URL : http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195 (visité le 05/10/2017) (cf. p. 22).

Stoecker, Diane K. (1999). « Mixotrophy among Dinoflagellates1 ». In : *Journal of Eukaryotic Microbiology* 46.4, p. 397–401. ISSN : 1550-7408. DOI : 10.1111/j.1550-7408.1999.tb04619.x. URL : http://onlinelibrary.wiley.com/doi/10.1111/j.1550-7408.1999.tb04619.x/abstract (cf. p. 4).

Stoecker, Diane K., Per Juel Hansen et al. (2017). « Mixotrophy in the Marine Plankton ». In : *Annual Review of Marine Science* 9, p. 311–335. ISSN : 1941-0611. DOI : 10.1146/annurev-marine-010816-060617 (cf. p. 4).

Stoecker, Diane K., Matthew D. Johnson et al. (2009). « Acquired phototrophy in aquatic protists ». In : DOI : 10.3354/ame01340. URL : http://darchive.mblwhoilibrary.org/handle/1912/4538 (visité le 15/10/2017) (cf. p. 29).

Stüken, Anke et al. (2011). « Discovery of Nuclear-Encoded Genes for the Neurotoxin Saxitoxin in Dinoflagellates ». In : *PLOS ONE* 6.5, e20096. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0020096. URL : http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0020096 (visité le 07/11/2016) (cf. p. 39).

Sunagawa, Shinichi et al. (2015). « Structure and function of the global ocean microbiome ». In : *Science* 348.6237, p. 1261359. ISSN : 0036-8075, 1095-9203. DOI : 10.1126/science.1261359. URL : http://science.sciencemag.org/content/348/6237/1261359 (visité le 01/10/2017) (cf. p. 11, 211).

Sunda, W. et al. (2002). « An antioxidant function for DMSP and DMS in marine algae ». In : *Nature* 418.6895, p. 317–320. ISSN : 0028-0836. DOI : 10.1038/nature00851. URL : http://www.nature.com/nature/journal/v418/n6895/full/nature00851.html?foxtrotcallback=true (visité le 30/09/2017) (cf. p. 39).

Suzuki, Noritoshi et Yoshiaki Aita (2011). « Radiolaria : achievements and unresolved issues : taxonomy and cytology ». In : *Plankton and Benthos Research* 6.2, p. 69–91. DOI : 10.3800/pbr.6.69 (cf. p. 30, 33).

Suzuki, Noritoshi et Fabrice Not (2015). « Biology and Ecology of Radiolaria ». In : *Marine Protists*. DOI : 10.1007/978-4-431-55130-0_8. Springer, Tokyo, p. 179–222. ISBN : 978-4-431-55129-4 978-4-431-55130-0. URL : https://link.springer.com/chapter/10.1007/978-4-431-55130-0_8 (visité le 02/10/2017) (cf. p. 31, 33, 183).

Suzuki, Noritoshi et Kazuhiro Sugiyama (2001). « Regular axopodial activity of Diplosphaera hexagonalis Haeckel (spheroidal spumellarian, Radiolaria) ». In : *Paleontological Research* 5.2, p. 131–140. DOI : 10.2517/prpsj.5.131 (cf. p. 33).

Takahashi, Osamu et al. (2004). « Molecular phylogeny of solitary shell-bearing Polycystinea (Radiolaria) ». In : *Revue de Micropaléontologie* 47.3, p. 111–118. ISSN : 0035-1598. DOI : 10.1016/j.revmic.2004.06.002. URL : http://www.sciencedirect.com/science/article/pii/S0035159804000297 (cf. p. 34).

Taylor, Gordon (1982). « The role of pelagic heterotrophic protozoa in nutrient cycling : A review ». In : *Annales De L Institut Oceanographique* 58, p. 227–241 (cf. p. 29).

Theis, Kevin R. et al. (2016). « Getting the Hologenome Concept Right : an Eco-Evolutionary Framework for Hosts and Their Microbiomes ». In : *mSystems* 1.2, e00028–16. ISSN : 2379-5077. DOI : 10.1128/mSystems.00028-16. URL : http://msystems.asm.org/content/1/2/e00028-16 (visité le 24/10/2017) (cf. p. 16, 17).

Tomczak, K., P. Czerwińska et M. Wiznerowicz (2015). « The Cancer Genome Atlas (TCGA) : an immeasurable source of knowledge. » In : *Contemporary oncology (Poznan, Poland)* 19.1, A68–77. ISSN : 1428-2526. DOI : 10.5114/wo.2014.47136. URL : http://europepmc.org/abstract/med/25691825 (visité le 28/10/2017) (cf. p. 10).

Toseland, Andrew et al. (2014). « Metatranscriptomes from diverse microbial communities : assessment of data reduction techniques for rigorous annotation ». In : *BMC Genomics* 15, p. 901. ISSN : 1471-2164. DOI : 10.1186/1471-2164-15-901. URL : https://doi.org/10.1186/1471-2164-15-901 (cf. p. 18).

Tu, Qichao, Zhili He et Jizhong Zhou (2014). « Strain/species identification in metagenomes using genome-specific markers ». In : *Nucleic Acids Research* 42.8, e67. ISSN : 0305-1048. DOI : 10.1093/nar/gku138. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4005670/ (cf. p. 209).

Valiadi, Martha et Debora Iglesias-Rodriguez (2013). « Understanding Bioluminescence in Dinoflagellates—How Far Have We Come ? » In : *Microorganisms* 1.1, p. 3–25. ISSN : 2076-2607. DOI : 10.3390/microorganisms1010003. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5029497/ (cf. p. 39).

Vargas, Colomban de, Stéphane Audic et al. (2015). « Eukaryotic plankton diversity in the sunlit ocean ». In : *Science* 348.6237, p. 1261605. ISSN : 0036-8075, 1095-9203. DOI : 10.1126/science.1261605. URL : http://science.sciencemag.org.insb.bib.cnrs.fr/content/348/6237/1261605 (visité le 19/10/2017) (cf. p. 212).

Vargas, Colomban de, Louisette Zaninetti et al. (1997). « Phylogeny and Rates of Molecular Evolution of Planktonic Foraminifera : SSU rDNA Sequences Compared to the Fossil Record ». In : *Journal of Molecular Evolution* 45.3, p. 285–294. ISSN : 0022-2844, 1432-1432. DOI : 10.1007/PL00006232. URL : https://link.springer.com/article/10.1007/PL00006232 (visité le 29/10/2017) (cf. p. 36).

Vaulot, Daniel et al. (2004). « The Roscoff Culture Collection (RCC) : A collection dedicated to marine picoplankton ». In : *Nova Hedwigia* 79, p. 49–70. DOI : 10.1127/0029-5035/2004/0079-0049 (cf. p. 6, 8).

Veldhuis, Marcel J. W., Terry L. Cucci et Michael E. Sieracki (1997). « Cellular Dna Content of Marine Phytoplankton Using Two New Fluorochromes : Taxonomic and Ecological Implications1 ». In : *Journal of Phycology* 33.3, p. 527–541. ISSN : 1529-8817. DOI : 10.1111/j.0022-

3646.1997.00527.x. URL : http://onlinelibrary.wiley.com/doi/10.1111/j.0022-3646.1997.00527.x/abstract (cf. p. 14).

WANG, Da-Zhi (2008). « Neurotoxins from Marine Dinoflagellates : A Brief Review ». In : *Marine Drugs* 6.2, p. 349–371. ISSN : 1660-3397. DOI : 10.3390/md20080016. URL : http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2525493/ (visité le 20/01/2017) (cf. p. 38).

WANG, Zhong, Mark GERSTEIN et Michael SNYDER (2009). « RNA-Seq : a revolutionary tool for transcriptomics ». In : *Nature Reviews Genetics* 10.1, p. 57–63. ISSN : 1471-0056. DOI : 10.1038/nrg2484. URL : http://www.nature.com/nrg/journal/v10/n1/abs/nrg2484.html (visité le 17/07/2014) (cf. p. 15).

WESTREICH, Samuel T. et al. (2016). « SAMSA : a comprehensive metatranscriptome analysis pipeline ». In : *BMC Bioinformatics* 17, p. 399. ISSN : 1471-2105. DOI : 10.1186/s12859-016-1270-8. URL : https://doi.org/10.1186/s12859-016-1270-8 (cf. p. 208).

WORDEN, Alexandra Z. et al. (2015). « Rethinking the marine carbon cycle : Factoring in the multifarious lifestyles of microbes ». In : *Science* 347.6223, p. 1257594. ISSN : 0036-8075, 1095-9203. DOI : 10.1126/science.1257594. URL : http://science.sciencemag.org/content/347/6223/1257594 (visité le 28/10/2017) (cf. p. 5).

YOON, Hwan Su et al. (2011). « Single-Cell Genomics Reveals Organismal Interactions in Uncultivated Marine Protists ». In : *Science* 332.6030, p. 714–717. ISSN : 0036-8075, 1095-9203. DOI : 10.1126/science.1203163. URL : http://science.sciencemag.org.insb.bib.cnrs.fr/content/332/6030/714 (visité le 25/10/2017) (cf. p. 206).

YUASA, Tomoko, Takeo HORIGUCHI et al. (2016). « Gymnoxanthella radiolariae gen. et sp. nov. (Dinophyceae), a dinoflagellate symbiont from solitary polycystine radiolarians ». In : *Journal of Phycology* 52.1, p. 89–104. ISSN : 1529-8817. DOI : 10.1111/jpy.12371. URL : http://onlinelibrary.wiley.com/doi/10.1111/jpy.12371/abstract (visité le 03/05/2016) (cf. p. 8).

YUASA, Tomoko et Osamu TAKAHASHI (2016). « Light and electron microscopic observations of the reproductive swarmer cells of nassellarian and spumellarian polycystines (Radiolaria) ». In : *European Journal of Protistology* 54, p. 19–32. ISSN : 1618-0429. DOI : 10.1016/j.ejop.2016.02.007 (cf. p. 33).

YUKI, Masahiro et al. (2015). « Dominant ectosymbiotic bacteria of cellulolytic protists in the termite gut also have the potential to digest lignocellulose ». In : *Environmental Microbiology* 17.12, p. 4942–4953. ISSN : 1462-2920. DOI : 10.1111/1462-2920.12945. URL : http://onlinelibrary.wiley.com/doi/10.1111/1462-2920.12945/abstract (cf. p. 27).

ZERBINO, Daniel R. et Ewan BIRNEY (2008). « Velvet : Algorithms for de novo short read assembly using de Bruijn graphs ». In : *Genome Research* 18.5, p. 821–829. ISSN : 1088-9051, 1549-5469. DOI : 10.1101/gr.074492.107. URL : http://genome.cshlp.org/content/18/5/821 (visité le 24/06/2014) (cf. p. 19, 46).

Zettler, Linda Amaral, Mitchell L. Sogin et David A. Caron (1997). « Phylogenetic relationships between the Acantharea and the Polycystinea : A molecular perspective on Haeckel's Radiolaria ». In : *Proceedings of the National Academy of Sciences* 94.21, p. 11411–11416. ISSN : 0027-8424, 1091-6490. URL : http://www.pnas.org/content/94/21/11411 (visité le 29/09/2017) (cf. p. 34).

Zhao, Qiong-Yi et al. (2011). « Optimizing de novo transcriptome assembly from short-read RNA-Seq data : a comparative study ». In : *BMC Bioinformatics* 12 (Suppl 14), S2. ISSN : 1471-2105. DOI : 10.1186/1471-2105-12-S14-S2. URL : http://www.biomedcentral.com/1471-2105/12/S14/S2/abstract (visité le 24/06/2014) (cf. p. 19, 21).

Zimorski, Verena et al. (2014). « Endosymbiotic theory for organelle origins ». In : *Current Opinion in Microbiology*. Growth and development : eukaryotes/ prokaryotes 22 (Supplement C), p. 38–48. ISSN : 1369-5274. DOI : 10.1016/j.mib.2014.09.008. URL : http://www.sciencedirect.com/science/article/pii/S1369527414001283 (cf. p. 27).

Zorita, Eduard, Pol Cuscó et Guillaume J. Filion (2015). « Starcode : sequence clustering based on all-pairs search ». In : *Bioinformatics* 31.12, p. 1913–1919. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btv053. URL : https://academic.oup.com/bioinformatics/article/31/12/1913/213875 (visité le 26/10/2017) (cf. p. 136).

Zwaan, G. J van der, F. J Jorissen et H. C de Stigter (1990). « The depth dependency of planktonic/benthic foraminiferal ratios : Constraints and applications ». In : *Marine Geology* 95.1, p. 1–16. ISSN : 0025-3227. DOI : 10.1016/0025-3227(90)90016-D. URL : http://www.sciencedirect.com/science/article/pii/002532279090016D (cf. p. 26).

Glossaire

autotrophe : qualifie un organisme capable de synthétiser de la matière organique à partir de matière minérale (*e.g.* oxygène, hydrogène, carbone, azote), en utilisant, soit l'énergie lumineuse (organisme photo-autotrophe photosynthétique), soit l'énergie chimique (organisme chimio-autotrophe).

chloroplaste : organite présent dans le cytoplasme des cellules eucaryotes, contenant de la chlorophylle, et où se déroule la première phase de la photosynthèse.

chimère (génomique) : séquence n'ayant pas de réalité biologique, souvent composée de deux ou plusieurs fragments de séquences d'origines différentes.

clique (théorie des graphes) : sous-ensemble de sommets (ou noeuds) d'un graphe dans lequel chaque noeud est adjacent (*i.e. voisin de degré 1* aux autres noeuds de ce sous-ensemble.

contig : séquence créée à partir de l'assemblage de lectures issues de séquençage d'ADN ou ARN, et dont la réalité biologique n'est pas encore établie.

composante connexe : graphe dont toute paire de sommets peut être reliée par au moins un chemin (une succession d'arêtes).

couverture de séquence (génomique) : proportion d'une séquence génomique sur laquelle peuvent être alignées une ou plusieurs autres séquences génomiques. Dans le contexte de mon étude, j'emploie la couverture de séquence pour évaluer la qualité des contigs notamment en comparant leur couverture de séquences par rapport aux lectures utilisées au cours de l'assemblage *de novo*.

faux-positifs (au sens statistique) : résultat d'un test statistique déclaré « vrai » (ou « positif ») alors qu'en réalité ce dernier est « faux » (ou « négatif »). Dans le contexte de la génomique, ce terme est employé pour désigner une fausse vérité à propos d'une séquence biologique (*e.g.* le résultat de l'assignation d'une lecture par un programme informatique à une séquence A alors qu'en réalité la séquence d'origine est B).

FPKM (*fragments per million bases of exon per million fragments mapped*) : valeur d'estimation de l'expression de chaque contig, normalisée par le nombre total de lectures alignées et par la longueur des contigs. Plus le nombre de lectures alignées sur un contig est élevé par rapport à la longueur de ce contig et au nombre total de lectures alignées sur tous les contigs, plus la valeur de FPKM est grande.

gène composite : séquence créé avec au moins deux fragments appartenant à des familles de gènes différentes (au sein d'un même organisme, voire appartenant à des

organismes distincts). De nombreux termes ont été utilisés pour désigner ce type de séquence : gène chimérique, gène de fusion, gène codant pour des protéines multi-domaines ou encore gène composite.

hétérotrophe : qualifie un organisme assimilant des substances organiques comme source de carbone.

***machine-learning* :** méthode (également appelée « apprentissage automatique » ou « apprentissage statistique ») permettant à une machine (au sens large) d'apprendre par l'expérience dans le but de prendre une décision adaptée.

***mapping* :** procédé permettant l'alignement des lectures sur un génome ou un transcriptome. Dans le cadre de la transcriptomique, le *mapping* est largement employé pour estimer le niveau d'expression de chacun des transcrits d'un transcriptome (cf. définition du FPKM).

méta-génomique/méta-transcriptomique : ensemble des génomes ou des transcriptomes présents dans un environnement ou un échantillon donné (*i.e.* sols, microbiote instestinal, océan).

métabarcode : séquence (relativement courte) d'ADN ou d'ARN cherchant à marquer et à différencier les différentes espèces au sein d'un échantillon environnemental.

métriques : série de mesures effectuées sur un génome ou un transcriptome nouvellement assemblé. Elles permettent d'évaluer la qualité des séquences nouvellement produites (Annexe).

mixotrophe : qualifie un organisme pouvant alternativement utiliser une source de carbone minéral (autotrophe) ou organique (hétérotrophe).

monophylétique : groupe sur un arbre phylogénétique incluant l'ancêtre commun ainsi que l'ensemble de ses descendants (partageant au moins un caractère dérivé).

N50 : longueur minimale de contig(s) nécessaire(s) pour couvrir 50% d'un génome/transcriptome (*i.e.* la somme des longueurs des contigs de taille supérieure ou égale au N50 contient au moins 50% des séquences du génome ou du transcriptome).

ORF (*open reading frame*) : "cadre de lecture ouvert" correspondant à la région d'une séquence nucléotidique à partir d'un codon START (ATG) jusqu'à un codon STOP.

organisme modèle : (définition propre à cette thèse) organisme pour lequel il existe une ou plusieurs référence(s) génomique(s) (*e.g.* génome, transcriptome, banque d'ESTs).

organisme non-modèle : (définition propre à cette thèse) organisme dont au moins un génome ou un transcriptome de référence n'est pas disponible.

PCR (*polymerase chain reaction*) : technique de biologie moléculaire permettant de dupliquer une séquence d'ADN ou d'ARN).

phylogénomique : méthode permettant d'inférer à partir de multiples données génomiques les relations de parenté entre différents objets (*e.g.* génomes, espèces, lignées du vivant).

protéome : ensemble des protéines d'un organisme dans des conditions données et à un instant donné.

***score phred* :** score de qualité assigné à chacune des bases issues du séquençage d'un génome/transcriptome.

SSN *Sequence Similarity Network* : structure de données correspondant à un graphe dans lequel les noeuds sont des séquences génomiques (*sensu largo*) et les arêtes représentent des alignements globaux (Figure 1.12).

Annexes

A Performances des assembleurs *de novo* de transcriptome Trinity et Velvet/Oases

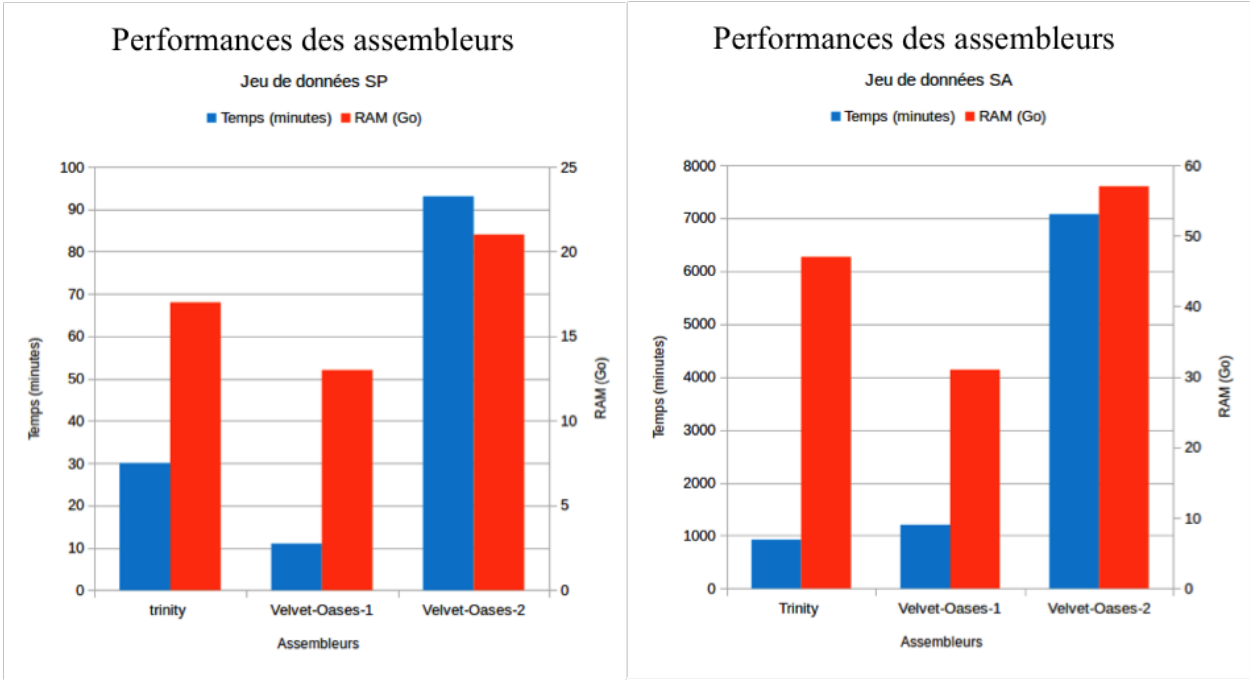


FIGURE 1 – Temps d'exécution et quantité de RAM consommée pour Trinity et Velvet/Oases (VO) appliqués aux jeux de données de RNA-seq de *Schizosaccharomyces pombe* Sp (4 millions de lectures pairés, longueur de 100 pb) et un jeu de données nommé SA (transcriptome simulé dans le cadre du concours Assemblathon [GALLO et al. 2014], 44 millions de lectures pairés, longueur de 100 pb). Deux gammes de valeurs de k ont été testées pour Velvet/Oases : $k \in [21,31]$ et un pas de 4 dans le cas de VO-1 et $k \in [21,63]$ et un pas de 4 pour VO-2.

B Métriques d'évaluation d'assemblage

Métriques utilisées pour l'évaluation des assemblages

Les métriques sont un ensemble de mesures permettant d'évaluer les assemblages de lectures. Elles sont couramment employées pour évaluer la qualité des assemblages de séquences ou encore pour comparer deux jeux de données | Ci-dessous, quelques-unes des métriques les plus utilisées sont présentées.

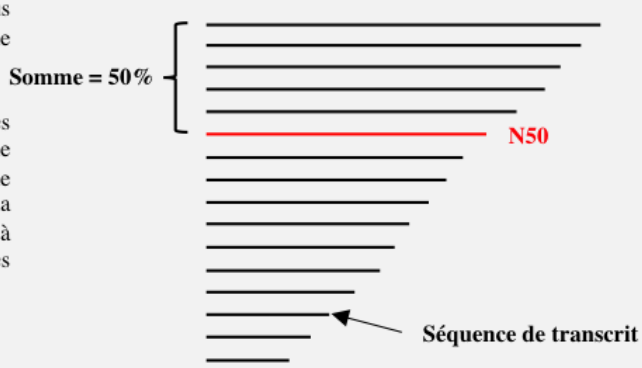
Nombre de transcrits :
Nombre de séquences assemblées à la fin de la procédure d'assemblage.

Taille totale de l'assemblage :
Nombre de nucléotides composant la totalité des séquences de transcrits.

Taille moyenne des transcrits :
Moyenne des tailles pour toutes les séquences assemblées.

N50 :
Longueur du transcrit à laquelle les séquences les plus grandes couvrent 50% de la longueur totale de l'assemblage.

Ex : Si on considère un ensemble de séquences de tailles aléatoires triées par longueur décroissante, et avec une longueur totale (somme des longueurs de séquences) de 2000 nucléotides. Sur l'ensemble des séquences, la valeur de N50 correspond à la taille de la séquence à partir de laquelle la somme des longueurs des séquences précédentes est de 1000 nucléotides.



Pourcentage GC (%GC) :
Proportion de nucléotides, pour la totalité des transcrits assemblés, étant soit une cytosine (C), soit une guanine (G).

FIGURE 2 – Présentation des métriques* d'assemblage

C Article en collaboration

SCIENTIFIC REPORTS

OPEN

Acclimation of a low iron adapted *Ostreococcus* strain to iron limitation through cell biomass lowering

Received: 17 October 2016
Accepted: 14 February 2017
Published online: 23 March 2017

Hugo Botebol[1], Gaelle Lelandais[2], Christophe Six[3], Emmanuel Lesuisse[2], Arnaud Meng[4], Lucie Bittner[4], Stéphane Lecrom[4], Robert Sutak[5], Jean-Claude Lozano[1], Philippe Schatt[1], Valérie Vergé[1], Stéphane Blain[1] & François-Yves Bouget[1]

Iron is an essential micronutrient involved in many biological processes and is often limiting for primary production in large regions of the World Ocean. Metagenomic and physiological studies have identified clades or ecotypes of marine phytoplankton that are specialized in iron depleted ecological niches. Although less studied, eukaryotic picophytoplankton does contribute significantly to primary production and carbon transfer to higher trophic levels. In particular, metagenomic studies of the green picoalga *Ostreococcus* have revealed the occurrence of two main clades distributed along coast-offshore gradients, suggesting niche partitioning in different nutrient regimes. Here, we present a study of the response to iron limitation of four *Ostreococcus* strains isolated from contrasted environments. Whereas the strains isolated in nutrient-rich waters showed high iron requirements, the oceanic strains could cope with lower iron concentrations. The RCC802 strain, in particular, was able to maintain high growth rate at low iron levels. Together physiological and transcriptomic data indicate that the competitiveness of RCC802 under iron limitation is related to a lowering of iron needs though a reduction of the photosynthetic machinery and of protein content, rather than to cell size reduction. Our results overall suggest that iron is one of the factors driving the differentiation of physiologically specialized *Ostreococcus* strains in the ocean.

Iron is an essential element for all living organisms and in particular for photosynthetic phytoplanktonic cells in which numerous iron-sulphur centre and heme containing proteins are involved in the photosynthetic electron transfer and the nitrate reduction reactions. The extensive use of iron in the cellular machinery of modern micro-organisms is inherited from the growth of their ancestors in the iron rich Archean ocean. However the geochemical evolution of the chemical composition of the ocean, largely driven by the evolution of life, has led to vanishing iron concentrations in the present ocean[1]. Thus, iron is a limiting resource in very large regions of the ocean.

Phytoplanktonic species are diverse in size, shape, phylogenetic origin and have evolved specific strategies to colonize ecological niches where iron is limiting (for a review see ref. 2). Physiological and molecular responses to iron limitation have been well studied in nano and microphytoplankton. Cultures of different species have revealed the critical role surface/volume ratio to deal with iron limitation[3,4] but several other adaptations to low iron conditions have also been reported. In oceanic diatoms, for example, the lowering of iron needs can be achieved by optimizing the architecture of the photosynthetic machinery through a decrease of the iron-rich

[1]Sorbonne Universités, Université Pierre et Marie Curie (Paris 06) & Centre National pour la Recherche Scientifique CNRS, UMR 7621, Laboratoire d'Océanographie Microbienne, Observatoire Océanologique, F-66650, Banyuls/mer, France. [2]Université Paris Diderot (Paris 07), Centre National de la Recherche Scientifique, Institut Jacques Monod, F–75013, Paris, France. [3]Sorbonne Universités, Université Pierre et Marie Curie (Paris 06) & Centre National pour la Recherche Scientifique, UMR 7144, Adaptation et Diversité en Milieu Marin, Equipe Marine Phototrophic Prokaryotes, Station Biologique de Roscoff, 29680, Roscoff Cedex, France. [4]Sorbonne Universités, UPMC Univ Paris 06, CNRS, Institut de Biologie Paris-Seine (IBPS), Evolution Paris Seine, F-75005, Paris, France. [5]Department of Parasitology, Faculty of Science, Charles University, 12844, Prague, Czech Republic. Correspondence and requests for materials should be addressed to S.B. (email: stephane.blain@obs-banyuls.fr) or F.-Y.B. (email: fy.bouget@obs-banyuls.fr)

complexes such as photosystem I and cytochrome b_6/f[5–7]. Iron storage into the ferritin protein pool also constitutes an efficient physiological strategy to cope with sporadic iron supply in pennate diatoms[8].

The acclimation and adaptation strategies are less understood in picophytoplankton, despite the fact that small phytoplanktonic cells, such as the tiny picocyanobacterium *Prochlorococcus*, numerically dominate oligotrophic water phytoplanktonic communities[9,10]. Metagenomic and physiological studies in iron depleted regions point to the presence of *Prochlorococcus* clades adapted to iron scarcity[11–13]. These adaptations could notably rely on the selection of specific gene repertoires during the evolution, such as the replacement of iron Super Oxide Dismutase (SOD) by Nickel SOD[14,15]. Together, genomic and phylogeographic studies strongly suggest the existence of iron adapted lineages in prokaryotic picophytoplankton. A recent study in *Synechococcus* revealed that a coastal strain acclimates to changes in Fe concentration by modulating iron uptake, storage, and photosynthetic proteins[16].

We recently unveiled a central role for ferritin in the day-night regulation of iron homeostasis in the picoeukaryotic alga *Ostreococcus tauri* and showed that the transcriptional response to iron limitation in this microorganism is tighly dependent on the diurnal cycle[17,18]. Still, the adaptive responses of eukaryotic picophytoplankton to low iron conditions remain largely unexplored. Within the photosynthetic picoeukaryotes, the order of *Mamelliales* (*Mamiellophyceae*, *Chlorophyta*), encompassing the genera *Ostreococcus*, *Bathycoccus* and *Micromonas*, has a worldwide geographic distribution and it has been shown to contribute significantly to primary production in coastal ecosystems[19–22]. Ribosomal DNA sequence phylogenies segregate *Ostreococcus* sp into 4 phylogenetic clades (A-D) and physiological studies have demonstrated the existence of high- and low-light adapted strains in the genus *Ostreococcus*[23–26]. Specific primers were successfully developed to amplify Ostreococcus clades A, B and C 18S ribosomal DNA, enabling the study of niche-partitioning in natural populations[27]. Coastal, meso- to eutrophic areas were dominated by clades A and C (renamed clade Ol), whereas open-ocean, oligotrophic regions were dominated by clade B (renamed clade OII)[27]. This study also indicated that the distribution of *Ostreococcus* lineages is not explained primarily by light irradiance, and that the specialization in different nutrient regimes along coast-offshore gradients might be a key driver of picoeukaryotes evolution/diversification[27].

In this paper, we compare the physiological responses to iron limitation of several *Ostreococcus* strains inhabiting different environments. Using a combination of physiological and transcriptomic approaches, we unveil several aspects of the specialization of *Ostreococcus* to environments with different levels of iron bioavailability. In particular, we have characterized a Mediterranean strain, RCC802, which shows very low iron requirement, and whose main adaptation trait to iron limitation is a marked cell biomass reduction.

Material and Methods

Algal Strains and culture conditions. *Ostreococcus* strains were obtained from the Roscoff Culture Collection (http://www.roscoff-culture-collection.org/): *Ostreococcus tauri* strain OTTH595 (RCC745; isolated from Thau Lagoon, France), *Ostreococcus* sp. RCC802 (isolated at 65 m in the Sicily channel, Italy), *Ostreococcus* sp. RCC809 (isolated from 105 m in the tropical Atlantic Ocean), and *Ostreococcus* sp. RCC789 (strain BL_82-7_ clonal, isolated from surface water of Barcelona harbour, Spain).

Cells were grown at 20 °C under 25 µmol quanta $m^{-2} s^{-1}$ of constant blue light (blue led, $\lambda_{MAX} = 465$ nm) in AQUIL medium[17,28] containing concentrations of Fe(III)-EDTA ranging between 5.4 and 270 nM, corresponding to the theoretical solubility limit of iron in sea water. All culture manipulations were conducted in clean room (class 10,000) equipped with a laminar flow hood (class 100). Synthetic ocean water (SOW) and solutions of inorganic nutrients (NO_3^- and PO_4^{3-}) were separately purified by removing trace metals using a Chelex 100 ion exchange resin (Bio-Rad). Final nutrients concentrations were 300 µM NO_3^- and 10 µM PO_4^{3-}. Trace metal solutions were buffered with 0.1 M of EDTA. For iron content determinations, cells were grown in modified F (Mf) medium[29,30] and iron was provided as radioactive ^{55}Fe(III)-EDTA (1:20).

Growth rates and cell diameter measurement. To remove contaminating iron and deplete intracellular iron storage, *Ostreococcus* strains were first acclimated during 5 days in 1.5 mL of AQUIL medium containing 5.4 nM Fe(III)-EDTA (1:1). To study the growth response function of iron supply, triplicate flasks containing various concentrations of Fe(III)-EDTA ranging from 5.4 nM to 270 nM Fe(III)-EDTA in Aquil medium, were inoculated with 10^6 cells mL^{-1} of iron-depleted cells. These cultures were grown at 20 °C, under 25 µmol quanta $m^{-2} s^{-1}$ of continuous blue light, as described above. Cell density and chlorophyll (Chl) red fluorescence were measured using a flow cytometer (BD Accury C6). Growth rates were computed as the slope of a Ln (Nt) *vs*. time plot, where Nt is the cell abundance (cell/mL) number at time t. Average cell diameter was estimated using a CASY multi-channel cell counting system. (Schärfe System GmbH, Germany).

Determination of photosynthetic parameters and chlorophyll quantitation. Light-response curves were recorded using a pulse amplitude modulated fluorometer (Phyto-PAM, Walz) connected to a chart recorder (Labpro, Vernier). Chlorophyll contents were determined using HPLC (Hewlett-Packard HPLC 1100 Series). Details are given in Supplementary information online.

Intracellular iron content measurements. Cells were grown in Mf medium devoid of iron for a week, and were then transferred to Mf medium containing 1 to 100 nM of ^{55}Fe (III) EDTA (1:20, 29,600 MBq/mg). The iron cell content was determined by scintillation counting after bleaching photosynthetic pigments with sodium hypochlorite, as previously described[17,29].

Protein extraction and quantitation. The different *Ostreococcus* strains were grown in AQUIL medium containing various concentrations of Fe(III)-EDTA, in 24-wells microplates, under 20 µmol quanta $m^{-2} s^{-1}$ of blue light irradiance at 20 °C. After 6 days, cells were harvested in 2 mL micro-tubes by centrifugation at 8 000 × *g* for 10 min. Dry pellets were frozen in liquid nitrogen ant stored at at −80 °C until extraction. All manipulations

Strain name	*OTTH595*	*BL_82-7_clonal*	*Eum16BBL_clonal*	*PROSOPE_44_clonal*
#RCC	745	789	809	802
Clade[1]	C	D	B	A
Clade[2]	OI		OII	OI
Latitude[3]	+43° 24′	+41° 23′	+21° 2′	+36° 29′
Longitude[3]	+3° 36′	+2° 10′	−31° 8′	+13° 19′
Depth (m)	Surface	Surface	105	65
Region	Thau lagoon, France	Barcelona harbour, Spain	Tropical Atlantic Ocean	Sicily channel, Italy
Trophic level	Meso/eutrophic	Mesotrophic	Oligotrophic	Oligotrophic

Table 1. Information regarding the *Ostreococcus* strains used in this study. [1]According to refs 24, 49. [2]According to refs 24, 49. [3]Localisation on a map in Supplementary Figure 1.

were carried out on ice. Cell pellets were ground in 50 µL extraction buffer (50 mM Tris-HCl pH 7.4, 100 mM NaCl, 5 mM EDTA, 0.5% (v/v) Nonidet NP40 (Sigma), 10% (v/v) glycerol, using a Tissue Lyser system (Quiagen). The samples were then centrifuged at 10 000 × *g* for 10 min to remove filter debris. The supernatant was collected, and the total protein concentration was determined using BCA (Bicinchoninic Acid) assay as described elsewhere[31].

Oxygen measurements. Oxygen levels were measured using a 24-channel sensor dish oxygen reader (Presens, Regenburg Germany). Cells were grown for one week in sealed OxoDishes® 4 mL vials, in AQUIL medium containing various concentrations of Fe(III)-EDTA. The sensor was placed at 20 °C under blue light. Before recording, light was switched off for 12 hours and the oxygen levels were then measured for 1 h under 20 µmol quanta $m^{-2} s^{-1}$ blue light. Oxygen consumption, *i.e.* respiration, was subsequently monitored during 1 hour in darkness. Positive control consisted of cell-free medium oxygenated for 1 h. The negative control (0% O_2) was obtained by dissolving 0.2 g of sodium L-ascorbate in 0.1 M NaOH.

RNA extraction and sequencing. Cell cultures were grown for 5 days in 100 mL AQUIL medium containing 5.4 nM (−Fe condition) or 270 nM (+Fe condition) Fe(III)-EDTA. Cultures were placed under 12:12 light/dark cycles of blue light at 40 µmol quanta $m^{-2} s^{-1}$. Three cultures were harvested at 3 h, 9 h, 15 h, 22 h (time 0 h corresponded to dawn and time 12 h to dusk), by centrifugation for 10 min at 8 000 × *g*. Pellets were frozen in liquid nitrogen and stored at −80 °C. RNA extraction was carried out as described by Moulager *et al.*[32]. All manipulations were carried out on ice. The 24 RNA samples (12 +Fe, 12 –Fe) were submitted to high-throughput RNA sequencing using Illumina HiSeq™ 2000 custom pair-end stranded sequencing (Fasteris, Switzerland). After quality controls and filtering of low quality reads (bases quality lower than 15 on sliding window size of 4 bases), clean reads from all samples were gathered, resulting in a dataset of around 300 millions reads. Since there is no reference genome in RCC802, webuild by de novo assembly a reference transcriptome with the TRINITY program in version 2.1.1[33] using the default parameters and paired-end method. The transcript expression levels in each sample was determined by mapping each read from the sample to the reference transcriptome using BOWTIE[34] and counting the number of aligned reads using BEDTOOLS[35]. We next applied the DEseq program[36] to identify differentially expressed genes and calculate an associated risk of error (p-value). Different filtering procedures were finally performed in order to remove potential chimera. Systematic sequence comparisons with *O. tauri* (http://bioinformatics.psb.ugent.be/orcae/overview/OsttaV2) coding sequences were performed using tblastx program in version 2.2.29+[37] with an evalue of $10 e^{-30}$. The whole procedure is detailed in Supplementary information online.

Results

Growth rates of *Ostreococcus* strains under iron limitation. We selected a panel of *Ostreococcus* strains isolated in contrasted trophic regimes and representative of each of the different clades (Table 1). All strains were isolated in the Mediterranean Sea, but the clade B low light ecotype RCC809, which was isolated in the tropical Atlantic Ocean at 105 m depth. Two coastal strains were included in the study, RCC789 (clade D), from Barcelona harbour, and the lagoon strain OTTH595 (clade C) known as *Ostreococcus tauri*, the first *Ostreococcus* species discovered in the *Thau* lagoon in 1995[38]. Both *Ostreococcus tauri* and RCC789 come from eutrophic areas where nutrient bioavailability is high, including iron. RCC802 (clade A) has been isolated at a 65 m depth between Sicily and Tunisia during the PROSOPE cruise[39]. Both RCC802 and RCC809 come from nutrient poor environments that display low Chl *a* concentrations in surface waters (Supplementary Figure 1)[40]. These geographic zones, however, are exposed to sporadic iron fertilization by aeolian mineral dust from the Sahara desert[41,42].

Ostreococcus strains were acclimated in AQUIL medium containing 5.4 nM Fe(III)-EDTA), to set the cells at the minimum iron level. The growth rates were then measured in response to a supply of various concentrations of total Fe, ranging from 5.4 to 270 nM. Upon transfer to iron replete conditions (270 nM total Fe), RCC802 displayed higher growth rates than the three other strains (Fig. 1). This strain outcompeted all other strains, maintaining high growth rates (~1 d^{-1}) for all Fe concentrations, except for 5.4 nM total Fe concentration, at which a 20% decrease in growth rate ($0.8 \pm 0.09\ d^{-1}$) was observed. The coastal strains OTTH595 and RCC789 were the most sensitive to iron limitation. Their growth rates dropped dramatically for total Fe concentrations lower than 270 nM ($0.42\ d^{-1} \pm 0.02$ at 270 nM to $0.16\ d^{-1} \pm 0.04$ at 54 nM total Fe for OTTH595; no growth of RCC789 from

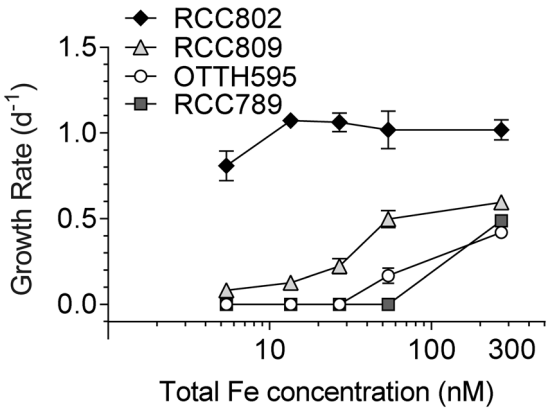


Figure 1. Iron requirements of *Ostreococcus strains*. Growth rates of *Ostreococcus* strains, OTTH595, RCC789, RCC809 and RCC802, were determined in response to various iron supply. Cells were first acclimated for one week in low-iron Aquil medium (5.4 nM Fe(III)-EDTA) before being transferred to Aquil medium containing various concentrations of Fe(III)-EDTA (5.4 nM to 270 nM Fe(III)-EDTA). Growth rates were determined in exponential phase. Mean ± SD of 3 experiments.

54 nM). RCC809 displayed an intermediate response with a growth rate of 0.5 d^{-1} and of 0.15 d^{-1} at 54 nM and 5.4 nM total Fe, respectively.

Photosynthetic parameters of iron limited cells. The efficiency of light energy conversion by photosynthesis was studied using PAM fluorimetry (Fig. 2). Since the coastal strains OTTH595 and RCC789 exhibited very similar responses to iron limitation in terms of growth rates (Fig. 1), only OTTH595 together with RCC802 and RCC809 were further studied. In our standard culture conditions, all strains displayed optimal F_V/F_M values of about 0.65, as reported in previous studies[26]. We compared the photosystem II quantum yield of the three strains in iron replete and limiting conditions, corresponding to the lowest iron concentrations at which cell growth was observed, *i.e.* 54, 27 and 5.4 nM of total Fe for OTTH595, RCC809 and RCC802, respectively (Fig. 1). Under iron limitation, the photosystem II quantum yield of OTTH595 and RCC809 strains decreased by about 50% compared to iron replete conditions (from 0.64 to 0.34 for OTTH595 and from 0.59 to 0.31 for RCC809). For RCC802, only a moderate decrease of about 10% (from 0.61 to 0.54) was observed (Fig. 2A).

Light response curves were recorded to evaluate photosynthetic capacities. The maximal relative Electron Transfer Rate ($rETR_{MAX}$) in photosystem II, a key photosynthetic complex located in the thylakoidal membranes of the plastid, dramatically dropped by more than 50% for both OTTH595 and RCC809 respectively. By contrast, no significant change in rETRmax values was observed for RCC802 (Fig. 2B). The α parameter, which corresponds to the initial slope of the light response curve, is related to the light harvesting efficiency of PSII at limiting light irradiance and thus provides indications on the functional size of the photosystem II antenna. Under iron limitation the α parameter of OTTH595 and RCC809 was 35% lower than under replete conditions, indicating reduced light harvesting capacities. As for $rETR_{MAX}$, RCC802 show less than 10% decrease in the α parameter between iron depleted and replete conditions (Fig. 2C).

Iron content. Cellular iron content was determined after 3 days of incubation of exponentially growing cells in the presence of various concentrations of ^{55}Fe(III)-EDTA (Fig. 3). Under iron replete condition of 270 nM total Fe, similar iron contents were detected in OTTH595 (1.1 ± 0.1 10^{-15} mol ^{55}Fe $cells^{-1}$), RCC809 (1.10 ± 0.09 10^{-15} mol ^{55}Fe $cells^{-1}$) and RCC802 (1.28 ± 0.09 10^{-15} mol ^{55}Fe $cells^{-1}$). In OTTH595 and RCC809 Fe contents remained fairly constant for Fe concentrations of 27 and 54 nM (between 0.85 ± 0.04 and 1.20 ± 0.05 10^{-15} mol ^{55}Fe $cells^{-1}$). In RCC802, in contrast, the iron content dropped progressively from 1.51 ± 0.04 down to 0.040 ± 0.003 10^{-15} mol ^{55}Fe $cells^{-1}$, corresponding to a 19 fold reduction. At 5.4 nM total Fe, RCC809 and OTTH595 also had low Fe contents (0.39 ± 0.02 and 0.22 ± 0.02 10^{-15} mol ^{55}Fe $cells^{-1}$), however no cell growth was observed at this Fe concentration for these two strains (see Fig. 1).

Cell size and cell biomass reductions in response to iron limitation. We assessed the variations of the cellular volume in response to extracellular iron concentrations (Fig. 4A). The cell surface was inferred from the average cell diameter determined using a CASY cell counter, assuming a spherical shape of these coccoid cells. When cells were transferred from iron replete conditions (270 nM total Fe) to limiting conditions, both OTTH595 and RCC809 reduced their cellular volume, from 2.72 μm^3 to 2.09 μm^3 under limitation (108 nM of total Fe) for OTTH595 and from 2.45 μm^3 down to 2.12 μm^3 at 54 nM of total Fe for RCC809. For lower iron concentrations, however, the cellular volume of RCC809 was similar to the initial value, *i.e.* in iron replete conditions (2.5 μm^3). In sharp contrast, the cellular volume of the RCC802 strain was about 4 fold smaller (~0.6 μm^3) than the two other strains, and remained constant independently of the external iron concentrations.

The Chl red fluorescence parameter (FL3), as determined by flow cytometry, provides a reliable proxy of the Chl content in *Ostreococcus* cells (Supplementary Figure 2). Figure 4B shows that for all strains, the cell Chl fluorescence decreased in response to iron limitation from 36.4 ± 0.4 down to 11.7 ± 0.1 in RCC802), from 59.7 ± 0.3

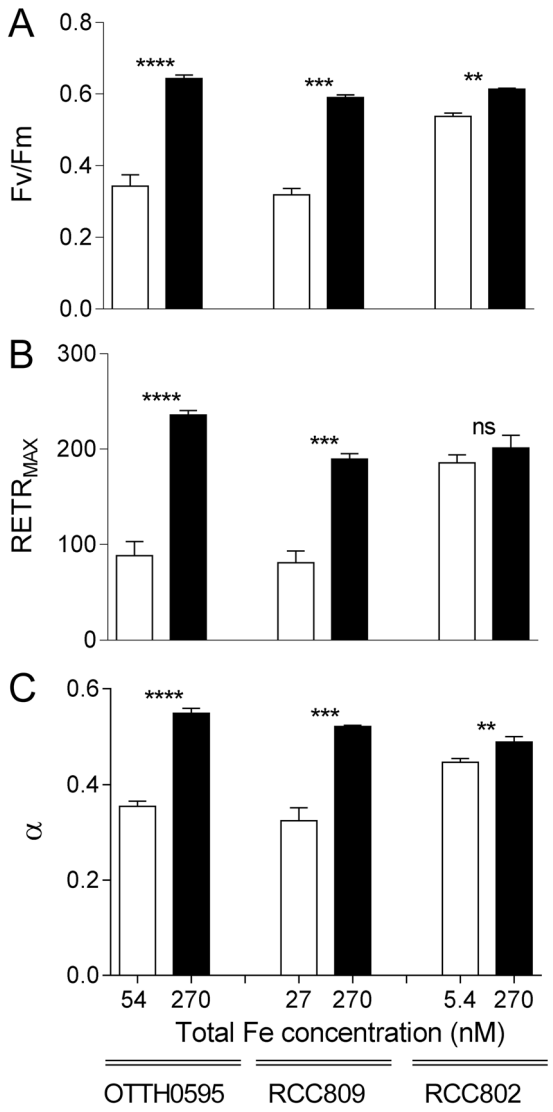


Figure 2. Effect of iron limitation on photosynthetic parameters as derived from light response curves. (**A**) Maximum photosystem II quantum yield of light photoconversion (F_V/F_M), measured in the dark, (**B**) Maximal relative Electron Transfer Rate ($rETR_{MAX}$) in photosystem II and (**C**) Initial slope of the light response curve (α), reflecting the functional photosynthetic antenna size. The white boxes correspond to iron limiting concentrations (OTTH595: 54 nM; RCC809: 27 nM; RCC802: 5.4 nM) and the black boxes to iron replete conditions (270 nM). Mean ± SD from 3 experiments.

down to 41.5 ± 0.3 in RCC809 and from 76.4 ± 0.8 down to 55.3 ± 1.3 in OTTH595. The cellular protein content was determined under the same conditions (Fig. 4C). Under iron replete conditions, the protein cell content was about 2 to 3 fold lower in RCC802 (0.170 ± 0.02 pg $cell^{-1}$) than in OTTH595 (0.464 ± 0.005 pg $cell^{-1}$) and RCC809 (0.3 ± 0.04 pg $cell^{-1}$). As for the cell Chl fluorescence, the protein content of RCC802 decreased dramatically in response to iron limitation (0.11 ± 0.01 pg $cell^{-1}$ at 5.4 nM Fe(III)-EDTA corresponding to a 35% decrease). In OTTH595 and RCC802, smaller reductions in protein contents were observed, *i.e.* from 0.31 down to 0.29 pg $cell^{-1}$ in RCC809 (5% decrease) and from 0.46 to 0.36 pg $cell^{-1}$ in OTTH595 (20% decrease).

Normalized oxygen production of RCC802. The respiration and the oxygen net evolving cellular rates were measured in the RCC802 strain, under iron depleted (5.4 nM) and replete conditions (108 nM). The oxygen production, resulting from the activity of photosystems II, decreased by 50% under iron limitation, while the consumption due to respiration was four-times lower (Supplementary Figure 3). The net oxygen production (production - consumption) was reduced from 7.97 to 3.59 fM min^{-1} $cell^{-1}$, corresponding to a 55% lowering (Fig. 5A). Interestingly, when normalized to the Chl red fluorescence, the net production of oxygen remained fairly stable in iron replete and depleted conditions (Fig. 5B).

Transcriptomic response of RCC802 to iron limitation. A transcriptomic study was conducted to get further insights into the regulation of mechanisms underlying acclimation and adaptation of RCC802 to iron depleted environments. RCC802 cultures were exposed to 12 h day:12 h night cycles under iron limited

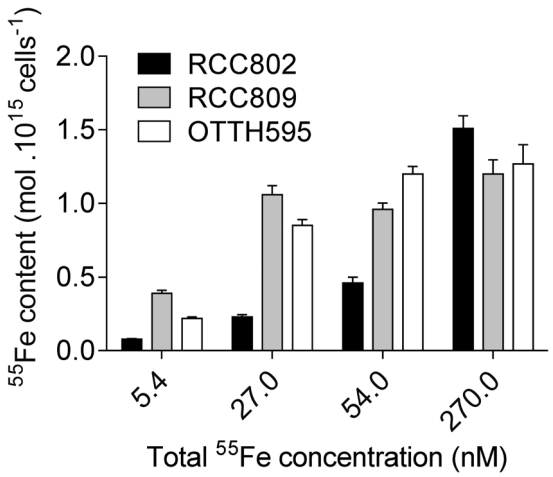


Figure 3. Iron content. *Ostreococcus* cells were grown for 3 days with 5.4 to 270 nM radioactive ^{55}Fe(III)-EDTA. Cellular radioactive iron content was determined by liquid scintillation. Mean ± SD from three experiments.

(5.4 nM Fe(III)-EDTA) and under iron replete conditions, for 7 days. RNA extracted from cells were harvested at 3, 9, 12, 15 and 22 h (i.e. 3 h before and after dawn and dusk) and subjected to RNAseq analysis, as described in Supplementary Figure 4. As there is no reference genome for RCC802, we undertook a *de novo* assembly approach to generate a full-length "reference transcriptome assembly" composed of all predicted transcripts produced by RCC802 in at least one experimental condition (Supplementary Figure 4). The number of reads, mapped on each reference transcript was used to calculate Log_2(Fold Change) (LogFC) values (comparing −Fe and +Fe growth conditions) for each time point. Out of 2233 differentially expressed transcripts 980 were found to be upregulated (LogFC > 1, p-value < 0.01) and 1253 downregulated (LogFC < −1, pvalue < 0.01). After isoform filtering and removal of fusion transcripts based on systematic sequences comparisons with OTTH595, 1251 differentially expressed transcripts were kept (see Supplementary Data File 1). Based on the observation that a vast majority the genes of *O. tauri* are strongly regulated by the day/night cycle[43], we focussed on genes that are induced or repressed in response to iron limitation at all day times. In iron depleted conditions 64 transcripts were downregulated (Fig. 6B, Supplementary Table 1). About one third of the sequences (23) were related to photosynthesis including (i) photosystem proteins such as chlorophyll a/b binding light-harvesting protein of PSI (Lhca proteins), and chlorophyll synthesis such as the Coproporphyrinogen III oxidase, (ii) components of the Calvin-Benson cycle such as the 1,5 ribulose bisphosphate carboxylase oxygenase (RubisCO) and the fructose-1,6-bisphosphatase. The second largest class of repressed transcripts (14) encodes enzymes involved in amino acid metabolism and components of the translation machinery such as translation elongation/initiation factors. Among the 15 remaining downregulated genes with putative homologues or functional domains in other organisms, we identified several iron binding proteins including cytochrome P450, cytochrome b_{561} and a putative mitochondrial ferric reductase (Supplementary Table 2).

Only 13 transcripts were induced across all daytimes. A RCC802 flavodoxin with no homologue in *O. tauri* was overexpressed in all iron-depleted conditions, whereas the iron-containing ferredoxin was repressed under the same conditions (Fig. 6C). A putative Basic Helix loop Helix transcription factor was the only transcription factor induced at all day times under iron limitation.

Discussion

Iron adapted strains in the genus *Ostreococcus*. Using iron limitation experiments, we were able to point out, among several *Ostreococcus* strains, differential abilities to grow in iron limiting conditions. Three different types of response are observed (Fig. 1). The coastal strains OTTH595 and RCC789 were capable of growing only under iron concentrations higher than 54 and 108 nM total Fe respectively, while the oceanic RCC809 strain showed growth down to 27 nM total Fe. RCC802, in contrast, grew well at all Fe concentrations tested with only a 20% reduction in growth rate at 5.4 nM total Fe. Comparison of growth rate inhibition between iron replete and limiting conditions clearly supports that RCC802 has the ability to grow under severe iron limitation like oceanic species such as the diatom *Thalassiosira oceanica* or the coccolithophorid *Emiliana huxlyeyi* (Table 2). In contrast, strains isolated in meso- to eutrophic environments showed much higher iron requirements. These differences in growth rate response were associated to differential physiological responses, such as the modulation of the chlorophyll and protein cell content, the photosynthetic activity, the iron cell content in response to iron limitation (Figs 1, 2, 3 and 4). RCC802 and to some extent RCC809 strains come from geographic areas where iron bioavailability is on average low, with only sporadic iron supplies from Sahara.

Our study overall strongly supports the existence of *Ostreococcus* strains or ecotypes physiologically adapted to environments where iron concentration may be low[41]. Interestingly, RCC802 showed higher growth rates than the other strains not only under iron depleted conditions, but also following iron addition to iron-limited cells (270 nM Fe(III)-EDTA condition in Fig. 1). This suggests that this strain may not be specialized to low iron environments but rather acclimates efficiently to fluctuating environmental iron bioavailability. It is also possible that OTTH595 and RCC809 did not perform as well as RCC802 under iron replete conditions because they had been

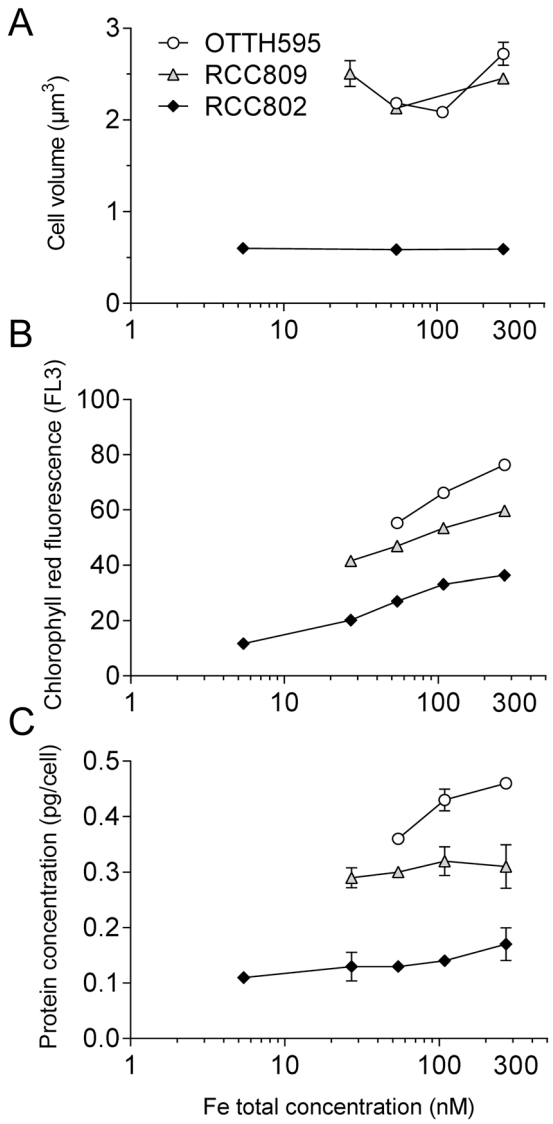


Figure 4. Cell biomass reduction in response to iron limitation. *Ostreococcus* strains were grown for one week in AQUIL medium containing various concentration of Fe(III)-EDTA ranging from 5.4 to 270 nM. (**A**) Cell volume inferred from cell diameter measurement. (**B**) Cellular chl fluorescence parameter measured by flow cytometry (**C**) Protein cell content. Mean ± SD from 3 experiments.

more severely limited during the acclimation phase. The cellular iron quotas of RCC802 varied by about 20 fold between iron replete and iron depleted conditions (while maintaining cellular growth rate), thereby exemplifying the plastic response of RCC802 to fluctuating iron bioavailability. OTTH595 and RCC809, in contrast, showed little variation in iron content between iron replete (270 nM) and the lowest iron concentration at which cell growth occurred (54 and 27 nM, respectively).

Field studies have revealed that *Ostreococcus* OI (clade A and C) and OII (clade B) are rarely abundant at the same location and that OI populations are dominant in coastal, cool and/mesotrophic areas wheras OII populations are more abundant in warm, deep oligotrophic regions[27]. In agreement with these data OTTH595 (clade C) exhibits higher iron requirements than the deep strain RCC809 (clade B). Surprisingly, however, RCC802 (clade A) which belongs to OI clade and was isolated at 65 m deep, has very low iron requirements compared to other strains. Together these data indicate that like light irradiance, iron is a factor which influences the ecological niche partitioning of *Ostreococcus* strains.

Cell biomass reduction: a strategy to cope with iron limitation. Our observations showing that under iron limiting conditions, RCC802 exhibits the highest growth rate and the smallest cell diameter (compared to RCC809 and OTTH595) are consistent with the hypothesis that cell iron supply is limited by diffusion[3,44,45]. The smallest cells are more efficient to use iron because the supply of iron per unit of cellular volume is inversely related to the square of the cellular radius. We observed, however, that the size of RCC802 did not change when iron concentration decreased. RCC802 cells may have reach a minimal cell size, may be due to physical constraints

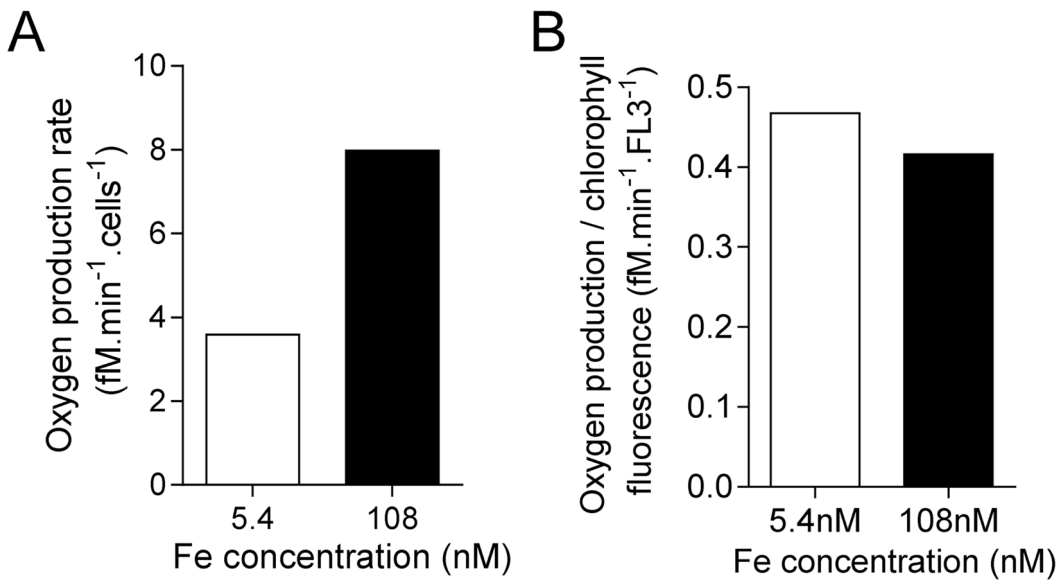


Figure 5. Oxygen evolving in RCC802. (**A**) Rate of net oxygen evolving in RCC802. Oxygen production resulting from photosynthesis was measured during one hour, at 20 °C under 25 µmol quanta $m^{-2} s^{-1}$ blue light (see Supplementary Figure 2). Oxygen consumption resulting from respiration during one hour in darkness was subtracted from production to obtain the net production. (**B**) Net oxygen production determined in (**A**) normalized to the chlorophyll red fluorescence parameter (FL3) measured by flow cytometry.

Organisms	*Growth rate ratio −Fe/+Fe*	*−Fe/+Fe conditions*
*Pelagomonas calceolata**	0.92	4.3 nM/232 nM
*Emiliana huxleyi**	0.88	3.9 nM/299 nM
*Thalassiosira oceanica**	0.70	4.2 nM/303 nM
*Prorocentrum minimum**	0.42	4.3 nM/299 nM
*Thalassiosira weissflogii**	0.00	4.3 nM/301 nM
Ostreococcus‡		
RCC802	0.79	5.4 nM/270 nM
RCC809	0.14	5.4 nM/270 nM
	0.38	27 nM/270 nM
OTTH595	0.00	27 nM/270 nM
	0.40	54 nM/270 nM
RCC789	0.00	54 nM/270 nM

Table 2. Comparison of growth rates in different phytoplanktonic organisms under iron limiting and iron replete conditions in aquil medium (Fe(III)-EDTA source). *From ref. 50. ‡This study.

such as mitotic spindle organization, which precludes further cell size reduction under iron limitation[46]. The main strategy of this strain to deal with iron limitation would, thus, rely on the optimisation of iron use rather than on cell size reduction. In agreement with this hypothesis we observed a nearly 20 fold decrease in intracellular iron content while maintaining cell size and growth rates under iron limitation.

Cell biomass, as estimated from the total amount of protein per cell, was much lower in RCC802 than in other strains. Compared to RCC809 and OTTH595, RCC802, showed a marked decrease of the Chl cell content, which was associated with a decrease of the oxygen net production rate per cell under iron limitation. When the oxygen net production was normalized by the Chl cell fluorescence (reflecting the Chl cell content), there was, however, no variation between iron depleted and iron replete conditions, demonstrating that the photosynthetic efficiency per Chl molecule was not affected. This is supported by the fact that the RCC802 photosystem II parameters did not significantly decrease upon iron limitation unlike in OTTH595 and RCC809, for which the growth rate decrease was associated to a pronounced drop of the photosystem II quantum yield and relative electron transport rate, strongly suggesting a global photosynthesis impairment (Fig. 2). This decrease likely originates from photosystem II photoinactivation (see e.g. ref. 47) and/or the induction of excess light energy dissipation mechanisms[23,26]. Moreover, iron depletion induced a decrease in the light harvesting capacities in these strains and a global decrease of the Chl cell content (Figs 2C and 4B). These observations indicate that OTTH595 and

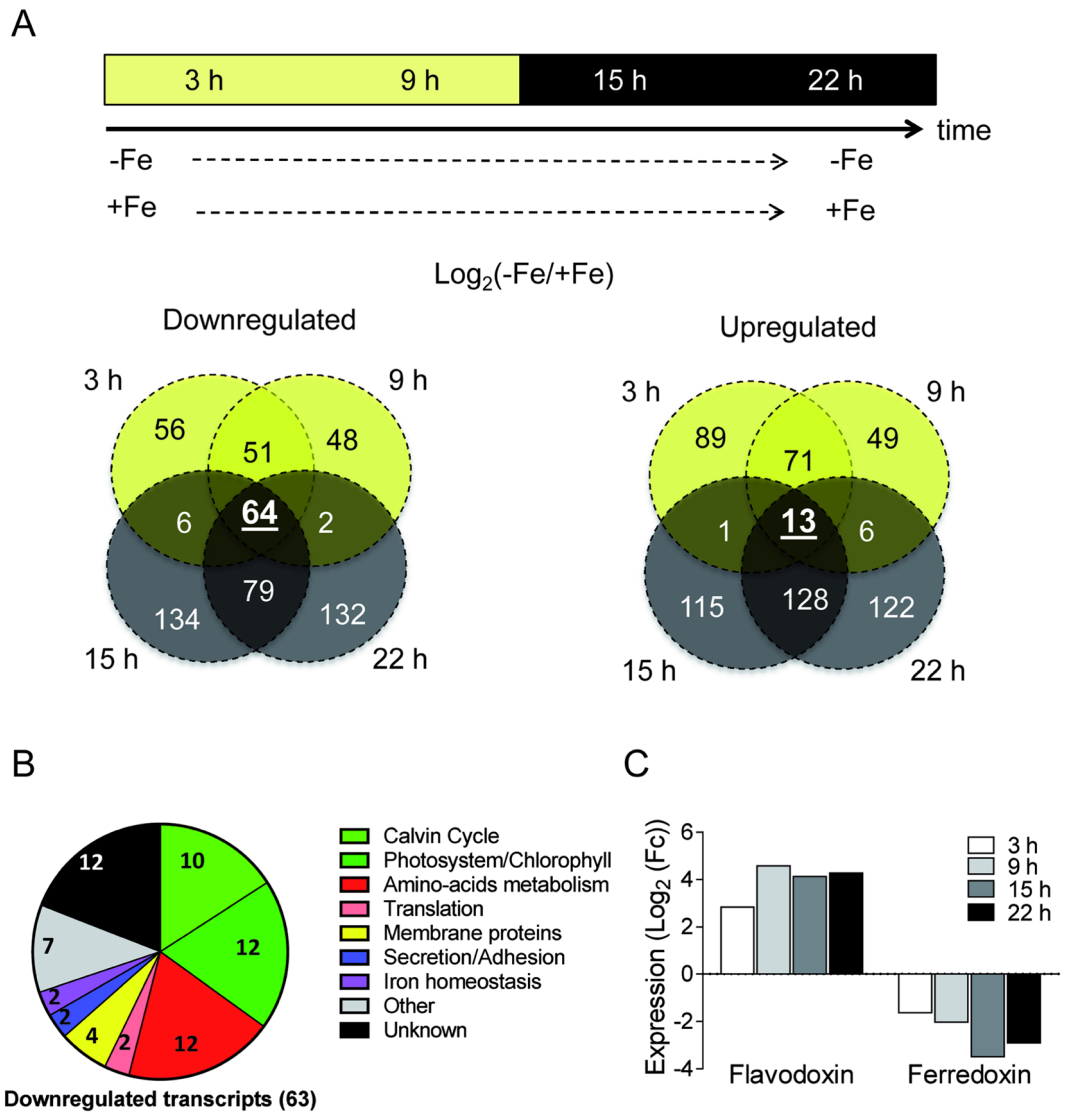


Figure 6. Transcriptomic response of RCC802. (**A**) RCC802 cells were grown for 5 days under 12 h light: 12 h dark cycles in +Fe (270 nM Fe(III)-EDTA) or −Fe (5.4 nM Fe(III)-EDTA) conditions. RNA samples extracted at 3 h, 9 h, 15 h and 22 h after dawn, were subjected to RNAseq analysis. Out of 16527 de novo assembled RNAs, 1251 were differentially expressed between iron depleted and replete conditions in at least one day time point ($-1 < \mathrm{Log_2}(-\mathrm{Fe}/+\mathrm{Fe}) <$ or $\mathrm{Log_2}(-\mathrm{Fe}/+\mathrm{Fe}) > 1$). Venn diagrams of differentially expressed genes. Day is represented in yellow, night in dark grey. (**B**) Functional clustering of the 64 genes upregulated in iron limiting conditions. Genes encoding photosynthesis (green) and protein synthesis proteins (red) represent more than 50% of downregulated genes. (**C**) At all day times a flavodoxin was induced while the ferredoxin was repressed.

RCC809 responded to iron limitation by drastically decreasing the amount of light energy utilized, which resulted in a strong decrease of the growth rate compared to RCC802 (Fig. 1). The downregulation of photosynthesis is probably the consequence of the high requirement for Fe of the photosynthetic apparatus, as characterized in other photosynthetic organisms such as diatoms[44].

Altogether, our results suggest that, in response to iron limitation, RCC802 decreases the quantity of photosynthetic complexes per cell, *i.e.* the 'volume' of the photosynthetic machinery while maintaining its efficiency. In agreement with this hypothesis, RNAseq analysis, showed that in response to iron limitation, about one third of down regulated transcripts during all day time encode proteins involved in various aspects of photosynthesis including Chl synthesis, photosystem composition and Calvin-Benson cycle. The overall reduction in protein content is also reflected at the transcriptional level by the repression of genes involved in protein synthesis. In the low iron ecotype *Thalassiosira oceanica*, a similar reduction of the cellular protein content was probably due to a smaller number of chloroplasts per cell[44]. *Ostreococcus* species, unlike diatoms, contain a single chloroplast per cell, which occupies about 50% of the cell volume. The reduction of protein and chlorophyll content per cell in RCC802 is therefore likely to reflect a global reduction of the photosynthetic machinery and of the chloroplast. When comparing the expression patterns of the 64 RCC802 downregulated transcripts with putative homologues in OTTH595, we found that most genes were not downregulated in this latter species (Supplementary Table 3).

This observation is consistent with the fact that OTTH595 does not reduce as much its cellular biomass under iron limitation.

Adaptations relying on the regulation of iron binding proteins are also suggested by the transcriptomic analyses. For example, an RCC802 flavodoxin, which had no homologue in OTTH595 or RCC809, is upregulated under iron limitation, while ferredoxin is downregulated (Fig. 6) suggesting that in RCC802 like in oceanic diatoms, the substitution of ferredoxin by flavodoxin may contribute to the ecological success under chronically low iron environments[5–7]. The single *BHLH* gene is the only transcription factor of RCC802 to be constitutively induced under iron limitation. BHLH, in contrast, was down regulated in OTTH595 under iron limitation. BHLH may, therefore, be a transcriptional regulator involved in the efficient acclimation of RCC802 to low iron environments, like the*BHLH Fer* gene of higher plants[48].

Conclusions

Our results establish the existence of an *Ostreococcus* "low iron requiring strain", which acclimates efficiently to low iron conditions. They support the field studies suggesting that picoeukaryotes should not be seen only as components of mesotrophic areas[27] and suggest that iron may drive the differentiation of physiologically specialized strains along coast-ocean gradients. The main acclimation to low iron environment by *Ostreococcus* sp. RCC802 appears to involve primarily a reduction of cell biomass, rather than the reduction of cell surface/volume ratio reported in nano and microphytoplankton.

References

1. Saito, M. A., Sigman, D. M. & Morel, F. M. M. The bioinorganic chemistry of the ancient ocean: The co-evolution of cyanobacterial metal requirements and biogeochemical cycles at the Archean-Proterozoic boundary? *Inorganica Chim. Acta* **356**, 308–318 (2003).
2. Morrissey, J. & Bowler, C. Iron Utilization in Marine Cyanobacteria and Eukaryotic Algae. *Front. Microbiol.* **3**, 43 (2012).
3. Sunda, W. G. & Huntsman, S. A. Interrelated influence of iron, light and cell size on marine phytoplankton growth. *Nature* **2051**, 1193–1197 (1997).
4. Lis, H., Shaked, Y., Kranzler, C., Keren, N. & Morel, F. M. M. Iron bioavailability to phytoplankton: an empirical approach. *Isme J.* **9**, 1003–1013 (2015).
5. Strzepek, R. F. & Harrison, P. J. Photosynthetic architecture differs in coastal and oceanic diatoms. *Nature* **431**, 689–692 (2004).
6. Peers, G. & Price, N. M. Copper-containing plastocyanin used for electron transport by an oceanic diatom. *Nature* **441**, 341–4 (2006).
7. Lommer, M. *et al.* Recent transfer of an iron-regulated gene from the plastid to the nuclear genome in an oceanic diatom adapted to chronic iron limitation. *BMC Genomics* **11**, 718 (2010).
8. Marchetti, A. *et al.* Ferritin is used for iron storage in bloom-forming marine pennate diatoms. *Nature* **457**, 467–470 (2009).
9. Scanlan, D. J. Physiological diversity and niche adaptation in marine Synechococcus. *Adv. Microb. Physiol.* **47**, 1–64 (2003).
10. Martiny, A. C., Huang, Y. & Li, W. Occurrence of phosphate acquisition genes in Prochlorococcus cells from different ocean regions. *Environ. Microbiol.* **11**, 1340–1347 (2009).
11. Venter, J. C., Rusch, D. B., Martiny, A. C., Dupont, C. L. & Halpern, A. L. Characterization of Prochlorococcus clades from iron-depleted oceanic regions. *Proc. Natl. Acad. Sci.* **107**, 16184–16189 (2010).
12. Huang, S. *et al.* Novel lineages of Prochlorococcus and Synechococcus in the global oceans. *ISME J.* **6**, 285–97 (2012).
13. Biller, S. J., Berube, P. M., Lindell, D. & Chisholm, S. W. Prochlorococcus: the structure and function of collective diversity. *Nat Rev Micro* **13**, 13–27 (2015).
14. Dufresne, A. *et al.* Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome. *Proc. Natl. Acad. Sci. USA* **100**, 10020–10025 (2003).
15. Palenik, B. *et al.* The genome of a motile marine Synechococcus. *Nature* **424**, 1037–1042 (2003).
16. Mackey, K. R. M. *et al.* Divergent responses of Atlantic coastal and oceanic Synechococcus to iron limitation. *Proc. Natl. Acad. Sci.* **112**, 9944–9949 (2015).
17. Botebol, H. *et al.* Central role for ferritin in the day/night regulation of iron homeostasis in marine phytoplankton. *Proc. Natl. Acad. Sci.* **112**, 14652–14657 (2015).
18. Lelandais, G. *et al.* Ostreococcus tauri is a new model green alga for studying iron metabolism in eukaryotic phytoplankton. *BMC Genomics* **17**, 1–23 (2016).
19. Vaquer, A., Troussellier, M., Courties, C. & Bibent, B. Standing stock and dynamics of picophytoplankton in the Thau Lagoon (northwest Mediterranean coast). *Limnol. Oceanogr.* **41**, 1821–1828 (1996).
20. Bec, B., Husseini-Ratrema, Collos, Y., Souchu, P. & Vaquer, A. Phytoplankton seasonal dynamics in a Mediterranean coastal lagoon: emphasis on the picoeukaryote community. *J. Plankton Res.* **27**, 881–894 (2005).
21. O'Kelly, C. J., Sieracki, M. E., Thier, E. C. & Hobson, I. C. A transient bloom of Ostreococcus (Chlorophyta, Prasinophyceae) in West Neck Bay, Long Island, New York. *J. Phycol.* **39**, 850–854 (2003).
22. 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).
23. Six, C., Finkel, Z., Rodríguez, F. & Marie, D. Contrasting photoacclimation costs in ecotypes of the marine eukaryotic picoplankter Ostreococcus. *Limnol. Oceanogr.* **53**, 255–265 (2008).
24. Rodríguez, F. *et al.* Ecotype diversity in the marine picoeukaryote Ostreococcus (Chlorophyta, Prasinophyceae). *Environ. Microbiol.* **7**, 853–9 (2005).
25. 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**, 643–59 (2013).
26. Six, C., Sherrard, R., Lionard, M., Roy, S. & Campbell, D. A. Photosystem II and pigment dynamics among ecotypes of the green alga Ostreococcus. *Plant Physiol.* **151**, 379–390 (2009).
27. Demir-Hilton, E. *et al.* Global distribution patterns of distinct clades of the photosynthetic picoeukaryote Ostreococcus. *ISME J.* **5**, 1095–107 (2011).
28. Price, N. M. *et al.* Preparation and Chemistry of the Artificial Algal Culture Medium Aquil. *Biol. Oceanogr.* **6**, 443–461 (1989).
29. Sutak, R. *et al.* A comparative study of iron uptake mechanisms in marine microalgae: iron binding at the cell surface is a critical step. *Plant Physiol.* **160**, 2271–2284 (2012).
30. Botebol, H. *et al.* Different iron sources to study the physiology and biochemistry of iron metabolism in marine micro-algae. *Biometals* **27**, 75–88 (2014).
31. Smith, P., Krohn, R. & Hermanson, G. Measurement of protein using bicinchoninic acid. *Anal. Biochem.* **150**, 76–85 (1985).
32. Moulager, M. *et al.* Light-dependent regulation of cell division in Ostreococcus: evidence for a major transcriptional input. *Plant Physiol.* **144**, 1360–1369 (2007).

33. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–52 (2011).
34. 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).
35. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
36. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
37. Camacho, C. *et al.* BLAST command line applications user manual. *BLAST® Help*. *Bethesda*, *MD Natl. Cent. Biotechnol. Inf.* (2008).
38. Courties, C., Vaquer, A. & Troussellier, M. Smallest eukaryotic organism. *Nature* **370**, 255 (1994).
39. Claustre, H. *et al.* Is desert dust making oligotrophic waters greener? *Geophys. Res. Lett.* **29(107)**, 1–4 (2002).
40. Partensky, F., Blanchot, J., Lantoine, F., Neveux, J. & Marie, D. Vertical structure of picophytoplankton at different trophic sites of the tropical northeastern Atlantic Ocean. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **43**, 1191–1213 (1996).
41. Guieu, C. *et al.* Impact of high Saharan dust inputs on dissolved iron concentrations in the Mediterranean Sea. *Geophys. Res. Lett.* **29**, 2–5 (2002).
42. Mills, M. M., Ridame, C., Davey, M., La Roche, J. & Geider, R. J. Iron and phosphorus co-limit nitrogen fixation in the eastern tropical North Atlantic. *Nature* **429**, 292–294 (2004).
43. Monnier, A. *et al.* Orchestrated transcription of biological processes in the marine picoeukaryote Ostreococcus exposed to light/dark cycles. *BMC Genomics* **11**, 192 (2010).
44. Lommer, M. *et al.* Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome Biol.* **13**, R66 (2012).
45. Nunn, B. L. *et al.* Diatom proteomics reveals unique acclimation strategies to mitigate fe limitation. *PLoS One* **8**, e75653 (2013).
46. Gan, L., Ladinsky, M. S. & Jensen, G. J. Organization of the smallest eukaryotic spindle. *Curr. Biol.* **21**, 1578–1583 (2011).
47. Murata, N., Takahashi, S., Nishiyama, Y. & Allakhverdiev, S. I. Photoinhibition of photosystem II under environmental stress. *Biochim. Biophys. Acta (BBA)-Bioenergetics* **1767**, 414–421 (2007).
48. Ling, H.-Q., Bauer, P., Bereczky, Z., Keller, B. & Ganal, M. The tomato fer gene encoding a bHLH protein controls iron-uptake responses in roots. *Proc. Natl. Acad. Sci. USA* **99**, 13938–13943 (2002).
49. 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).
50. Sunda, W. G. & Huntsman, S. A. Iron uptake and growth limitation in oceanic and coastal phytoplankton. *Mar. Chem.* **50**, 189–206 (1995).

Acknowledgements

Funding is acknowledged from ANR "PhytoIron" (ANR 11BSV7 018 02) to F.Y.B. and E.L. H.B. was supported by a fellowship from the Institut National des Sciences de l'Univers (INSU CNRS). We thank Audrey Gueuneugues for help with preparing Aquil Medium.

Author Contributions

S.B. and F.Y.B. designed the experiments, H.B., E.L., C.S., J.C.L., P.S. and V.V. performed the experiments. G.L., A.M. L.B., S.L. performed RNAseq analysis. Daniel Vaulot kindly provided RCC789, RCC802 and RCC809 strains. All authors contributed to writing.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-00216-6

Competing Interests: The authors declare that they have no competing interests.

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

D *Curriculum vitae*

Arnaud Meng

Date of Birth:	27 January 1989
Address:	3 rue Paul Bert, 77400, Lagny-sur-Marne, France
Phone:	+33 6 89 09 80 96
email:	arnaud.meng@gmail.com

Work Experience

Current
Oct 2014

Ph.D. Student in Bioinformatics
University of Pierre and Marie Curie (Paris VI) | Paris, France
Team High-Throughput Sequencing Data Analysis in Genomics at "Institut de Biologie Paris-Seine"
Supervized by Stéphane Le Crom, Fabrice Not and Lucie Bittner
"Study of symbiosis in marine plankton by a transcriptome and meta-transcriptome approach"

My researches are based on *de novo* transcriptome assembly of non-model organisms using data from **high-throughput sequencing technologies**. I therefore contributes to the **establishment of bioinformatics tools** and of a friendly-user pipeline for transcriptome assembly and analysis, in order to explore the genomic data of organisms for which our knowledge is limited in the context of symbiosis conditions. (https://github.com/arnaudmeng)

Sep 2014
Mar 2014
(6 months)

Master Degree internship in Bioinformatics
University of Pierre and Marie Curie (Paris VI) | Paris, France
Team High-Throughput Sequencing Data Analysis in Genomics at "Institut de Biologie Paris-Seine"
Supervized by Stéphane Le Crom and Lucie Bittner
"Implementation of a de novo assembly pipeline for transcriptomic data"

My objective was to evaluate and choose the most suitable *de novo* assembler on marine plankton transcriptomic data and **to implement a Python pipeline** to automate assembly and analysis processes.

Aug 2013
Mar 2013
(5 months)

Master Degree internship in Bioinformatics
University of Evry-Val-d'Essone | Evry, France
Team Analysis and Modeling for Biology and Environment
Supervized by Nathalie Basdevant
"Coarse-grained modeling of protein-protein interactions"

My objective was to perform a benchmark on two coarse-grained force fields and to automate the protein-protein docking process with **the establishment of a Python pipeline**.

Education

2012-2014 **Master of Science** in Bioinformatics (1/12, with honors)
Paris Diderot University (Paris VII) | Paris, France

2008-2012 **Bachelor of Science** in Bioinformatics
Paris Diderot University (Paris VII) | Paris, France

2007-2008 **DUT** in Electrical engineering and industrial IT
Conservatoire National des Arts et Métiers (C.N.A.M.) | La Plaine Saint-Denis, France

2007 **Baccalaureate** in Science, life science option
Pierre de Coubertin | Meaux, France

Skills

Computing skills

Programming languages: **Python**, **Perl**, **bash**, **R**, HTML, C, JavaScript, BioPerl, BioPython
Bioinformatics Tools: NCBI toolkit (BLAST), *de novo* assembly (Trinity, Velvet/Oases), Phylogeny (MAFFT, Gblocks, RAxML), Functional analysis (InterproScan 5, Transdecoder), Network analysis and statistics (igraph R package)
Operating systems: Linux, MacOS X, Windows
System administration: DELL server under Ubuntu 14.04

Mentoring

Jul 2016 – Mar 2016 (5 months) **Supervision of Master intern** in Bioinformatics
Student : Quentin Letourneur, Bioinformatic Master Student (Paris VII)
"Optimizing transcript selection from de novo assembly pipeline considering gene expression dynamic"

Jul 2015 – Mar 2015 (5 months) **Supervision of Master intern** in Bioinformatics
Student : Anita Annamale, Bioinformatic Master Student (Paris VII)
"Implementation of Python module dedicated to RNA-seq data filtering"
https://github.com/upmcgenomics/PREMSEQ

Teaching

March 2017 **Introduction to programming and algorithmics** (PhD, master, permanent researchers)
4 hours of practical exercices using Python.
Organizers: Ingrid Lafontaine and Philippe Lopez
University of Pierre and Marie Curie (Paris VI) | Paris, France

Languages

French: Native
English: Fluent
Spanish: Basic

Grants

2016 | International fellowship grant - 500€ (French Society of Bioinformtics (SFBI)) to attend ECCB 2016 at The Hague, Netherlands

2014 | "Interface pour le vivant" 3-years Ph.D. grant - 90k€ (from the French minister of research in bioinformatics and evolution) at University of Pierre and Marie Curie

Publications

Key functions involved in the establishment and the maintenance of marine plankton symbiosis revealed by a meta-transcriptome approach.
A. Meng, E. Corre, C. Marchet, P. Peterlongo, A. Alberti, C. Da Silva, P. Wincker, I. Probert, N. Suzuki, S. Le Crom, L. Bittner and F. Not.
in prep.

A *de novo* approach to disentangle/decouple partners identity and function in holobiont systems.
A. Meng, C. Marchet, E. Corre, P. Peterlongo, A. Alberti, C. Da Silva, P. Wincker, E. Pelletier, I. Probert, J. Decelle, S. Le Crom, F. Not and L. Bittner.
in prep.

Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome-based sequence similarity network.
A. Meng, E. Corre, I. Probert, A. Gutierrez-Rodriguez, R. Siano, A. Annamale, A. Alberti, C. Da Silva, P. Wincker, S. Le Crom, F. Not and L. Bittner
under review at *Molecular Ecology*

Acclimation of a low iron adapted Ostreococcus strain to iron limitation through cell biomass lowering.
H. Botebol, G. Lelandais, C. Six, E. Lesuisse, **A. Meng**, L. Bittner, S. Lecrom, R. Sutak, J. Lozano, P. Schatt, V. Vergé, S. Blain and F.Y. Bouget.
In ***Scientific Reports***, 7(1):327, Mar 2017. **IF: 5.228**
5th/13 authors - PMID: 28336917

Oral communications

A highly scalable data structure for read similarity computation and its application to marine plankton holobionts
Camille Marchet†, A. Meng†, E. Corre, S. Le Crom, F. Not, L. Bittner and P. Peterlongo
Workshop RCAM: Recent Advances in Metagenomics, October 9-10 2017, Paris, France.

The use of sequence similarity network to explore functional diversity of dinoflagellates.
A. Meng, E. Corre, I. Probert, A. Gutierrez-Rodriguez, R. Siano, A. Annamale, A. Alberti, C. Da Silva, P. Wincker, S. Le Crom, F. Not and L. Bittner.
National conference in environmental genomics - Networks thematic day, June 2 2017, Nantes, France.

A transcriptomic approach to study marine plankton holobionts.
A. Meng, E. Corre, P. Peterlongo, C. Marchet, A. Alberti, C. Da Silva, P. Wincker, I. Probert, N. Suzuki, S. Le Crom, L. Bittner and F. Not.
International Conference on Holobiont, April 19-21 2017, Paris, France.

Posters

***De novo* transcriptome assembly dedicated pipeline and its specific application to non-model marine planktonic organisms**.
A. Meng, L. Bittner, E. Corre, F. Not and S. Le Crom.
European Conference on Computational Biology 2016, September 3-7 2016, The Hague, Netherlands.
(Web link)

***De novo* assembly pipeline for transcriptomic analysis**.
A. Meng, L. Bittner, Anita Annamale, E. Corre, F. Not and S. Le Crom.
Journées Ouvertes de Biologie, Informatique et Mathématiques 2015, July 6-9 2015, Clermont-Ferrand, France.
(Web link)

A modular pipeline for *de novo* transcriptome assembly.
A. Meng, L. Jourden, L. Bittner and S. Le Crom.
European Student Council Symposium 2014, September 6 2014, Strasbourg, France.
(Web link)

MISCELLANEOUS

Trainings
Statistical analysis of networks, December 1-3 2017, CNRS, Paris, France.
Summer School in Metagenomics, September 12-16 2017, Pasteur Institute, Paris, France

Organizational activities
"Journée Des Doctorants de l'unité Evolution", March 3 2017, University of Pierre et Marie Curie, Paris, France.
"Journée Des Doctorants de l'unité Evolution", February 18 2016, University of Pierre et Marie Curie, Paris, France

Affiliations
Member of the French Society of Bioinformtics (SFBI)

Résumé

Les relations symbiotiques entre organismes sont essentielles pour l'évolution de la biodiversité et le fonctionnement des écosystèmes. En milieu terrestre (*i.e.* mycorhize) ou en milieu marin benthique (*i.e.* récifs coralliens) les symbioses sont assez bien décrites. Si dans le plancton marin, les relations entre hôtes hétérotrophes et symbiontes photosynthétiques (photosymbioses) sont des phénomènes observés fréquemment dès le 19ème siècle, les mécanismes fonctionnels qui régissent ces symbioses restent largement inconnus. C'est notamment le cas de l'association symbiotique entre certaines espèces de radiolaires et leurs photosymbiontes dinoflagellés. Il s'agit là du modèle symbiotique, composé de deux unicellulaires eucaryotes, sur lequel je me suis concentré au cours de ce travail de thèse. Ces deux organismes sont connus pour être largement répandus dans les océans ainsi que pour leur importance au sein des écosystèmes marins, et il est donc important de mieux caractériser ces événements de symbiose afin d'approfondir nos connaissances de ces organismes. Grâce aux technologies de séquençage haut-débit il est désormais possible d'obtenir, pour ces organismes unicellulaires non cultivables mais isolés depuis l'environnement, une quantité sans précédent d'information génomique. Ces approches représentent donc une opportunité unique de décrypter finement les mécanismes à l'oeuvre dans ces interactions symbiotiques entre unicellulaires. Mon travail de thèse a combiné la mise en place de protocoles et d'outils bioinformatiques dédiés à l'assemblage et à l'analyse de jeux de données de transcriptomique sans référence génomique des holobiontes (couple hôte-symbiontes) de radiolaires et dinoflagellés. Outre le développement d'outils de bioinformatique innovants pour l'étude d'organismes non-modèles et non cultivables, ce travail de doctorat contribue à une meilleure compréhension des mécanismes d'adaptation fonctionnelle et évolutive des organismes photosymbiotiques marins.

Abstract

Symbiotic associations between organisms are essentials in biodiversity evolution and ecosystems functioning. In terrestrial environments (*e.g.* mycorrhiza) or in the benthic marine environment (*e.g.* coral reefs), the symbioses encountered are fairly well described and studied. In the marine plankton, photosymbioses (associations between heterotrophic hosts and photosynthetic symbionts) are phenomena described and observed frequently since the 19th century. However, if the actors of these associations begin to be identified, the fundamental functional mechanisms for the establishment and the maintenance of these symbioses remain largely unknown. This is particularly true for the symbiotic association between symbiotic radiolarians and their dinoflagellate photosymbionts, two unicellular eucaryotes, which I was interested in during this thesis. These two organisms are known to be widespread in the oceans and for their key role in marine ecosystems, and it is therefore important to better characterize these symbiotic events in order to deepen our knowledge of these organisms. Thanks to high-throughput sequencing technologies (RNAseq type) it is now possible to obtain an unprecedented amount of transcripts for these unicellular organisms that are not cultivable and need to be directly isolated from the environment. These new technologies thus represent a unique opportunity to better characterized the mechanisms involved in these intimate cellular interactions. My Ph.D. work has combined the implementation of bioinformatics protocols and tools dedicated to the assembly and analysis of RNA-seq data without genomic reference as well as to the study of holobiont transcriptomes (the host-symbionts pair) of radiolarians and dinoflagellates. In addition to the development of innovative bioinformatics tools for the study of non-model and non-culturable organisms, this thesis contributes to a better understanding of the mechanisms of functional and evolutionary adaptation of marine photosymbiotic organisms.