On the use of networks to link molecules to ecosystems: Investigating the uncultured lineages and/or the unknown functions from protists for a better understanding of ecosystems

Lucie Bittner

Habilitation à Diriger des Recherches

Faculté de Biologie - UFR 927 des Sciences de la Vie

Lucie Bittner

Associate Professor at Sorbonne Université

Evolution Paris Seine UMR7138, Institut de Biologie Paris Seine

On the use of networks to link molecules to ecosystems

Investigating the uncultured lineages and / or the unknown functions from protists for a better understanding of ecosystems

December 21st, 2018, in Paris

Rapporteurs: Dr Denis Faure, Pr Sébastien Monchy, Dr Naiara Rodríguez-Ezpeleta

Jury:

M. Didier Debroas	PR Université Clermont Auvergne
M. Denis Faure	DR CNRS Paris-Saclay
M. Sébastien Monchy	PR Université du Littoral Côte d'Opale
M. Hervé Moreau	DR CNRS Sorbonne Université, Banyuls-sur-Mer
Mme. Emma Rochelle-Newall	DR IRD Sorbonne Université, Paris

Abstract / Résumé

Our understanding of the microbial world is living a true paradigm shift. The advent of next generation sequencing in molecular biology, offers, via metagenomics and metatranscriptomics, unprecedent measures of both taxonomic and functional biodiversity of hitherto hidden worlds. However, as current inference tools mostly rely on species names or function names, a significant part of the available meta-omic sequences are currently ignored. Improving culture methods and increasing the number of model organisms help to reduce the proportion of unknown sequences, but these solutions remain expensive and time-consuming. At the present time, bioinformatics methods exist that can exploit the massive amount of known and unknown sequences, thus allowing to go beyond our still uncomplete view of the microbial communities. In this memoire, I expose bionformatics methods that are currently used in my team in order to mine the microbial (meta-)omic *dark matter*. I notably present the work I have recently supervised based on the use of sequence similarity networks and co-oocurence networks. These networks enable to study without *a priori* the adaptative and evolutionary processes shaping the taxonomical and functional diversity of protists in the environment.

Notre compréhension du monde microbien vit un changement de paradigme. La révolution technologique du séquençage dit haut-débit en biologie moléculaire, offre, via la métagénomique et de la métatranscriptomique, des mesures sans précédent de la biodiversité taxonomique et fonctionnelle des communautés microbiennes. Cependant, comme les outils d'inférence actuels reposent principalement sur des noms d'espèces ou des noms de fonctions, une partie importante des séquences (méta-)omiques actuellement disponibles est ignorée. Les progrès des méthodes de culture et l'augmentation du nombre d'organismes modèles aident à réduire la proportion de séquences inconnues, mais ces solutions restent coûteuses et prennent beaucoup de temps. À l'heure actuelle, il existe des méthodes bioinformatiques capables d'exploiter l'énorme quantité de séquences connues et inconnues, et qui nous permettent ainsi de dépasser notre vision encore incomplète des communautés microbiennes. Dans ce mémoire, j'expose les méthodes bionformatiques actuellement utilisées dans mon équipe afin d'explorer la « matière noire » omique microbienne. Je présente notamment les travaux que j'ai récemment encadrés et qui utilisent des réseaux de similarité de séquences et des réseaux de co-occurrence. Ces réseaux nous permettent d'étudier sans *a priori* les processus adaptatifs et évolutifs qui façonnent la diversité taxonomique et fonctionnelle des protistes dans l'environnement.

Contents

List of figures

1. Foreword - Why and how I became a (demanding) environmental genomicist

> *''It is of great use to the sailor to know the length of his line, though he cannot with it fathom all the depths of the ocean. It is well he knows that it is long enough to reach the bottom at such places as are necessary to direct his voyage, and caution him against running upon shoals that may ruin him.''*
> Essay Concerning Human Understanding (1689), John Locke
> Locke was writing about understanding the tools of human thought, but it is every bit as sensible to understand the limits of those tools that contributed to the creation of a body of data as well, because these affect our interpretation every bit as acutely. (Keeling & del Campo, 2017)

> *"scio me nihil scire"* statement attributed to the greek philosopher Socrate
> … but doubt pushes me forward !

This manuscript is obviously an excellent occasion to "force me" to take few minutes to sit and to have a look at the past (since my PhD defense already nine years ago), to make a summary of my present research, and to clarify my goals for the coming years.

Since I was a child, I am attracted by Natural sciences and by the exploration of our surrounding living world. For my higher education, I looked towards Biology and, quickly and clearly, nothing made sense to me except in the light of Evolution. I made a master in Systematics and Evolution at the University Pierre et Marie Curie, in Paris, in which I definitely enjoyed learning about phylogeny, genetics, molecular ecology, and diversity of eukaryotic organisms. At this time, I discovered Paris, its national history museum, and I also travelled for the first time to Brittany for a 2 weeks course hosted at the Marine Biological Station from Roscoff. Sampling marine macro-organisms and discovering the beautiful tidal variations, under the eye of a group of kind and extremely pedagogic professors, showed me the way for my future. A few months later, I obtained a PhD bursary and I was able to study for the next three years the evolutionary biology and molecular diversity of a red macro-algae lineage. Thanks to my supervisors, in the first months of my PhD, I took part to an exotic sampling in the Vanuatu and in New Caledonia to collect new specimens, which then kept me busy in Paris to get their DNA sequences and to perform various analyses. My Phd thus consisted in building and resolving phylogenetic trees and testing methods to delineate species. I rapidly realize the

importance of developing pragmatic and pluralistic approaches to reveal the processes leading to organisms' environmental diversity. I definitely enjoyed working with these non-model macro-organisms and with (relatively for this period, i.e. 2006-2009) big molecular datasets. In the second part of my third and last PhD year, my participation to the Marine Ecological and Evolutionary Genomics summer course in Roscoff in May 2009, initiated me to metagenomics for the first time. It was obviously a decisive event in my career. During this course, Dr. Chris Bowler presented the Tara Oceans expedition which was about to start 4 months later. I may seem presumptuous, but I knew at this exact moment that this is exactly where I wanted to be for the next coming years: working with a growing amount of datasets, with poorly/unknown and molecularly complex organisms (i.e. unicellular eukaryotes), tackling diversity, adaptation and evolutionary questions from these microorganisms in their environment, exchanging and building a common language with scientists from different fields. The same week, I discussed with Dr. Colomban de Vargas, and fortunately for me, they were looking for motivated post-docs.

During my post-docs, from 2010 to 2013, I expanded the scope of my research to microbial eukaryotic lineages, studying their environmental diversity using large high-throughput sequencing datasets. I jumped into the metagenomics and I notably developed skills in microbes field sampling and also in bioinformatics. These post-docs years were extremely stimulating and rich in connexions. Since 2010, I work in the international Tara Oceans consortium (focusing on open-ocean planktonic communities) and I took part to the european BioMarKs consortium (focusing on coastal planktonic and benthic communities), which has enabled me to create a strong interdisciplinary network of collaborators, specialized in microbial diversity and ecology, functional genomics, bioinformatics, oceanography and notably with experts from the biogeochemists and modelers. These consortia offered me 3 years and a half of post-doc contracts in excellent conditions, joining consecutively 3 teams in different institutes (cf. CV).

In March 2013, during the beginning of my third post-doc, I wrote a project aiming at developing a protocol to study the evolutionary and functional adaptations of protists (unicellular eukaryotes) via the analyses of transcriptomes and metatranscriptomes. I submitted it and I was consequently recruited at Sorbonne Université (former Université Pierre et Marie Curie) as associate professor in the team High-Throughput Sequencing Data Analysis in Genomics from Professor Stéphane Le Crom. Since September 2013, I am teaching Bioinformatics, Evolution and Environmental Microbiology to bachelor and master students,

while conducting research on various questions. I must admit, that on one hand, I did not yet implement the integrality of my initial research project, principally because I was not successful in my grants application (e.g. ANR, ERC), but on the other hand my fruitful collaborations make me work on exciting questions, which also allow me to develop my own research interests and offer me the opportunity of (co-)supervising master and PhD students. My general line is, as long as the analyses are involving unicellular eukaryotic molecular data and are somehow related to evolution, adaptation and ecology, I am interested in playing with them !

Since my PhD, I have developed the idea that the majority of the studies from non-model organisms (e.g. non cultivable organisms, and/or organisms for which no or just few molecular data were produced, and/or for which life-cycle is poorly known), and in particular, from protists, are still biased by a metazoa-like/anthropocentric vision (del Campo et al., 2014; Sibbald and Archibald, 2017) (Figure 1). For instance, diversity studies are still currently majoritarily based on metabarcoding data, which might make sense for multicellular lineages, but which potentially largely underestimate the number of eukaryotic unicellular lineages because of their large effective population sizes (Piganeau et al., 2011), and most certainly skewed our vision about their functional and adaptation potentials. A further example, even if eukaryotic sex (i.e. cell fusion-making diploid or meiosis-producing haploid cells meiosis) is widespread and might be already present in the last eukaryotic common ancestor, the general mode of existence of protists is best described by clonally propagating cell lines with episodic sex triggered by external or internal clues (Speijer et al., 2015). A 'classical' biological species concept (sensu Mayr 1942 (De Queiroz, 2007)) may rarely be operational to define evolutionary or diversity units (Figure 2). In fact, I am deeply convinced that a metazoa-like/anthropocentric (Caron et al., 2009) as well as a prokaryotic vision (Keeling and Campo, 2017) might be both barriers to the study and understanding of protists adaptive and evolutionary processes (Figure 3).

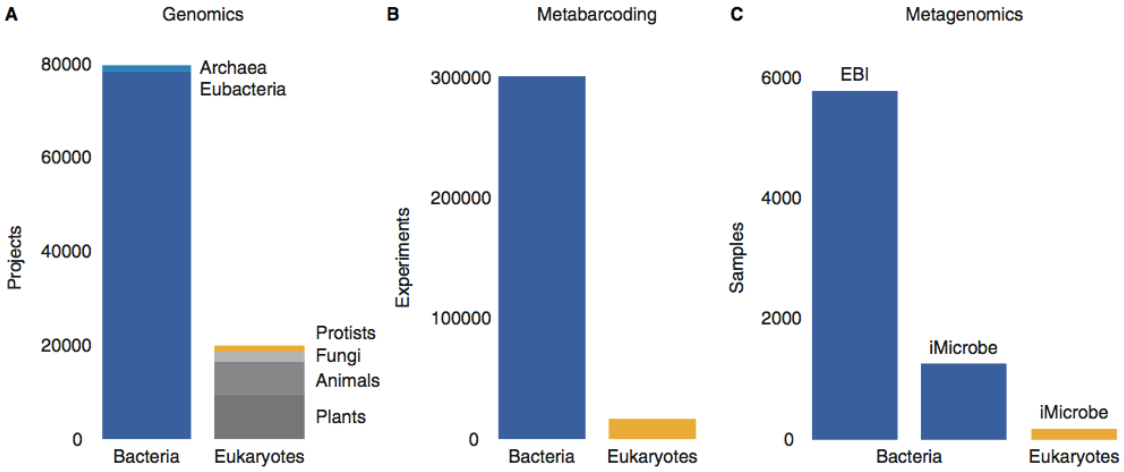


Figure 1 - Current microbial genomic data are heavily biased towards bacteria (Keeling and Campo, 2017) (A) The current number of completed genomic projects (retrieved from the Genomes OnLine Database (GOLD) in 2017 (Pagani et al., 2012)). (B) The current number of tag sequencing studies using 16S rRNA for bacteria as opposed to 18S rRNA for eukaryotes (retrieved from the Sequence Read Archive (SRA) (Leinonen et al., 2011)). (C) The current number of environmental metagenomics and metatranscriptomic studies in EBI Metagenomes database (Mitchell et al., 2016) and iMicrobe database (http://imicrobe.us/ (Hurwitz et al., 2017)).

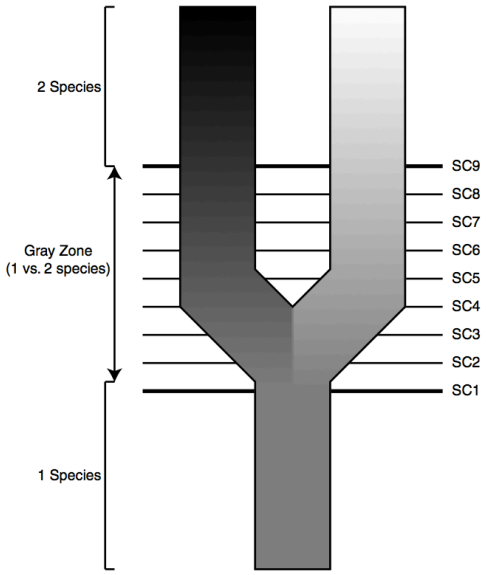


Figure 2 - A monist explanation of speciation processes (from De Queiroz, 2007).The process of metapopulation lineage divergence (speciation) is here illustrating. Progressive darkening and lightening of the daughter lineages represent their progressive divergence through time (bottom to top), and the numbered lines labeled SC (species criterion) 1– 8 represent the times at which the daughter lineages acquire different properties relative to one another (e.g., when they become phenetically distinguishable, diagnosable by a fixed character difference, recipro- cally monophyletic, reproductively incompatible, ecologically distinct, etc.). Before evolution of the first property (SC1), authors will agree there is a single species, and after evolution of the last property (SC8), they will agree there are two. Between these events, however, there will be disagreement among authors about whether one vs. two species are involved. The 'gray zone' is the conflicting zone. Disagreements result from authors adopting different contingent properties (species criteria) as the basis for their species definitions.

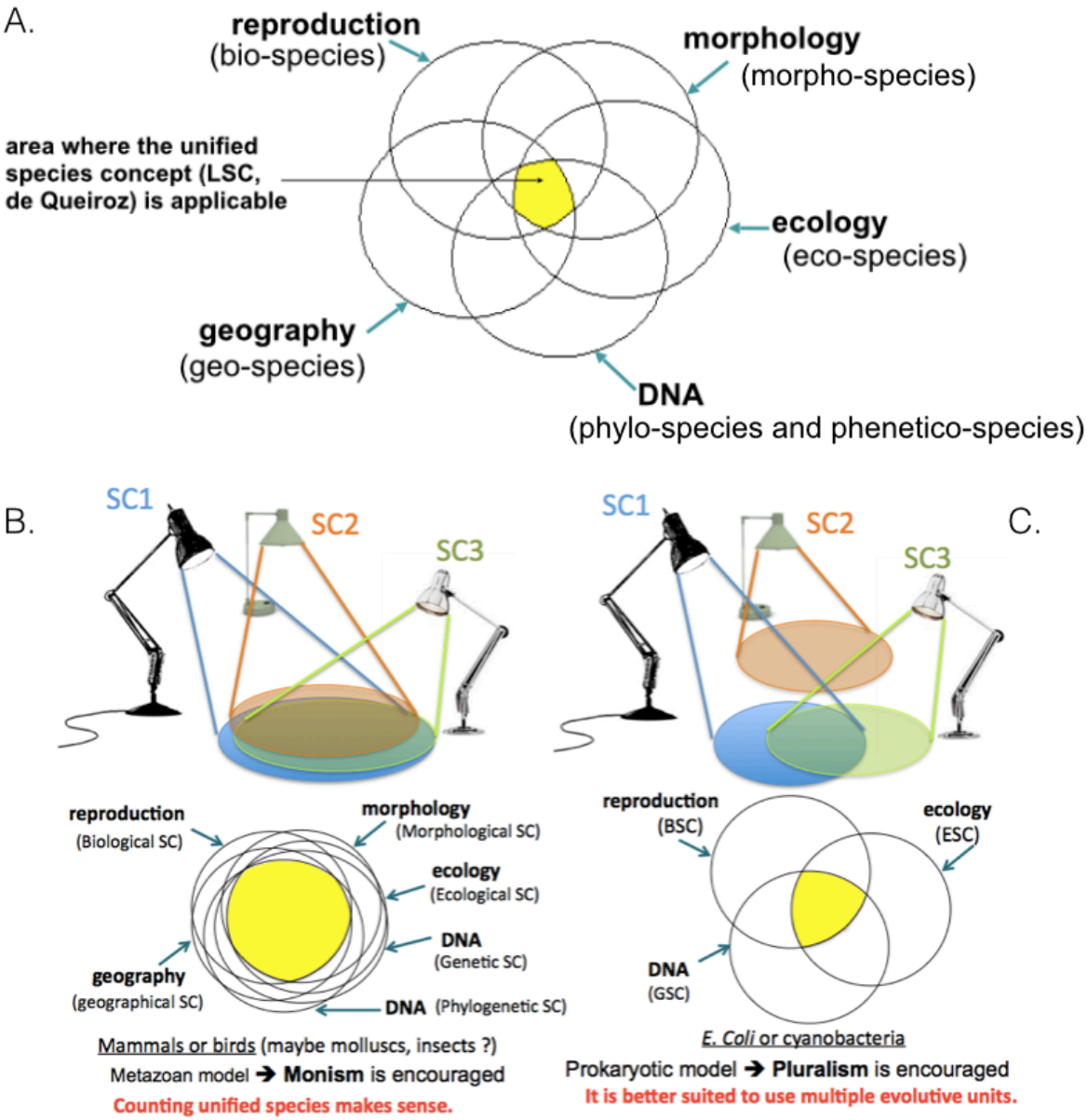


Figure 3 - Schemes of the area of application from the lineage species concept (LSC) proposed by (De Queiroz, 2007) (symbolized in yellow) (adapted from Bittner Lucie's PhD Manuscript and from (Bittner et al., 2010)). Figure SC means species concepts. (A) is a theoretical scheme. Circles symbolize the species that have been delimited using different data (in bold) and different species concepts (indicated into brackets). Other data and concepts could have been used (i.e. behavioral data, caryological data). DNA data can also be split in several circles (for instance one circle per marker). The size of the circle is also proportional to the amount of data available (size can vary between the different species concept and can change through time). The more the circles are overlapping, the less conflicts between species delimitation there are, the more applicable and . If circles are only overlapping on a small common area, the LSC is not pertinent. (B) and (C) correspond to application of the theoretical scheme to current studied lineages. My vision is that protists study framework can not be *a priori* defined, and that it should be addressed for each protistan lineage.

Due to the vast number of protists lineages, their various size range (appromatively between 0.8 µm to 10 nm), their diverse morphologies, their great deal of behaviours, trophic strategies and metabolisms, as well as the complexity and the potential huge size of their genomes (Caron et al., 2017, 2009; Carradec et al., 2018; Vargas et al., 2015), defining a single and simple model (i.e. either a Metazoan-like model or a Prokaryotic-like model) to formalise the framework of diversity and evolution studies from all protist lineages seems unrealistic. In my opinion some lineages might follow a more metazoan-like trend, some others a prokaryotic-like trend, and some others might mix characteristics from both with varying degrees. As datasets issued from cultures and experiments (e.g. (Keeling et al., 2014)) and from natural populations (e.g. (Carradec et al., 2018; Vargas et al., 2015)) are currently growing up, it is now the adequate time to study, without *a priori*, the adaptative and evolutionary processes shaping protistan lineages and functional diversity thanks notably to the mining of high-throughput molecular datasets. However, studying bigger datasets is a chance, but it does not mean there could not be biased. Implementing tests, adapting and building rigorous and reproducible analyses pipelines, favoring the cross-checking of information from various sources, as well not over-interpreting the results, remain the base to avoid many pitfalls.

In summary, since a decade, the common thread of my research lies in, not only using cutting-edge bioinformatic tools, but also developing new methodologies and concepts which can overcome the current (methodological and conceptual) pitfalls of classical methods used to study non-model and non-cultivable organisms. From a very broad point of view, I have tried since then, along with my collaborators (permanents and non-permanents, i.e. masters & PhDs) to investigate in an innovative way the two following questions:

- what is the diversity and how do eukaryotic microbes evolve and adapt in the environment? And how can we answer to these questions by investigating high-throughput omics datasets?
- as the majority of protist (and microbial in general) lineages are and will remain uncultured (Burki and Keeling, 2014; Keeling and Campo, 2017; Waller et al., 2018), can we develop in silico methods to analyze and make inferences about the natural populations while taking into account the massive quantity of taxonomical and/or functional unknown sequences?

2. From collaborations to supervisions

These past years, I have learned a lot from my collaborators. Dr. Fabrice Not convinced me that symbiosis was the most interesting (and beautiful) phenomenon to study for micro-organisms. Dr. Lionel Guidi initiated me to the biological carbon pump issue. Dr. Sakina-Dorothée Ayata enrolled me for linking omics and functional traits, towards population and biogeochemical modeling. Dr. Chris Bowler offered me to work in an original way several times with diatoms datasets. Dr. Stéphane Le Crom showed me how to conduct a research project in a realistic, reliable and professional way.

Since 2014, thanks to my collaborators, I co-supervised three PhDs (who all obtained grants from the French minister of research) and nine Master internships, who allowed me to develop several research questions (more details are given in the supervision section of the CV).

With Dr. Arnaud Meng (PhD grant 2014-2017, defense in December 2017; co-supervision with Dr. Fabrice Not and Dr. Stéphane Le Crom), we worked in revealing the genomic bases of symbiosis in marine planktonic protists. We developed a bioinformatic pipeline dedicated to *de novo* transcriptomic assembly of non model organisms and notably of holobionts (Meng et al., 2018b)°. We also developed an original analysis strategy through the use of sequence similarity networks, which enable to detect Open Reading Frames (ORFs) linked to symbiosis from un-annotated sequences of cultivated protists (Meng et al., 2018a)° in Molecular Ecology, or of protists isolated from the environment (Meng et al. in prep.).

Anne-Sophie Benoiston (PhD grant 2016-2019; co-supervision with Dr. Lionel Guidi and Dr. Stéphane Le Crom) is revisiting the Biological Carbon Pump issue in the global ocean by using an high-throughput and omics vision of the planktonic communities (Guidi et al., 2016)°. We are implementing co-occurrence networks calculations on metabarcoding and metagenomic datasets in order to detect the entire community of planktonic sequences (taxa, functions, and/or unknown sequences) as well as the main drivers of the net primary production, the carbon export and the remineralization processes. We are trying to integrate as much as possible non-annotated sequences in the analyses. In this way, un-annotated sequences can be linked to functions in the ecosystem and their abundance information give information about the importance and variability of these processes in the ocean.

Emile Faure (PhD grant 2017-2020; co-supervision with Dr. Sakina-Dorothée Ayata and Dr. Dominique Higuet) is investigating the impacts of planktonic diversity on oceanic biogeochemical cycles by integrating high-throughput sequencing data into marine ecosystems models. We focused recently on the environmental diversity and structure of mixotrophic protist lineages, which are thus far poorly or non involved in biogeochemical models. Our first analyses

were based on metabarcoding data, and will be in the coming month extended to metagenomics species (MGS or Co-Abundance gene Groups (CAGs) (Nielsen et al., 2014)) and to metagenome-assembled genomes (MAGs (Delmont et al., 2018; Parks et al., 2017)), which will enable to build a more exhaustive and accurate picture of mixotrophy in the global ocean, while also integrating taxonomical and/or functional unknown sequences.

Finally, Ophelie Da Silva who performed her master internships in the lab (spring 2017 and 2018; co-supervision with Dr. Sakina-Dorothée Ayata and Dr. Lionel Guidi), obtained in July 2018 a PhD bursary from the French minister of research. For the three coming years (October 2018-September 2021), she will be based in the Laboratoire Océanologique de Villefranche-sur-mer with Dr. Fabien Lombard and Dr. Lionel Guidi, and we will continue on collaborating to work on population (meta)genomics analysis from non-model planktonic organisms, and notably using sequences which are currently only found in the environment.

Exchanging and interacting with all these people makes my day every single working day. I am pleased to present and develop in the following memoire the results of our collaborative work.

N.b. During the last decade, I have explored many questions from phylogeny to ecology. In this manuscript, I choose to focus on a reduced number of research that I have made these last years, and I will here only develop the research in which I am involved as a supervisor. Throughout this manuscript, articles in which I have been involved as a collaborator are indicated with a bubble (°), and an exhaustive list of my publications can be found in the CV section.

3. State of the art (or brief introduction of my favorite bugs)

3.1. Meta-omics studies and their current caveats

Meta-omics or environmental genomic *sensu largo* correspond to the data and knowledge acquired on present or past organisms and ecosystems, by analyzing the sequences from a sample. The sequences (e.g. DNA, RNA, proteins) reveal the presence and / or the expression of organisms and / or functions for a given environment. Environmental genomic focuses mainly on microbial populations (i.e. organisms corresponding to prokaryotic and protistan lineages), and its evolution is intimately linked to technological advances (e.g. sampling methods, sequencing techniques). Therefore, a 'revolution' of the study of microbial communities has been taking place since the 2000s and is accelerating concurrently with the popularization in laboratories of high-throughput sequencing (HTS) or next-generation sequencing (NGS) techniques (**Figure 4**).

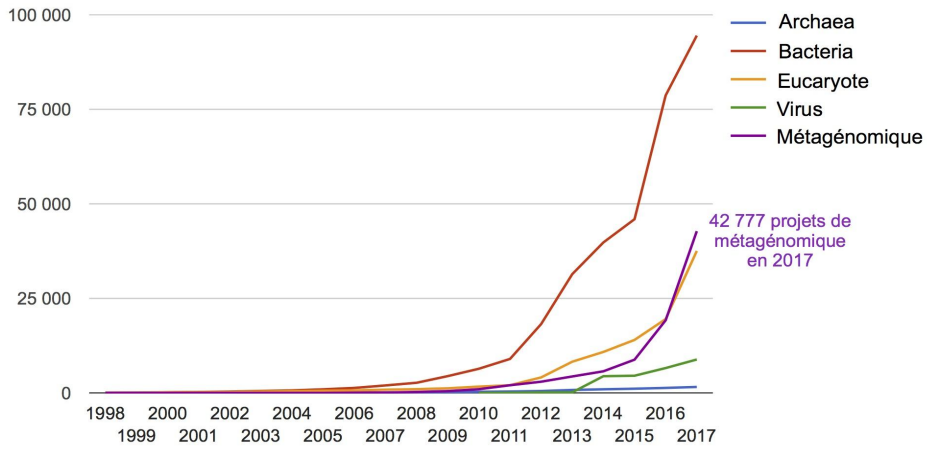


Figure 4 – Evolution of the number of sequencing projects from 1998 to 2017 (source JGI GOLD database (Pagani et al., 2012) https://gold.jgi.doe.gov/ - data from February 2018)

The first NGS/HTS methods appeared in 2005 (Reuter et al., 2015). These methods, from which Illumina technology is currently the most used (i.e. in 2018), can sequence millions of sequences simultaneously, reducing significantly the time required to obtain a complete genome. In 1996, 1 Mb (thousands or 10^6 bases) was sequenced per day, whereas in 2015, a sequencing experiment can read up to 500 Gb (billion or 10^9 bases) per day (Reuter et al., 2015). The methods are faster and their cost has also decreased (**Figure 5**): reading a single base costed 10 $ in 1985, whereas reading of 1 Mb costs 0.05 $ today (Pettersson et al.,

2009) (information from the National Human Genome Research Institute (NIH) on https://www.genome.gov/). In parallel, a paradigm shift is also observed in the cost of the different steps of a sequencing project across time (Figure 6).

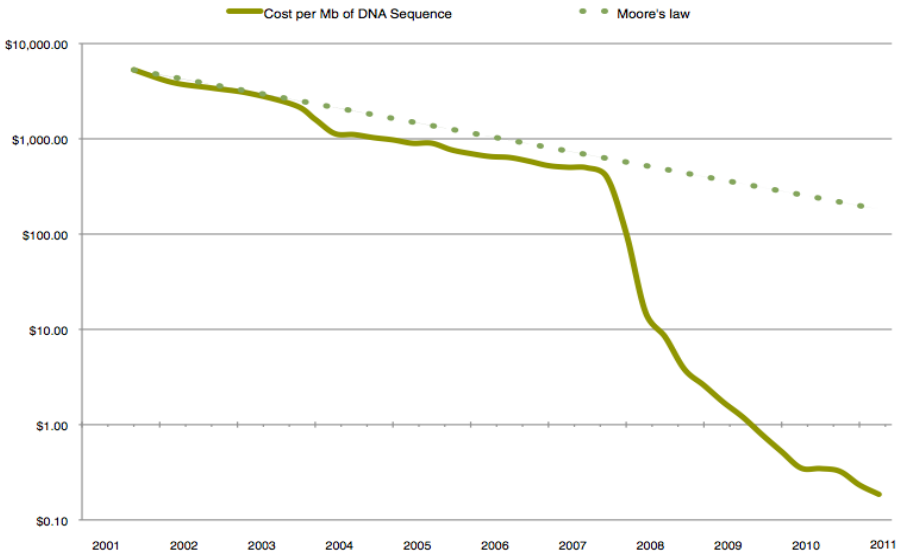


Figure 5 – Decreasing cost of DNA sequencing in the past 10 years compared with the expectation if it had followed Moore's law (Sboner et al., 2011)

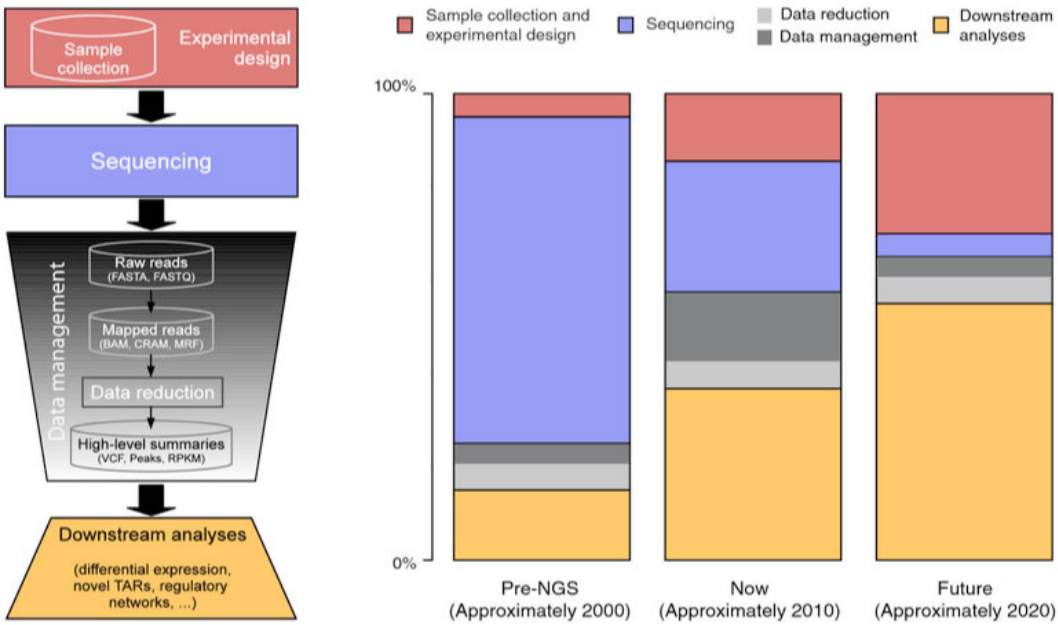


Figure 6 – Contribution of different factors to the overall cost of a sequencing project across time (Sboner et al., 2011). Left, the four-step process: (i) experimental design and sample collection, (ii) sequencing, (iii) data reduction and management, and (iv) downstream analysis. Right, the changes over time of relative impact of these four components of a sequencing experiment. (BAM, Binary Sequence Alignment/Map; BED, Browser Extensible Data; CRAM, compression algorithm; MRF, Mapped Read Format; NGS, next-generation sequencing; TAR, transcriptionally active region; VCF, Variant Call Format.)

This increasing productivity of sequencing methods coupled with bioinformatic analyzes has enabled large-scale projects to emerge. In genomics, one can notably mention the "1000 human genomes project" (initiated in 2008, http://www.internationalgenome.org/ (Siva, 2008; The 1000 Genomes Project Consortium, 2015)) the TCGA project (The Cancer Genome Atlas, initiated in 2005, https://cancergenome.nih.gov/ (Tomczak et al., 2015)), the "1000 Plant genomes project" (announced in 2008 and initiated in 2012, www.onekp.com (Matasci et al., 2014)), or the very recently Earth BioGenome Project (EBP, launched by the Sanger Institute on November the 1st, 2018, which aims to sequence, catalogue and categorise the genomes of all of Earth's eukaryotic biodiversity over a period of ten years, https://www.earthbiogenome.org/). These large initiatives, as well as individual or smaller genome sequencing projects, have led to the exponential creation of reference genomes since 2005 (**Figure 4**). Furthermore, HTS technologies represent the most evident solution to extend our knowledge to relatively unexplored and non currently / uncultivable microbial lineages, insofar as they can be punctually isolated from the rest of the community in which they live (Hug et al., 2016).

In the field of meta-genomics, large projects are also increasing over the years (Knight et al., 2012) (**Figure 4**): e.g. the HMP (Human Microbiome Project) / meta-Hit project (Metagenomic of Human intestinal tract, initiated in 2008, http://www.metahit.eu/ (Qin et al., 2010)), the *Tara* Oceans expedition as early as 2009 (https://www.embl.de/tara-oceans/ (Bork et al., 2015; Karsenti et al., 2011)), the Ocean Sampling Day (OSD (Kopf et al., 2015)), the EMP and EMP500 project (Earth Microbiome Project launched in 2010 http://www.earthmicrobiome.org/ (Gilbert et al., 2010; Thompson et al., 2017)), the MetaSUB project (The Metagenomics and Metadesign of the Subways and Urban Biomes, http://metasub.org/, initiated in 2015, (The MetaSUB International Consortium, 2016)). These projects often starts with metabarcoding approach (based on amplicons or on miTAGs (Logares et al., 2014b)), and continue with metagenomics and metatranscriptomics analyses. These projects also focus generally on prokaryotic microbes which size is inferior to 3 µm (i.e. size fractions dominated by Eubacteria and Archaea), and rarely on eukaryotic microbes / protists (which size is predominantly superior to 3 µm). Regarding the ocean microbiome, in 2015, a first article based on the analysis of 68 stations and 246 samples of the *Tara* Oceans expedition provided more than 7.2 Tb (terabases or 10^{12} bases) of raw metagenomic data from which resulted a first reference genes catalog of marine picoplanktonic organisms. The analyzed fraction was between 0.2 and 3 µm, and was mainly dominated by Eubacteria, but involved also Archaea, virus/girus and pico-eukaryotes sequences (Sunagawa et al., 2015). Simultaneously, the diversity and structure of eukaryotic fractions from the Tara Oceans

expedition (i.e. planktonic samples which size was comprised between 3 and 2000 µm) were analysed through metabarcoding at 334 samples from 47 stations (Vargas et al., 2015)°. Three years later, the metatranscriptomics from these eukaryotic fractions obtained at 68 stations produced 16.5 Tb of raw data from which 116 million of 'unigenes' were infered, representing as of now the largest reference collection of eukaryotic transcripts from any single biome (Carradec et al., 2018). Even if these *Tara* Oceans catalogues were built on snapshot samples taken irregularly during a four-year cruise, hence allowing no proper seasonal variations investigations, these exhaustive and large / global scale environmental genomic catalogues offer however the unique opportunity to study the evolution and adaptation of microbes in their living environments.

3.2. Protists : from History to current knowledge

Diversity of morphologies, sizes, trophic modes and roles of protists in the ecosystems

The term 'protist' is historical, and refers to a taxonomic rank ('the kingdom of protists') in Whittaker's tree of life (Whittaker, 1969), which is currently commonly used to refer to all eukaryotic lineages that are neither plants, nor animals, nor fungi (Pawlowski et al., 2012). Protists do not form a monophyletic group (i.e. a common hypothetical ancestor and all its descendants) and correspond to the majority of eukaryotic phylogenetic lineages (Baldauf, 2003; Burki et al., 2016; Burki and Keeling, 2014). The protists are unicellular, but some lineages form colonies with a coordinated behavior (Caron et al., 2017). Discussions persist whether or not the term protists should be extended to multicellular eukaryotic lineages with weak / undifferentiated tissues and non-specialized cells (e.g. red macroalgae (Rhodophyta), Phaeophyceae (Adl et al., 2018, 2012)). Most protists are microscopic, but collectively they cover more than five orders of magnitude (from 10^{-1} to 10^{3} µm) (Caron et al., 2017, 2009) (Figure 7).

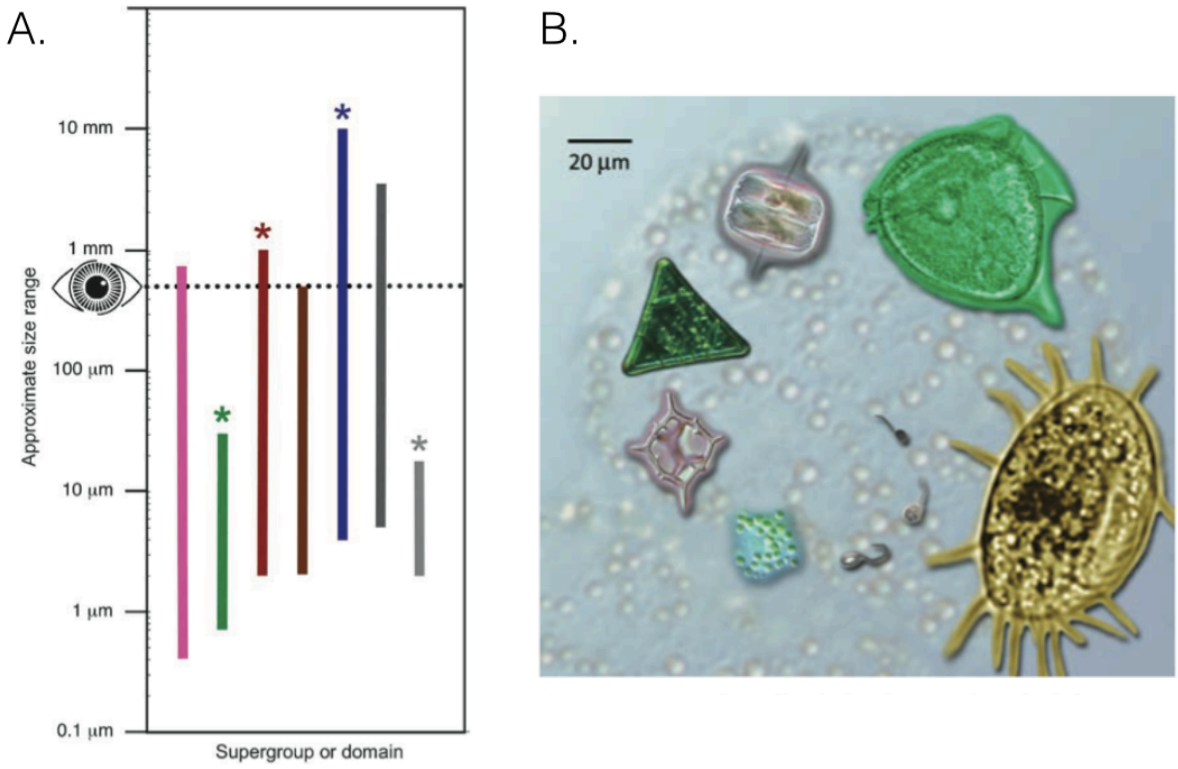


Figure 7 – Approximate size ranges for protists (Caron et al., 2009) (A) as currently placed within eukaryotic supergroups, as well as Eubacteria. Colored columns represent approximate size ranges

among taxa within each group. The dotted line indicates the approximate limit of resolution of the human eye. Note that the overall range of single-celled eukaryotic organisms spans several orders of magnitude and can be smaller than 1 mm. Examples of small and large organismal sizes within each supergroup are (columns from left to right): Eubacteria, *Pelagibacter ubique* and *Thiomargarita namibiensis* (pink); Archaeplastida, *Ostreococcus tauri* and *Chlamydomonas sp.* (green); Chromalveolata, Cafeteria roenbergensis and *Stentor roeseli* (red); Excavata, *Bodo saltans* and *Euglena sp.* (brown); Rhizaria, *Bigelowiella natans* and *Hastigerina pelagica* (blue); Amoebozoa, *Platyamoeba sp.* and *Pelomyxa palustris* (dark gray); Opisthokonta, *Encephalitozoon intestinalis* and *Diaphanoeca grandis* (gray). Asterisks indicate the existence of colonial forms within these supergroups (for example, *Volvox*, a member of the Archaeplastida, which can be 2mm). New protists are discovered each year, we have only molecular marker sequences and no morphological characterization for some of these; therefore this figure is meant only to provide a rough overview of cell sizes. Note the log y-axis scale. (B) A single cell of the largest described bacterium, *Thiomargarita namibiensis* (diameter E180 mm) forms the back-drop for a variety of phototrophic and heterotrophic protists that are shown at the same scale as *T. namibiensis*. Counterclockwise from the lower right, the protists include a ciliate, a dinoflagellate, two diatoms, a silicoflagellate, a colony of small chlorophytes and three minute heterotrophic flagellates.

Protists display a huge variety of morphologies, behaviors and nutritional modes (**Figure 8**). They are present in all environments, but preferentially in aquatic environments (e.g. oceans, lake) and they notably play a central role in marine ecosystems (Falkowski, 2012; Falkowski et al., 2008; Worden et al., 2015). The nutritional modes of protists range from pure phototrophy (n.b. aquatic lineages are traditionally referred as phytoplankton, and phytoplankton involves cyanobacteria and all protists lineages using chlorophyll for photosynthesis)) to pure heterotrophy (referred as protozoa, or also in aquatic environment as zooplankton). Protozoa usually get their nutrition from the ingestions of bacteria, archaea, or other eukaryotes, or can be parasites of others protists and metazoans. In addition, many protists can perform both photosynthesis and phagocytosis simultaneously, they are called mixotrophs (e.g. Chrysochromulina species (Prymnesiophyceae, Haptophyta) have chloroplasts and performed photosynthesis but they also consume preys (Jones et al., 1993)) (n.b. more information on mixotrophy are developed in annexe 4).

Protists reproduce through mitotic division rather than through sexual reproduction, which enables populations to double in a few hours to a few days (Caron et al., 2017). Their rapid growth rates enable them to contribute to important ecosystem functions. Photosynthetic protists have been recognized as major contributors to the existing stock of biomass and primary production in almost all aquatic ecosystems. In the ocean, photosynthetic protists accounts for about 40 Pg C $year^{-1}$ of primary production (i.e. the photosynthetic production of organic carbon from carbon dioxide - inorganic carbon - dissolved in the water), rivaling that of terrestrial plants which is about 50 Pg C $year^{-1}$ (Caron et al., 2017; Field et al., 1998). Protists are partners in various symbiotic relationships and, at multiple trophic levels, as links between the small Metazoa (e.g. shrimps, copepods) that prey on them and the vast numbers of bacteria, archaea, protists and even some metazoa that they consume (**Figure 8**). These links in

the food web are essential for the biogeochemical cycles of the ocean, including the transport of carbon to the deep ocean.

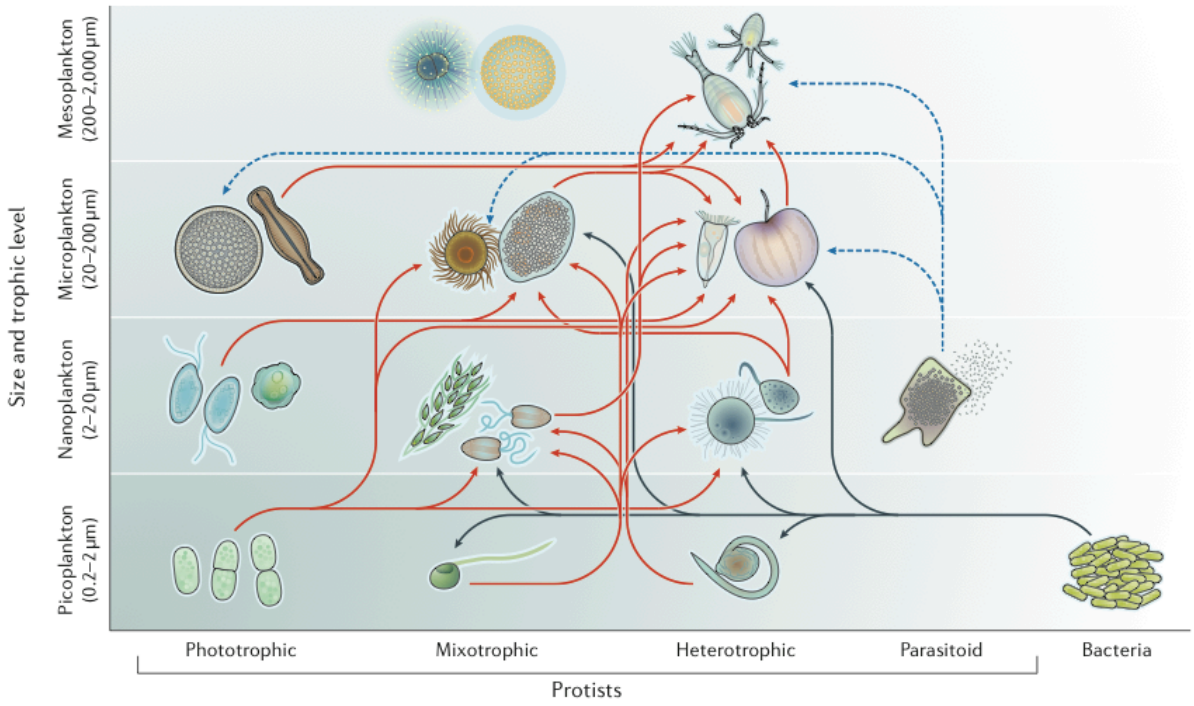


Figure 8 – Ecological and biogeochemical roles of protists in the marine plankton (Caron et al 2017). Protists are an important part of living biomass in oceanic ecosystems and are central to a wide array of food web processes and biogeochemical cycles. Phototrophic, heterotrophic and mixotrophic protists span several trophic levels at, and near, the base of the food web. Complex predator–prey relationships that involve protists make photosynthetically produced organic material (red arrows) and bacterial biomass (black arrows) available to higher trophic levels and also remineralize a substantial fraction of this material back to inorganic nutrients and carbon dioxide to support new primary production. Protistan parasites are also important players in marine food webs, as they prey on microplanktonic and mesoplanktonic species (blue dashed arrows). In addition to the myriad of predator–prey relationships, the interactions of protists with other microorganisms and multicellular organisms include a wide range that are not depicted here (involving competition, commensalism and mutualism).

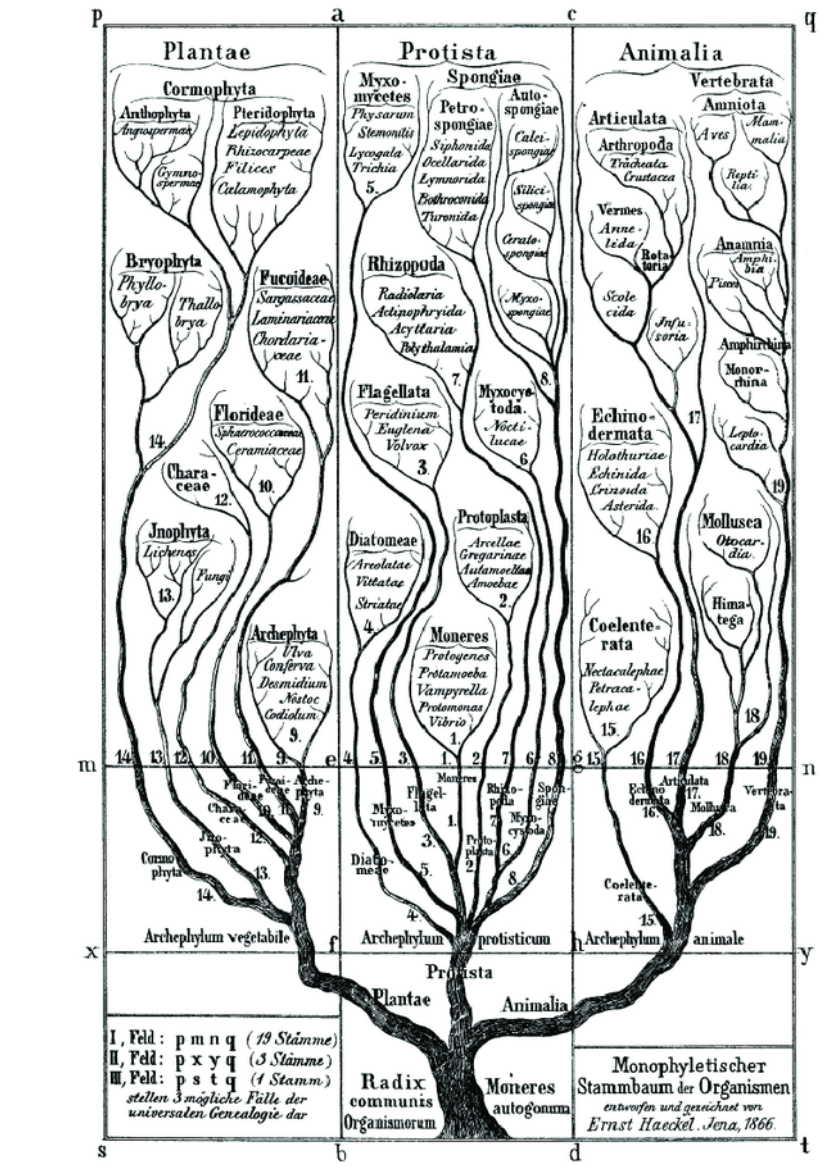
History of protists delineation

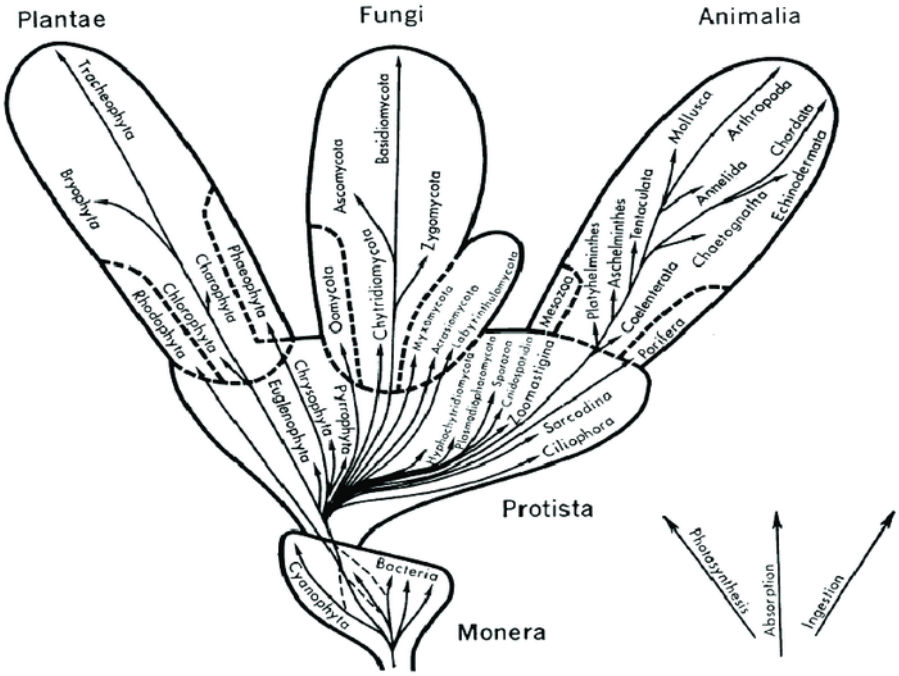
Protists were among the first microbial taxons to be observed and described by Anton van Leeuwenhoek in 1670 and by other pioneers of microbiology in the 17th century (Ereshefsky, 2000). The protists were described as microscopic organisms assimilated to small plants or small animals. The description and inventory of a wide variety of forms and functions in these organisms developed over the following centuries. Among these descriptions, the illustrations of Ernst Haeckel (1899-1904) (Haeckel, 1899) during the nineteenth century are some of the most beautiful scientific illustrations ever made, focusing largely on Radiolarians. As

early as the 19th century, biologists noticed that many micro-organisms were not just animals or unicellular plants: for example, there were Plants-like Animals (Phytozoa) and Animal-like Plants (e.g. the green alga *Volvox* (Chlorophyta) moves like an animal with a flagellum). Haeckel, who was conscious of the problem of classification posed by these organisms not presenting more convincing affinities with the Animals than with the Plants, erected in 1866 a third kingdom: the Protista. This branch included at the time most eukaryotic microorganisms (e.g. diatoms, radiolarians) but also the Monera (Bacteria), and the sponges. The ciliates, on the other hand, were classified with the animals (because Haeckel thought them multicellular), and the cyanobacteria, the green, red and brown macro-algae, the mushrooms and the lichens formed the Inophyta within the Plantae.

The definition of protists evolved also with Haeckel. At first, he classified unicellular and multicellular, as well as organisms with and without nucleus, within the Protista. Then in 1869, he redefined this kingdom on the basis of the absence of sexual reproduction (Whittaker, 1969). Haeckel also considered unicellularity as a criterion for defining protists. It also uses the terms protozoa (Protozoa, etymologically the first animals) to designate the unicellular animals, and protophytes (Protophyta, the first plants) to designate the unicellular plants, considering these groupings as sub-kingdoms of the Protista. Haeckel's tree is cited as a reference today, but in his time, the suggestion of a third kingdom ran counter to mainstream ideas. The Protista kingdom gained credibility in 1911, with what is considered to be the birth of Protistology : a founding article by Dobell explained that Prostista do not have a simple organization but a rather alternative one, very different from Plants or Animals. For Dobell, it is obvious that the current Protists (and in his mind the Protozoa) are in no way the ancestors of the contemporary Metazoans. In conclusion, the Protista can not be called with misleading terms such as « simple », « inferior » or « primitive », and because of their own specific organizations, they should be studied specifically.

Figure 9 - Tree of life in 1866 and 1969. (On the upper part) Tree of Life according to Haeckel (1866). First tree represented with a common ancestry to all living organisms. The branch represented at the center includes all members of the Protista sensu Haeckel group (i.e. prokaryotes and protists). (On the lower part) Tree of Life according to Whittaker (1969) composed of five kingdoms. For Whittaker, protists are an evolutionary stage and a transition between Monera (prokaryotes) and Plants, Fungi and Animals.


Plantae
Protista
Animalia
Cormophyta
Anthophyta
Angiospermae
Gymnospermae
Pteridophyta
Lepidophyta
Rhizocarpeae
Filices
Calamophyta
Bryophyta
Phyllobrya
Thallobrya
Fucoideae
Sargassaceae
Laminariaceae
Chordariaceae
Florideae
Sphaerococcaceae
Ceramiaceae
Characeae
Jnophyta
Lichenes
Fungi
Archephyta
Ulva
Conferva
Desmidium
Nostoc
Codiolum
Myxomycetes
Physarum
Stemonitis
Lycogala
Trichia
Spongiae
Petrospongiae
Autospongiae
Siphonida
Ocellarida
Lymnorida
Eothroconida
Turonida
Calcispongiae
Silicispongiae
Cerutospongiae
Myxospongiae
Rhizopoda
Radiolaria
Actinophryida
Acyttaria
Polythalamia
Flagellata
Peridinium
Euglena
Volvox
Myxocystoda
Noctilucae
Diatomeae
Areolatae
Vittatae
Striatae
Protoplasta
Arcellae
Gregarinae
Autamoebae
Amoebae
Moneres
Protogenes
Protamoeba
Vampyrella
Protomonas
Vibrio
Vertebrata
Amniota
Aves
Mammalia
Reptilia
Anamnia
Amphibia
Pisces
Amphirhina
Monorhina
Leptocardia
Articulata
Arthropoda
Tracheata
Crustacea
Vermes
Annelida
Rotatoria
Scolecida
Infusoria
Echinodermata
Holothuriae
Echinida
Crinoida
Asterida
Mollusca
Otocardia
Himatega
Coelenterata
Nectaculephae
Petracalephae
Archephylum vegetabile
Archephylum protisticum
Archephylum animale
Radix communis Organismorum
Moneres autogonum
I, Feld: p m n q (19 Stämme)
II, Feld: p x y q (3 Stämme)
III, Feld: p s t q (1 Stamm)
stellen 3 mögliche Fälle der universalen Genealogie dar
Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von
Ernst Haeckel. Jena, 1866.


Plantae
Fungi
Animalia
Tracheophyta
Bryophyta
Phaeophyta
Chlorophyta
Rhodophyta
Charophyta
Ascomycota
Basidiomycota
Zygomycota
Oomycota
Chytridiomycota
Myxomycota
Acrasiomycota
Labyrinthulomycota
Plasmodiophoromycota
Hyphochytridiomycota
Sporozoa
Cnidosporidia
Zoomastigina
Chrysophyta
Euglenophyta
Pyrrophyta
Sarcodina
Ciliophora
Protista
Cyanophyta
Bacteria
Monera
Mesozoa
Platyhelminthes
Aschelminthes
Tentaculata
Coelenterata
Porifera
Mollusca
Annelida
Arthropoda
Chaetognatha
Chordata
Echinodermata
Photosynthesis
Absorption
Ingestion

Overview of protistan phylogeny and evolution (modern studies)

Our vision of evolutionary relationships between protistan lineages has changed dramatically in recent decades. Whittaker's system (Figure 9, Figure 10) (Whittaker, 1969), which prevailed for two decades, grouped the organisms into five kingdoms (Monera, Protista, Plantae, Fungi and Animalia) according to their mode of nutrition (i.e., photosynthesis, absorption, ingestion). Unicellular eukaryotes were contained in the Protista, and were themselves organized according to their morphology and nutritional preferences. This practical system, influenced by an ecological and gradual vision, however, separated closely related taxa with different nutritional modes (e.g. dinoflagellates predominantly photosynthetic and dinoflagellates predominantly heterotrophic). The use of ultrastructural informations and DNA sequences allowed to build and infer a new system, placing all eukaryotes in a single domain, with animals, plants, and fungi represented as minor branches among a wide variety of protist lineages (Sogin, 1991) (Figure 10). Nearly 20 years of reorganization followed (Adl et al., 2018, 2012; Baldauf, 2003; Burki, 2014; Simpson and Roger, 2004; Tekle et al., 2009), in which all eukaryotes were classified as "supergroups". This pattern continues to be revised, to a smaller extent, in order to resolve the order of the roots and ramifications of various lineages (Figure 10) (Brown et al., 2018; Burki et al., 2016). Current studies of protistan evolution incorporate now large amounts of genomic and transcriptomic information (Brown et al., 2018; Burki et al., 2016; Grant and Katz, 2014).

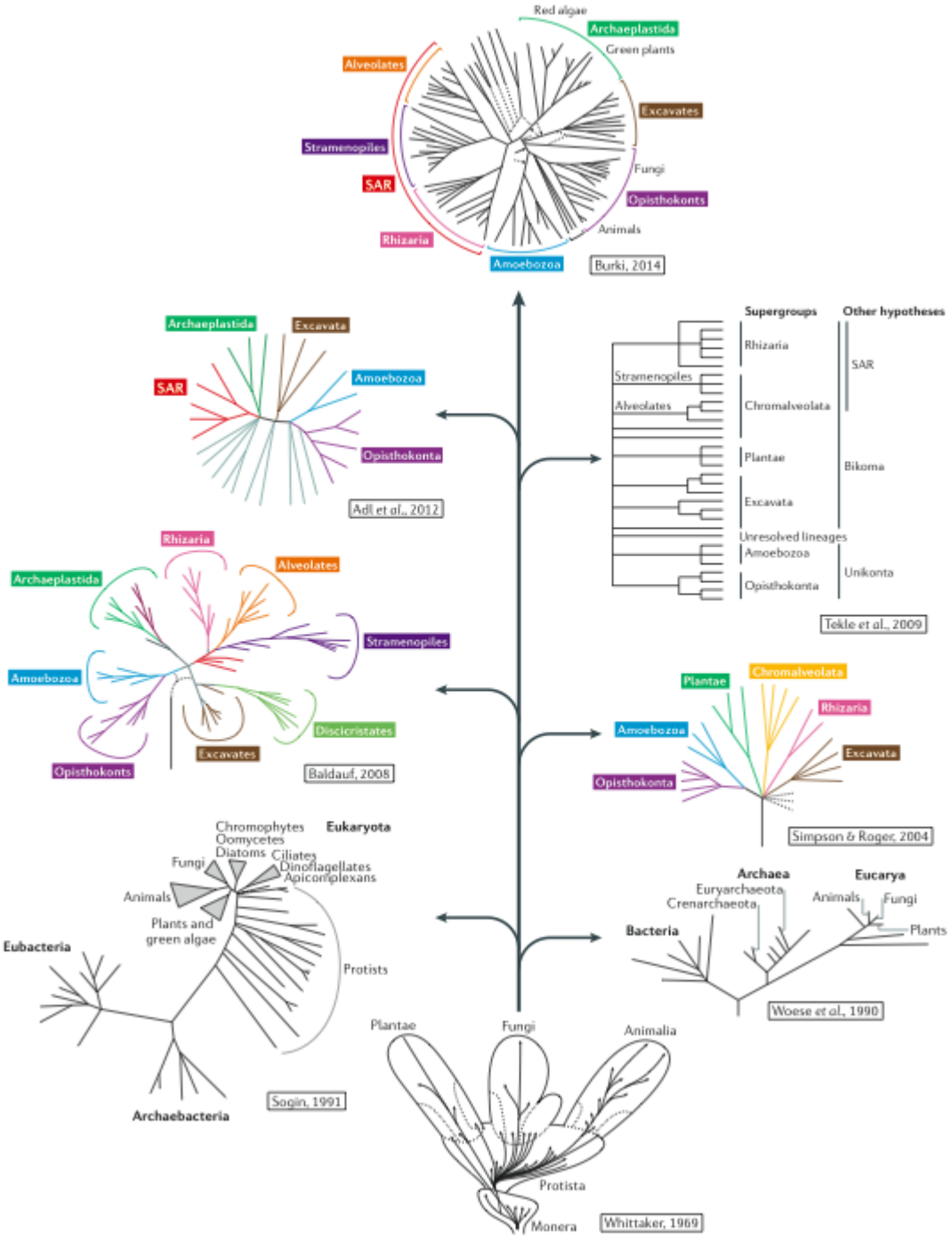


Figure 10 – The rapidly changing landscape of protistan phylogeny (Caron et al., 2017). From Whittaker (1969, a four kingdoms vision of eukaryotes: animals, plants, fungi and protists) to Burki (2014, eight 'super-groups'): DNA sequences and ultrastructural observations from electron microscopy have substantially changed our view of the evolutionary relationships among protistan groups. Our present understanding of protistan evolution (Brown et al., 2018; Burki, 2014; Burki et al., 2016; Grant and Katz, 2014) now relies on large amounts of genomic and transcriptomic information.

Genomic and transcriptomic of protists

The study of eukaryotes arises from an anthropocentric view of life. More than 96% of the described eukaryotic species are either Metazoa (animals), Fungi, or Embryophyta (land plants) (Pawlovski et al. 2012) (**Figure 11**), which correspond to multicellular organisms (even though the Fungi include unicellular members such as the yeasts) and which collectively represent two of the eukaryote super-group (**Figure 10**, phylogenetic tree from (Burki, 2014)). However, in 2014, these three lineages (i.e. Metazoa, Fungi, Embryophyta) only represent 62% of the 18S rDNA Genbank sequences (which is also a biased sample) and 23% of all operational taxonomic units (OTUs) in environmental surveys (**Figure 11**). In genomics this bias is amplified: in February 2017, the Genomes On-Line Database (GOLD) (Pagani et al., 2012) listed more than 7,500 eukaryotic genome sequencing projects as completed or underway, but nearly 90% of them correspond to Metazoa, Embryophyta or Fungi. Moreover, while most of the eukaryotic diversity is in the microbial domain, about 90% of the 200 protist genomes sequenced to date belong to parasites. In fact, genomics studies have first focused on protists impacting humans (e.g. human and other animal parasites, plant pathogens, or lineages of economic interest) (table 1). The 13.4 Mb genome of the yeast *Saccharomyces cerevisiae* (Fungi, Opisthokonts), which has enormous beneficial attributes and is well known for application in the food industry, was the first completely sequenced from a eukaryote (Goffeau et al., 1996). The project started in the early 1990s, and was achieved by an international consortium of researchers from 19 countries working in 94 laboratories using several different sequencing methods and technologies. The 23 Mb genome of the parasite *Plasmodium falciparum* (Alveolata), which causes the malaria in humans and which directly or indirectly causes up to 2.7 million deaths per year (Caron et al., 2009) was sequenced in 2002 (initiative launched in 1996 and published in (Gardner et al., 2002)). The 26 Mb genome of *Trypanosoma brucei* (Excavata), on other human parasite which causes sleeping sickness, was sequenced in 2005 (Berriman et al., 2005). The 240 Mb genome from the plant pathogen *Phytophthora infestans* (Oomycetes, Stramenopiles), which is the causal agent of the late potato blight disease and also infects tomatoes and other members of the Solanaceae, was published in 2009 (Haas et al., 2009).

Many eukaryotic lineages still have only one or a few representatives with sequenced genomes, and many remain completely unsampled (**Figure 11**). Archaeplastida and Stramenopila have more cultured species than other eukaryotes as a result of a long phycological tradition and of phycological culture collections (e.g. RCC (Roscoff Culture Collection, Roscoff, France, http://www.roscoff-culture-collection.org), ATCC (American Type

Culture Collection; Manassas, Virginia, USA, http://www.atcc.org), CCAP (Culture Collection of Algae and Protozoa; Oban, Scotland, UK, http://www.ccap.ac.uk)), and also because they are easier to maintain in culture than heterotrophs. A comparatively larger number of genome projects targets photosynthetic stramenopiles (Armbrust et al., 2004; Bowler et al., 2008; Cock et al., 2010) and, owing to their economic relevance in the agriculture, oomycetes (Pais et al., 2013) (Pais et al. 2013) (**Figure 11**, **Figure 12**). The apicomplexans (Alveolata) are also relatively well studied at the genomic level because they contain parasites (e.g. *Plasmodium*, *Toxoplasma*). When looking at the number of sequenced strains rather than species, these biases are increased further. A significant proportion of the retrieved cultures and genomes corresponds to different strains of the same dominant species, so a pool of species have been redundantly cultured and sequenced (**Figure 12**).

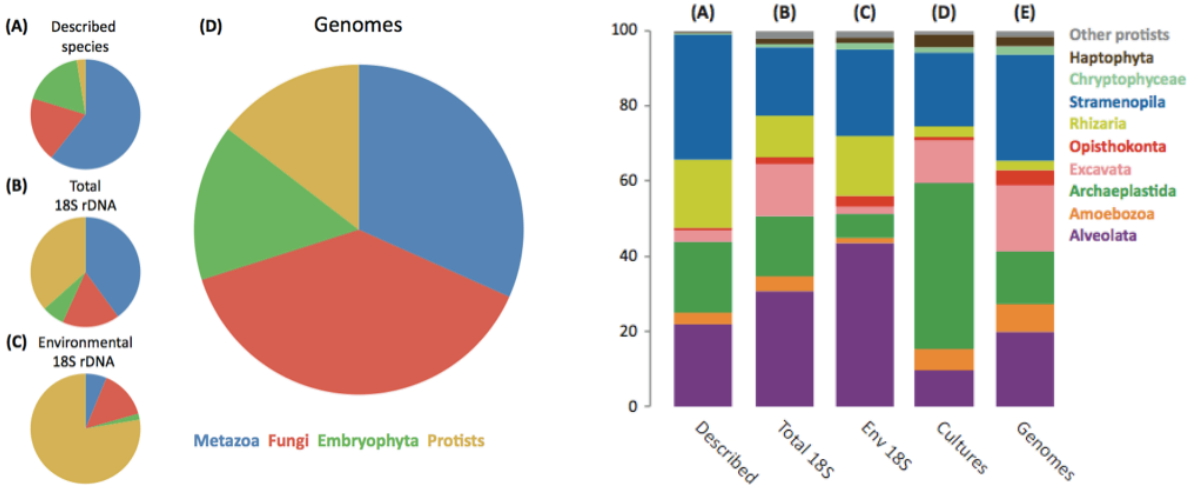


Figure 11 - Relative representation of protists in current databases (del Campo et al., 2014). On the left, pie charts showing the relative representation of metazoans, fungi, and land plants versus all the other eukaryotes in different databases. (A) Relative numbers of described species according to the CBOL ProWG (Consortium for the Barcode of Life Protist Working Group, Pawlowski et al . 2012) (n = 2 001 573). (B) Relative numbers of 18S rDNA OTU97 (operational taxonomic unit at >97% sequence identity) in GenBank in 2014 (n = 22 475). (C) Relative number of environmental 18S rDNA OTU97 in GenBank in 2014 (n = 1165). (D) Relative number of species with a genome project completed or in progress according to GOLD in 2014, per eukaryotic group (n = 1758). On the right, barplots illustrating the relative representation of eukaryotic supergroup diversity in different databases (excluding metazoans, fungi, and land plants). (A) Percentage of described species per eukaryotic supergroup according to the CBOL ProWG. (B) Percentage of 18S rDNA OTU97 (operational taxonomic unit at >97% sequence identity) per eukaryotic supergroups in GenBank. (C) Percentage of environmental 18S rDNA OTU97 per eukaryotic supergroups. (D) Percentage of species with a cultured strain in any of the analyzed culture collections (details in del Campo et al. 2014). (E) Relative numbers of species with a genome project completed or in progress according to GOLD, per eukaryotic group.

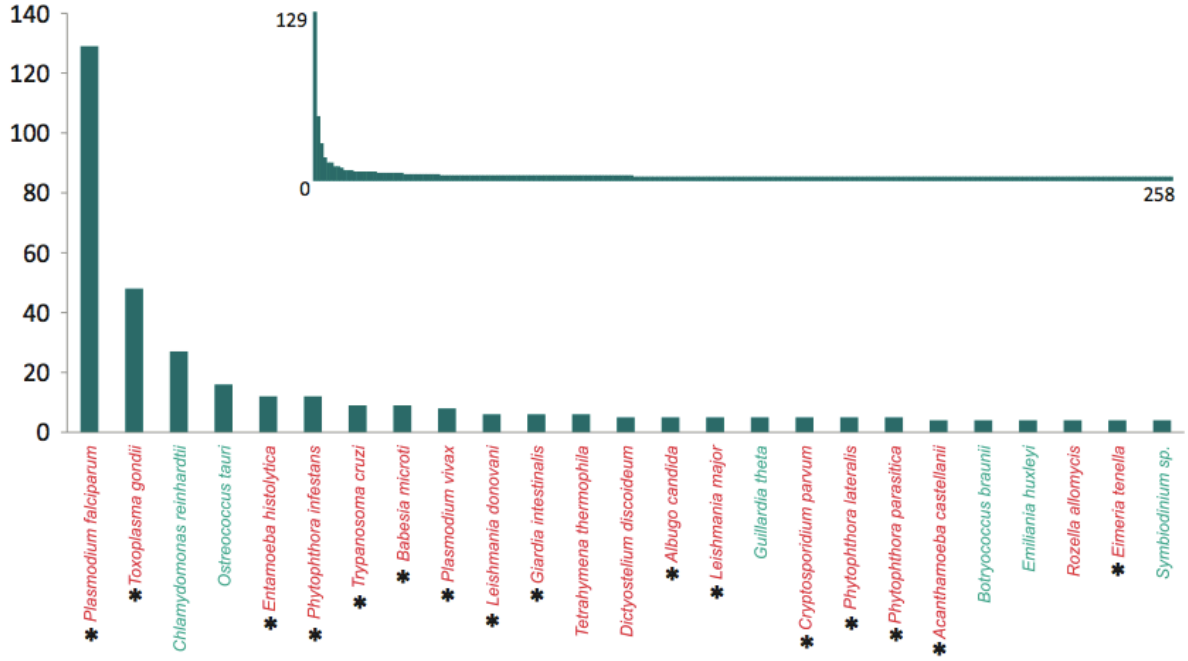


Figure 12 – Top 25 eukaryotic lineages among databases (del Campo et al. 2014) (some strains are not described at the species level and have been grouped by genus, so they may represent more than a single species) [*Parasitic organisms , red = colorless, green = pigmented]

Unlike most Eubacteria, Archaea and Viruses, the genomes of protists tend to be larger and more complex, which is a limiting factor for genomic level of assembly and annotation steps. The genome sizes of protists are also highly variable, covering about six orders of magnitude (from 10^6 to 10^{11} bp (table 1), e.g. the Dinoflagellates genome sizes are estimated as between 0,5 and 40 times the size of the human genome (Wisecaver and Hackett, 2011); the smallest protist genome reaches 2.3 Mb and belongs to the parasite *Encephalitozoon intestinalis* (microsporidia, Fungi or related to Fungi (Keeling, 2014)) (Corradi et al., 2010); the biggest eukaryotic genome belongs to *Polychaos dubium*, an amoeba, and is supposed to reach 670 Gb (Parfrey et al., 2008)).

Taxa	Taxonomy (based on Burki et al. 2016 phylogeny)	Notes	Genome size estimation	Number of protein coding genes (estimation)	References
Encephalitozoon intestinalis	Obazoa, Microsporidia, Fungi or related to Fungi	parasite, smallest eukaryote genome	2.3 Mb	1,833	(Parfrey et al., 2008)
Escherichia coli	Eubacteria	strain K-12 (non pathogen strain)	4.6 Mb	4,288	(Blattner et al., 1997)
Ostreococcus tauri	Archaeplastida, Prasinophyceae	smallest free-living eukaryotic genome	12.6 Mb	8,166	(Derelle et al., 2006)
Saccharomyces cerevisiae	Obazoa, Fungi	yeast, economic interest	13.4 Mb	6,091	(Goffeau et al., 1996) [first publication], then re-analysis : (Lin et al., 2013)
Plasmodium falciparum	Alveolata	causes malaria	22.8 Mb	5,268	(Gardner et al., 2002)
Entamoeba histolytica	Amoebozoa	anaerobic parasitic, pathogen	23.7 Mb	9,938	(Loftus et al., 2005)
Trypanosoma brucei	Excavata	parasite, causes sleeping sickness	26.1 Mb	9,068	(Berriman et al., 2005)
Dictyostelium discoideum	Amoebozoa	soil-living amoeba, 'social' amoeba	33 Mb	12,500	(Eichinger et al., 2005)
Oxytricha trifallax	Alveolata, Ciliates	analysis of the macronucleus	50 Mb	18,400	(Swart et al., 2013)
Toxoplasma gondii	Alveolata	parasite, causes toxoplasmosis	60-80 Mb		(Bontell et al., 2009; Lau et al., 2016)
Arabidopsis thaliana	Archaeplastida, Embryophyta		115 Mb	28,000	(Arabidopsis Genome Initiative, 2000)
Emiliania huxley	Haptophyta	Pan-genome effect	142 Mb	30,569 (but pan-genome effect)	(Read et al., 2013)
Symbiodinium spp.	Alveolata, Dinoflagellates	sometimes involve in coral symbiosis	1,03-4,8 Gb	57,000	Lajeunesse et al 2005; Bayer et al 2012; Lin et al 2015; Liu et al 2018
Homo sapiens	Obazoa, Metazoa		3 Gb	23,000	(Lander et al., 2001)
	Alveolata, Dinoflagellates		1.03-112 Gb	33,000 - 76,000	(Lie et al., 2018; Murray et al., 2016)
Psilotum nudum	Archaeplastida, Embryophyta	fern, polyploïdy	250 Gb		(Hidalgo et al., 2017)
Polychaos dubium or *Amoeba dubia*	Amoebozoa	freshwater amoeboid, the numbers are controversial, certainly polyploïd	670 Gb		(Friz, 1968; Parfrey et al., 2008)

Table 1. Non-exhaustive list of the genome size of protists (for comparative purpose *Homo sapiens* and *Escherichia coli* information were also indicated). Rows are sorted by increasing genome size.

Spectacular differences are observed between eukaryotic closely related organisms and also within species (Parfrey et al., 2008; Read et al., 2013). Genome size is not, as it is often too simply assumed, linked to the 'complexity' of an organism or even to the number of genes in its genome, an observation known as the C-value paradox (Keeling, 2007; Mirsky and Ris, 1951). The study of eukaryotic and, more precisely of protistan genome heterogeneity (e.g. based on the %GC content; the length, structure and distribution of introns; the abundance, structure and function of non-coding DNA; composition, abundance and dispersal of repeats and/or transposable elements) constitutes a wide and continuously promising field of research. In light of our current view of eukaryotic (nuclear and organellar) genomes, one of the most striking features is , for instance, that the 12.6 Mb genome of the pico-eukaryotic prasinophyte green alga *Ostreococcus lucimarinus* (which is also reported as the smallest free living eukaryotic genome), has one of the highest gene densities known in eukaryotes, yet it contains many introns (Derelle et al., 2006; Lanier et al., 2008) (**Figure 13**). Phylogenetic studies suggest this unusually compact genome (12.6 Mb) is an evolutionarily derived state among prasinophytes. An other important trend was observed in diatoms (based on *Thalassiosira pseudonana* (32.4 Mb (Armbrust et al., 2004)) and *Phaeodactylum tricornutum* (27.4 Mb (Bowler et al., 2008)) genomes): approximately 95% of their DNA is reported as non-coding (Vardi et al., 2008)), which stimulates epigenomic studies for these lineages (Loftus et al., 2005; Rogato et al., 2014).

An other interesting pattern was highlighted by the sequencing of 14 strains from the same species, *Emiliania huxleyi* (Haptophyta, coccolithophore) (142 Mb (Read et al., 2013)), which revealed a pan genome (core genes plus genes distributed variably between strains) probably supported by an atypical proportion of repetitive sequences. *E. huxleyi* genome is indeed dominated by repetitive elements, constituting more than 64% of the sequence, much greater than seen for instance in sequenced diatoms (i.e. 15% in *P. tricornutum*). This extensive genome variability, reflected in different metabolic repertoires, seems to underpin the capacity to thrive in a broad scale of habitats and to form large-scale episodic blooms under a wide variety of environmental conditions.

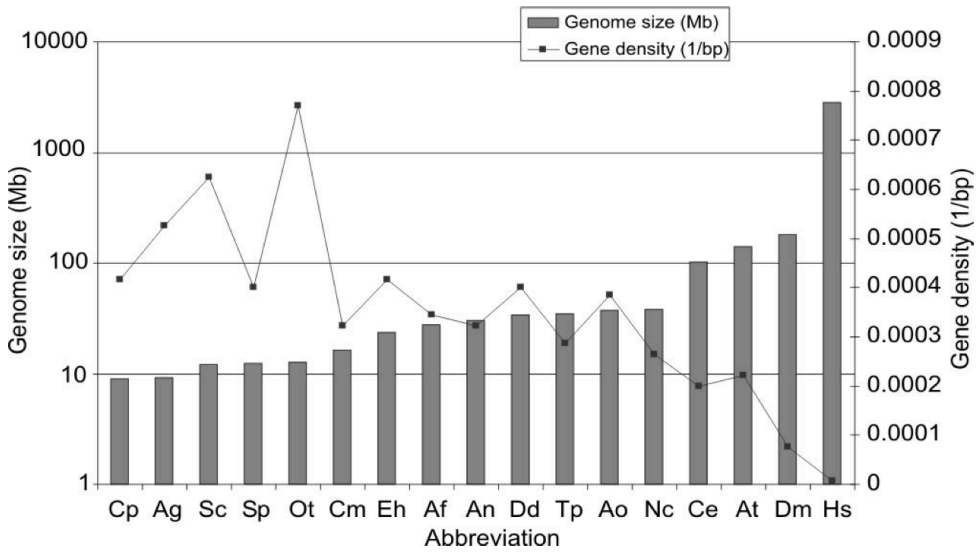


Figure 13 - Genome size and gene density for various eukaryote genomes (Derelle et al., 2006). Cp, *Cryptosporidium parvum*; Ag, *Ashbya gossypii*, Sp, *Schizosaccharomyces pombe*; Sc, *Saccharomyces cerevisiae*; Ot, *Ostreococcus tauri*; Cm, *Cyanydioschyzon merolae*; Eh, *Entamoeba hemolytica*; Af, *Aspergillus fumigatu*; An, *Aspergillus niger*; Dd, *Dictyostelium discoidum*; Tp, *Thalassiosira pseudonana*; Ao, *Aspergillus oryzae*; Nc, *Neurospora crassa*; Ce, *Caenorhabditis elegans*; At, *Arabidopsis thaliana*; Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*.

Each sequenced genome of an unicellular eukaryote has provided a bevy of new and unexpected insights (e.g. (Armbrust et al., 2004; Bowler et al., 2008; Derelle et al., 2006; Read et al., 2013; Worden et al., 2009). However, because nuclear genomes can be large, difficult to sequence and to assemble (e.g. when repeat elements constitute a large proportion of a genome - as in eukaryotic genomes - *de novo* assembly methods failed (Haas et al., 2013; Koch et al., 2014), the most obvious alternative to generate new datasets from non-model organisms is transcriptomics. Transcriptomics (currently called RNASeq when it is obtained by NGS) correspond to the large-scale sequencing of an organism's mRNA. RNASeq allows the rapid and efficient characterization of expressed genes without spending sequencing resources on the large intergenic regions, introns, and repetitive DNA, while at the same time eliminating many problems with assembly as well as gene prediction and modeling. As a first step, transcriptomes from pure cultures or from single-cell organisms isolated from the environment (i.e. either from pool of similar cells, or from single-cell alone - with or without random amplification steps) are suitable building blocks to begin to assemble reference databases for eukaryotic microbial ecology. This approach generates a large number of coding sequences (in the form of assembled contigs) from a known organism.

N.b. Genomic and transcriptomic approaches have different strengths and weaknesses and are better viewed as complementary rather than ''either/or.'' Indeed, nuclear genome sequencing generally requires substantial transcript sequencing to inform gene prediction

algorithms. As sequencing and computational methods grow increasingly powerful, many of the challenges to genome sequencing are being reduced. Nevertheless, until more genomes are available, transcriptomes from a sufficient number of representative species from a given environment could provide a valuable benchmark against which environmental data can be analyzed.

From 2008, the year of the first RNA-Seq publication on the Sequence Read Archive (SRA) database, an increasing and exponential growth of transcriptomics has been conducted (e.g. 74 runs in 2008, 208,892 runs in 2014 (Jazayeri et al., 2014)). In 2014 was officially launched the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP) (Keeling et al., 2014). This global community effort targeted to augment the available DNA sequence information on ecologically relevant protists. The overall goal was to substantially increase the sequence datasets from cultured, well-curated protistan lineages (originated from all around the world (**Figure 14**), to provide a resource for gene annotation throughout the Eukarya (**Figure 15**, **Figure 16**) and for the interpretation of metagenomic and metatranscriptomic datasets from diverse marine ecosystems. Supported by the Gordon and Betty Moore Foundation, the project directly involved nearly 70 laboratories engaged in protistan research, and more than 200 investigators worldwide participated by submitting material for sequencing. The project entailed the sequencing and assembly of 678 transcriptomes that encompassed 210 unique genera, 305 species and 396 strains of protists, and the results were released to the public in June 2014 and are now publicly available through the iMicrobe website (https://www.imicrobe.us) (n.b. a new re-assembly was published in by (Johnson et al., 2018)).

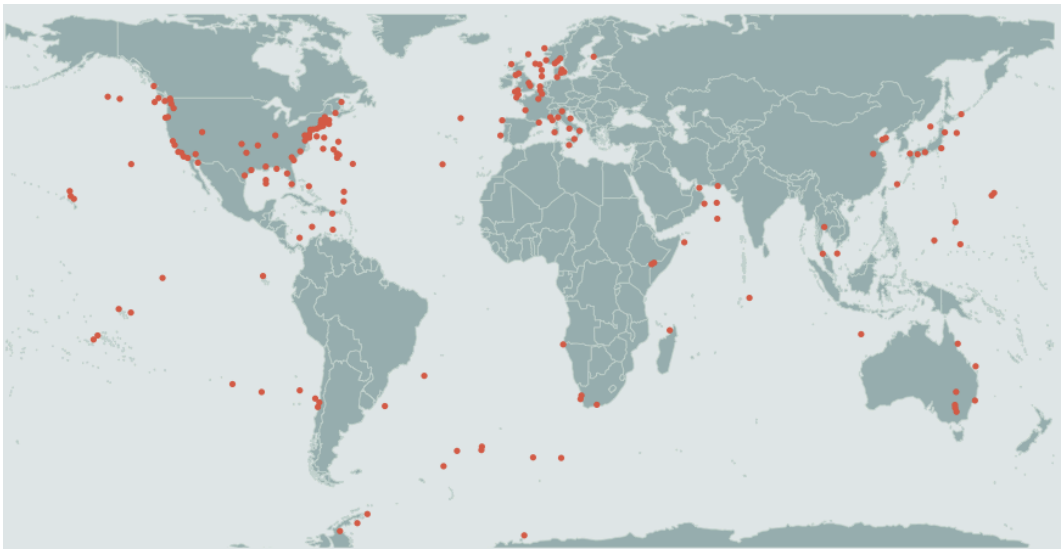


Figure 14 – Geographical origins of approximately half of the strains from the MMETSP (Keeling et al., 2014), indicating the degree of global coverage (Caron et al., 2017).

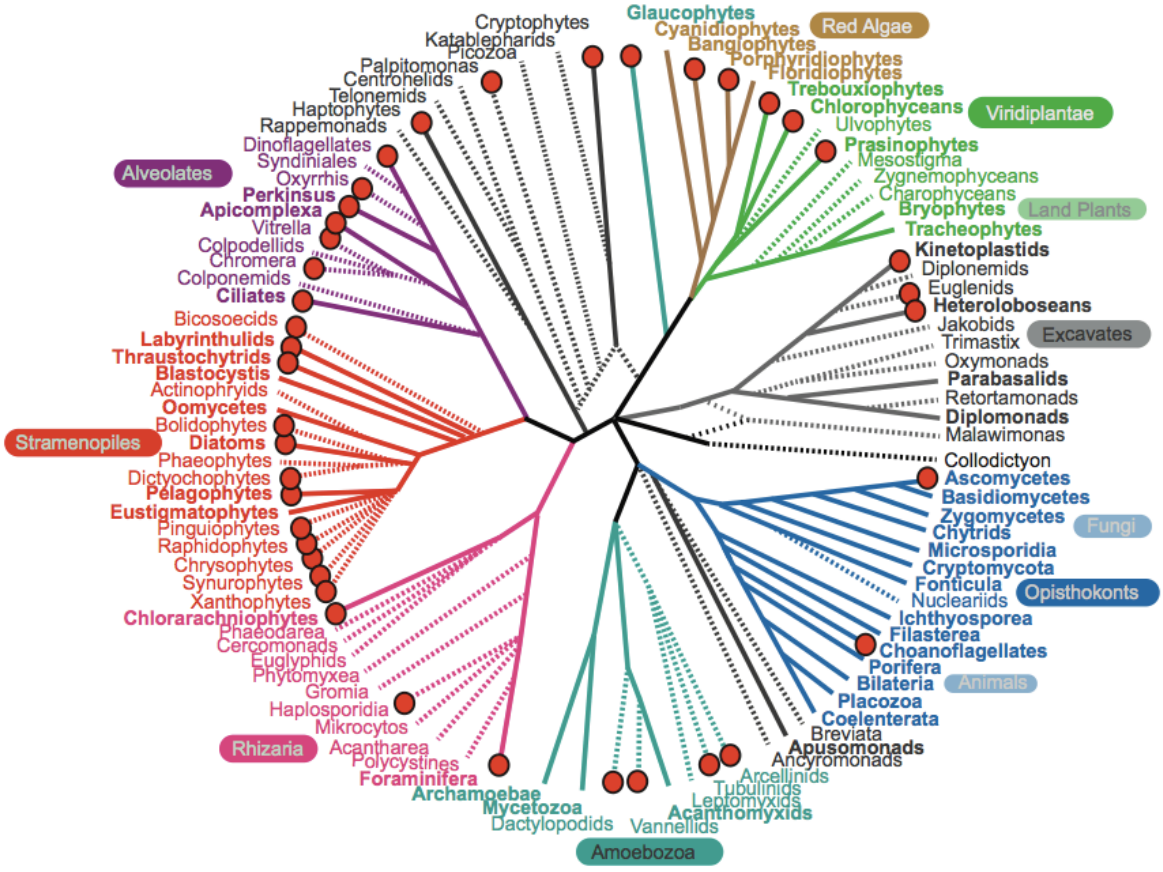


Figure 15 - A schematic of the major lineages in the eukaryotic tree of life (Keeling et al., 2014), showing the relationships between lineages for which genomic resources are currently available and those that have been targeted by the MMETSP. Lineages with complete genomes according to the GOLD database, are indicated by a solid line leading to that group, whereas lineages with no complete genome are represented by a dashed line. Lineages where at least one MMETSP transcriptome is complete or underway are indicated with a red dot by the name.

As a conclusion, our eukaryotic and protistan genomic view is thus currently biased (Figure 11, Figure 12, Figure 16), however, in the past few decades, a lot of DNA sequence information has become available for protistan marker genes (from reference databases (Guillou et al., 2013)°, since 2018 http://www.eukref.org/, and from the environment (e.g. Vargas et al., 2015°), transcriptomes (e.g. Keeling et al., 2014), genomes (almost 20 complete genomes in total) (Del Campo et al., 2014), metatranscriptomes (e.g. Keeling et al., 2014; Carradec et al., 2018) and even few on metagenomes (e.g. (Cuvelier et al., 2010)).

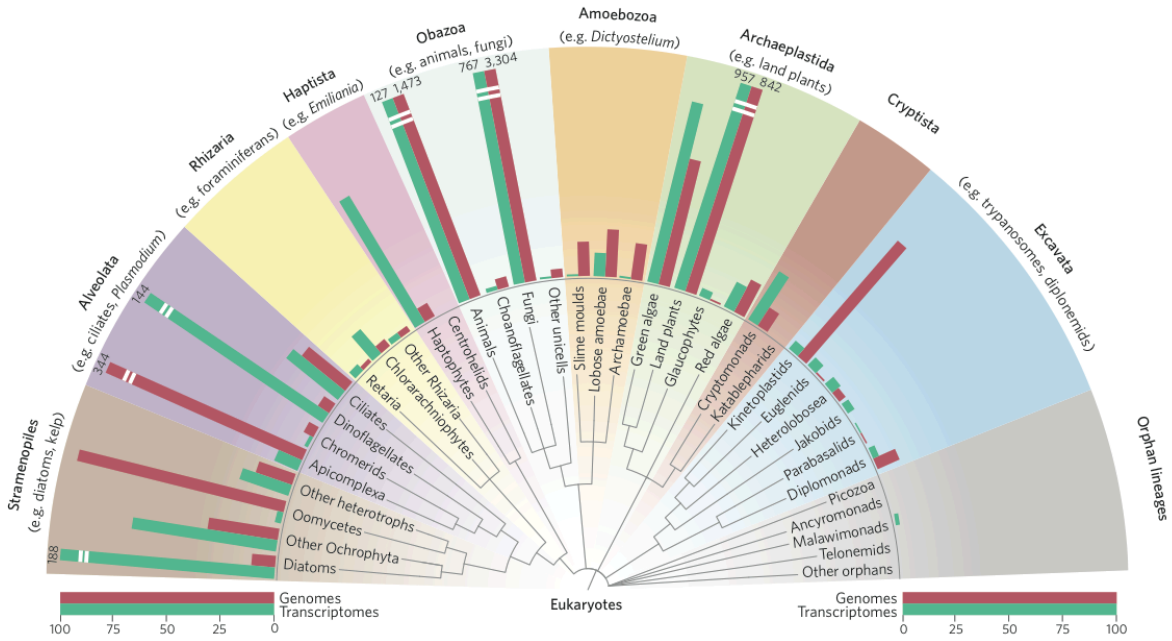


Figure 16 – Genomes and transcriptomes across the eukaryotic tree of life (Sibbald and Archibald, 2017). The cladogram summarizes the diversity of eukaryotes based on phylogenetic relationships in Burki et al. 2016 The histogram shows the number of genome and transcriptome sequencing projects listed as complete or in progress in the GOLD as of February 2017. Transcriptome data also include sequences from the MMETSP (Keeling et al., 2014).

Metagenomic and Metatranscriptomics of protists

The application of culture-independent molecular approaches to the study of natural assemblages of protists has revealed a 'hidden world' of microbial eukaryotic diversity (Caron et al., 2009). Such studies started with the use of plastid-targeted primers, both for the plastid-derived 16S rRNA genes (Rappé et al., 1998) and rbcL genes (Pichard et al., 1997). Subsequently, publications mainly based on the nuclear encoded 18S ribosomal (r)DNA genes (i.e. Small Subunit (SSU) ribosomal DNA), have established the presence of a large number of 'undescribed, un-cultured' taxa and whole lineages in natural protistan assemblages (e.g. (Countway et al., 2007; Cuvelier et al., 2008; Groisillier et al., 2006; López-García et al., 2003, 2001; Massana et al., 2002; Moon-van der Staay et al., 2001; Not et al., 2007; Romari and Vaulot, 2004; Shi et al., 2009). These first molecular surveys (from 2000 to 2010) were based on clone libraries of near full-length 18S rDNA followed by Sanger sequencing of a subset of the clones. The resulting, often manually checked, environmental sequences have been crucial for the phylogenetic placement of novel clades (e.g. the supposedly heterotrophic Picozoa (Not et al., 2007b; Seenivasan et al., 2013), the marine parasitic alveolates (MALV) (Guillou et al.,

2008) and the marine bacterivorous stramenopiles (MAST) (Massana et al., 2004)). These 'long' environmental sequences, associated with sequences from monoclonal cultures, are the current basis of reference rDNA databases (e.g. (Guillou et al., 2013)° (Decelle et al., 2015)). However, it was then commonly thought that traditional clone libraries only capture the most dominant species in the community (Pedrós-Alió, 2006), a limitation which should be bypassed by high-throughput sequencing (HTS) methods which is providing in comparison deeper inventories on a larger number of samples. Since 2009, metabarcoding approaches using HTS (i.e. 454 and then Illumina sequencing technologies) has been applied to study protist diversity in a wide variety of ecosystems, including surface and deep marine waters (e.g. (Amaral-Zettler et al., 2009; Edgcomb et al., 2011) (Bittner et al., 2013; Egge et al., 2015; Logares et al., 2014a; Massana et al., 2015; Vargas et al., 2015)°), marine sediments (e.g. (Bik et al., 2012) (Forster et al., 2016a; Massana et al., 2015)°), lakes (e.g.(Mangot et al., 2013)), soils (e.g. (Bates et al., 2013; Fiore-Donno et al., 2016; Mahé et al., 2017)) and metazoan hosts (e.g. (He et al., 2014)). These surveys used environmental mainly DNA and more occasionally RNA as template for PCR amplification, and it has been shown that using both can provide a different picture of biodiversity (Not et al., 2009; Stoeck et al., 2007) and useful complementary information (Blazewicz et al., 2013) (Bittner et al., 2013)°.

The choice of the marker gene for the survey or of the relatively short targeted marker subregion (e.g. (Dunthorn et al., 2012; Stoeck et al., 2010) (Bittner et al., 2013)°) or of the PCR primers sets (e.g. (Amaral-Zettler et al., 2009; Antoine-Lorquin et al., 2016; Egge et al., 2013; Stoeck et al., 2010)), the nature of the sequencing platform (e.g. (Mahé et al., 2015)), as well as the pipeline to analyse the metabarcodes (e.g. quality filtering parameters (removal of low quality sequences and / or singletons), (Egge et al., 2013); clusters OTU building algorithms and potential similarity threshold, (Callahan et al., 2017; Egge et al., 2013; Forster et al., 2016b)), have all a strong and significant impact on the inferred results. Morevover, only very few studies are concretely addressing the issues about the link between the abundance and richness of metabarcodes, the number of species present in the corresponding sample, as well as their relative abundance in terms of cell numbers and biomass (e.g. (Egge et al., 2013; Malviya et al., 2016)°). These relationships are complicated by the extreme variation in the number of marker gene copies per nucleus among species ((Godhe et al., 2008; Zhu et al., 2005); which can differ even within a single species, (Wang et al., 2017)) and the number of nuclei per individual. While bacteria have between one and ten copies of the SSU in their genomes, microbial eukaryotes can have a number of copies that differ by many orders of magnitudes. Some groups, like dinoflagellates, can have up to 12,000 copies (Zhu et al., 2005) while others, like MAST-4, have only 30 (Rodríguez-Martínez et al., 2009). A current practical

solution to this problem is to compare environmental communities only in terms of relative taxon abundance (e.g. normalizing sequence reads per OTU).

Despite these numerous, now well-known biases, today in 2018, metabarcoding studies involving PCR amplification steps (i.e. which involve more artefacts than the miTags approach sometimes used in prokaryotic metabarcoding studies (Logares et al., 2014b; Sunagawa et al., 2015), are routinely (and pragmatically) used to assess the diversity and structure of eukaryotic microbial communities. Metabarcoding at a very broad scale is also again stimulating hypotheses regarding the protistan rare biosphere (Figure 17) — diverse taxa present at low relative abundances in virtually all natural ecosystems — which may play important parts in the evolution of eukaryotes as well as in the stability and functional resilience of microbial communities (Logares et al., 2014a; Ser-Giacomi et al., 2018). Furthermore, a clear and recurrent pattern is found in the environmental surveys: lineages without cultured representatives dominate in the environment (e.g. (Liu et al., 2009; Shi et al., 2009) (Bittner et al., 2013 ; Forster et al., 2015)° (Figure 18): any newly produced environmental sequences are in majority more closely related to previous environmental sequences than to sequences obtained from organisms relatively well-studied in the laboratories.

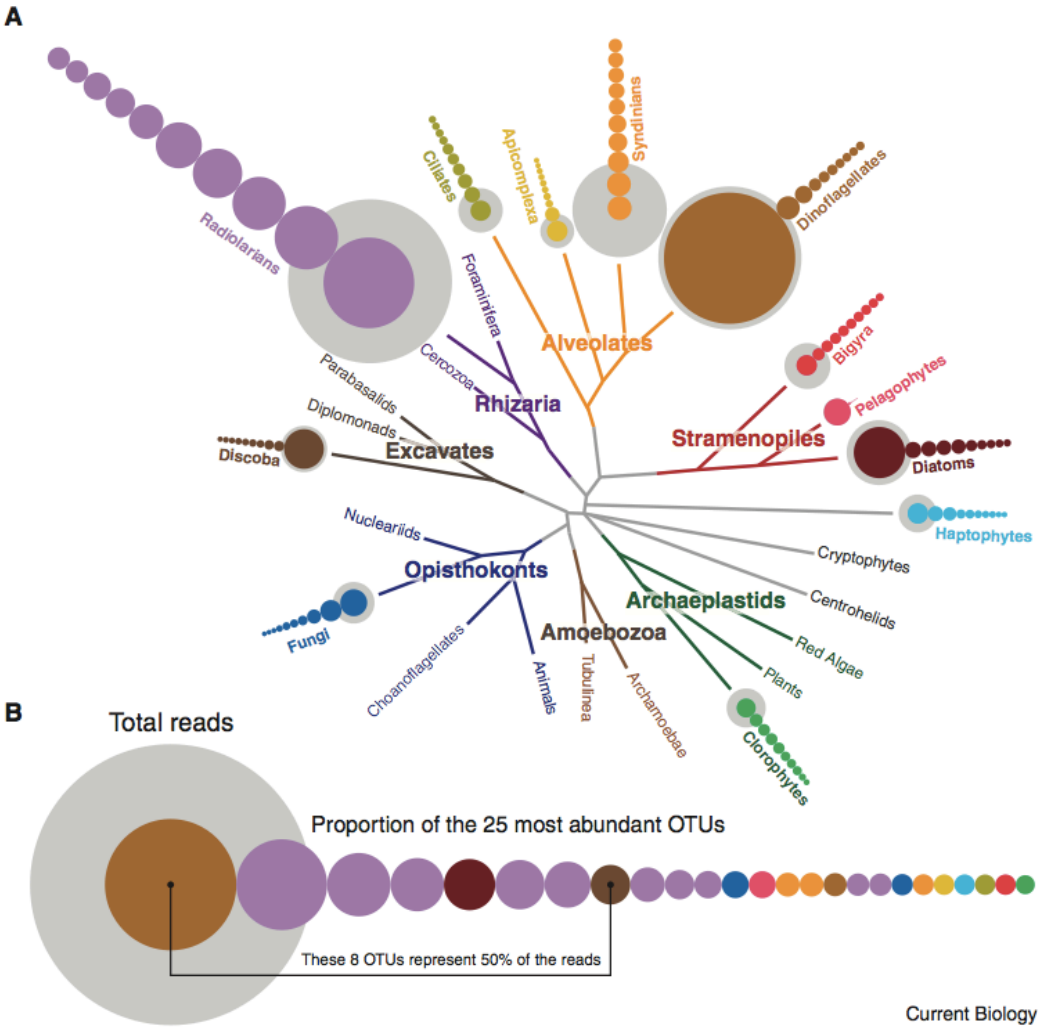


Figure 17 - Relative abundance patterns of the most common microbial eukaryotic OTUs in Tara Oceans (Keeling and Campo, 2017). (A) The relative abundance of the most common operational taxonomic units (OTUs) within 12 well-defined and diverse protist lineages (Vargas et al., 2015). For each lineage, the total number of reads for the entire group is shown as a grey circle (the size shown to scale between lineages), and the ten most common individual OTUs are shown as coloured circles of descending size. (B) The 25 most abundant protist OTUs in the entire *Tara* Oceans metabarcoding eukaryotic data set, in which as above, the grey circle represents the size of the whole data set, while the coloured circles represent individual OTUs, colour coded according to lineage as in panel A. The first eight OTUs account for over 50% of the total number of reads in the entire data set.

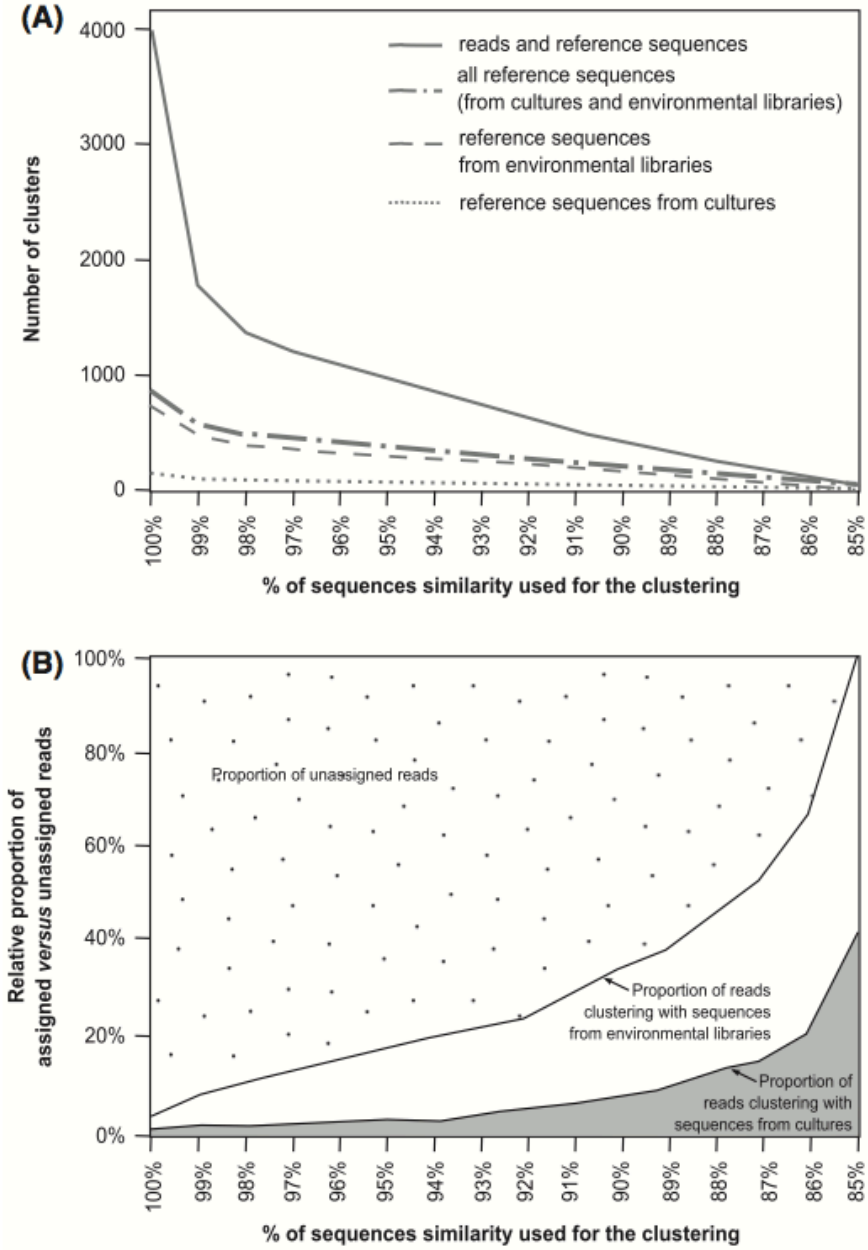


Figure 18 – The known versus the unknown omic diversity from environmental Haptophyta. Metabarcoding study of LSU markers (i.e. 28S rDNA) targeting planktonic Haptophyta (Bittner et al., 2013). (A) Number of clusters as a function of clustering level. (B) Proportion of unassigned vs. assigned reads as a function of clustering level. Full lines indicate the proportion of reads clustering at least with one reference sequence obtained by Sanger sequencing of environmental or cultured samples. Dashed lines indicate the proportion of reads clustering with reference sequences from cultured Haptophyte strains. This figure clearly illustrates that newly produced environmental sequences are in majority more closely related to previous environmental sequences than to sequences obtained from organisms relatively well-studied in the laboratories.

Large genome sizes and complexity of their structure (e.g. predominance of non coding and / or repeated sequences; cf. previous section on the Genomics of protists), considerable diversity and largely unexplored physiology, sequencing depth and cost of sequencing have limited the number of available metatranscriptomic and metagenomic datasets of natural eukaryotic assemblages. These approaches are becoming every day more tractable, in part owing to the still, constant decreasing costs of HTS and in part through the improvements and growing amount of reference datasets for gene annotation. Most of the studies yet published are focusing on communities with limited species richness and are restrained to small size fractions (e.g. for the planktonic communities, only the pico- and (rarely) the nano-eukaryotes are analyzed, and sequences were also obtained serendipitously from studies focusing on prokaryotic communities, (Cuvelier et al., 2010; Delmont et al., 2015; Ottesen et al., 2011; Piganeau et al., 2008)). By contrast, global surveys of the functional potential of microbes which size is inferior to 3 µm and double- stranded DNA viruses are advancing rapidly (e.g. gut, ocean) mainly because of the availability of comprehensive gene catalogs (e.g. (Brum et al., 2015; Qin et al., 2010; Roux et al., 2016; Sunagawa et al., 2015)). However, a cornerstone study in the environmental omics study of marine protists has been recently published (Carradec et al., 2018): compiling informations from reference transcriptomes and genomes, and newly produced metagenomes and metatranscriptomes from the *Tara* Oceans expedition, a eukaryotic gene catalogue of 116 million unigenes (i.e. a complete or partial transcript assembled from metatranscriptomic reads of at least one *Tara* Oceans station) has been created based on the analysis of 16.5 Tb of raw data. Four main organismal size fractions (from 0.8 to 2000 µm) at 68 stations corresponding to 441 samples were sampled independently to optimize the recovery of comprehensive metatranscriptomes from protists to zooplankton and fish larvae (Alberti et al., 2017; Pesant et al., 2015). One of the main results is that the gene repertoire of planktonic eukaryotes is massive and diverse, much more than the prokaryotic gene space (i.e. rarefaction analysis revealed that the sampling effort did not result in near saturation of the eukaryotic gene space, contrasting with the metagenomic results obtained from the prokaryote-enriched size fractions (Sunagawa et al., 2015)). This catalog unveils thus at a broad and global scale the functions expressed by planktonic communities, and constitutes an unprecedented resource to study the functional biogeography of protists. For the principal groups of phytoplankton, it was possible to obtain insights between adaptive and acclimatory processes underlying organismal responses to their environment using as proxies the contrasts between metagenomics and metatranscriptomics. The results indeed suggested that nutrient limitations are dealt with in different ways among the main photosynthetic taxa, either by a genotypic commitment to a specific regime (observed in Diatoms), or by the

maintenance of transcriptional flexibility (observed in Haptophyta, Chlorophyta and Pelagophyta). This catalog is thus an extremely useful resource to distinguish the strategies of any plankton group to adapt to environmental conditions when transcript regulation or gene copy number is implicated.

3.3. Next challenge: investigating the microbial omic *dark matter* of protists

Thanks to the advances of HTS and to the growing amount of environmental genomic datasets from various environments and of reference genomes/transcriptomes, 'classical' microbial genomic questions, such as *who is there ? what is done ? with whom ? in which conditions ? in which abundance ? in which intensity ?* can be tackled at a very broad and global scale. However as the majority of the current analyses (e.g. diversity analyses, trait-based analyses) are relying on a species name or on a function name, a significant part of the newly produced sequences in an environmental sample are ignored. Depending if one is looking at species or functions, and depending on the annotation strategy and / or on the sequence similarity threshold set up for the assignation, in average between 30 and 80% of the microbial environmental data are generally considered as « unknown » sequences (Sunagawa et al., 2015; Carradec et al., 2018; Bernard et al., 2018). This is particularly true for protists, for which even with metabarcodes studies focusing on the currently best referenced marker (i.e. 18S rDNA, Guillou et al., 2013), a significant amount of unknown lineages can still be found. For instance in 2015, the V9 metabarcoding global survey of the protistan planktonic fractions was highlighting between ¼ and ⅓ of taxonomically unknown OTUs (Figure 19). In 2017 these data were re-analyzed and by contrast, the most abundant V4 metabarcodes and OTUs from neotropical soil communities display only about 70% similarity with reference databases (Figure 20 (Mahé et al., 2017)).

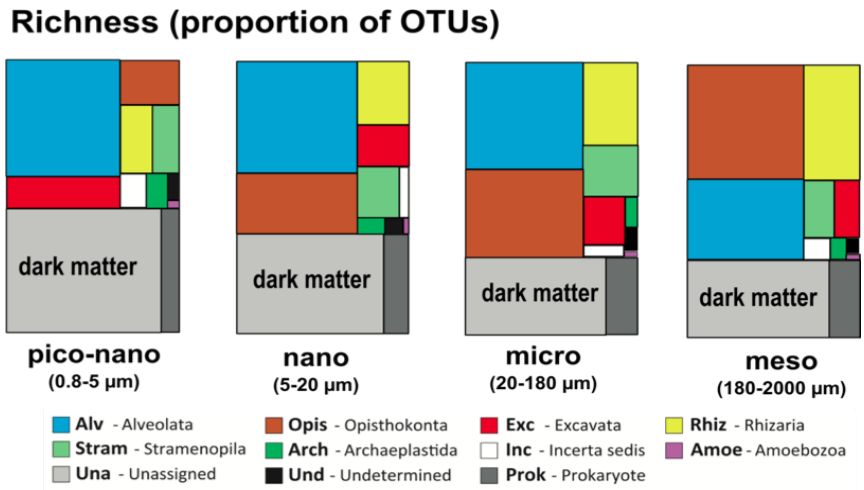


Figure 19 - Unknown and known components of eukaryotic plankton diversity (adapted from De Vargas et al., 2015). All Tara Oceans V9 rDNA reads (metabarcodes) were clustered in OTUs using Swarm (Mahé et al., 2014) and were then classified among the recognized eukaryotic supergroups, plus the known but unclassified deep-branching lineages (incertae sedis). The relative richness of the different eukaryotic supergroups in each organismal size fraction, from pico to meso, is here displayed. Note that ~5% of metabarcodes were assigned to prokaryotes, essentially in the piconano fraction, witnessing the universality of the eukaryotic primers used. Metabarcodes are "unassigned" (i.e. correspond to taxonomical microbial *dark matter*) when sequence similarity to a reference sequence is <80% and "undetermined" when eukaryotic supergroups could not be discriminated (at similarity >80%).

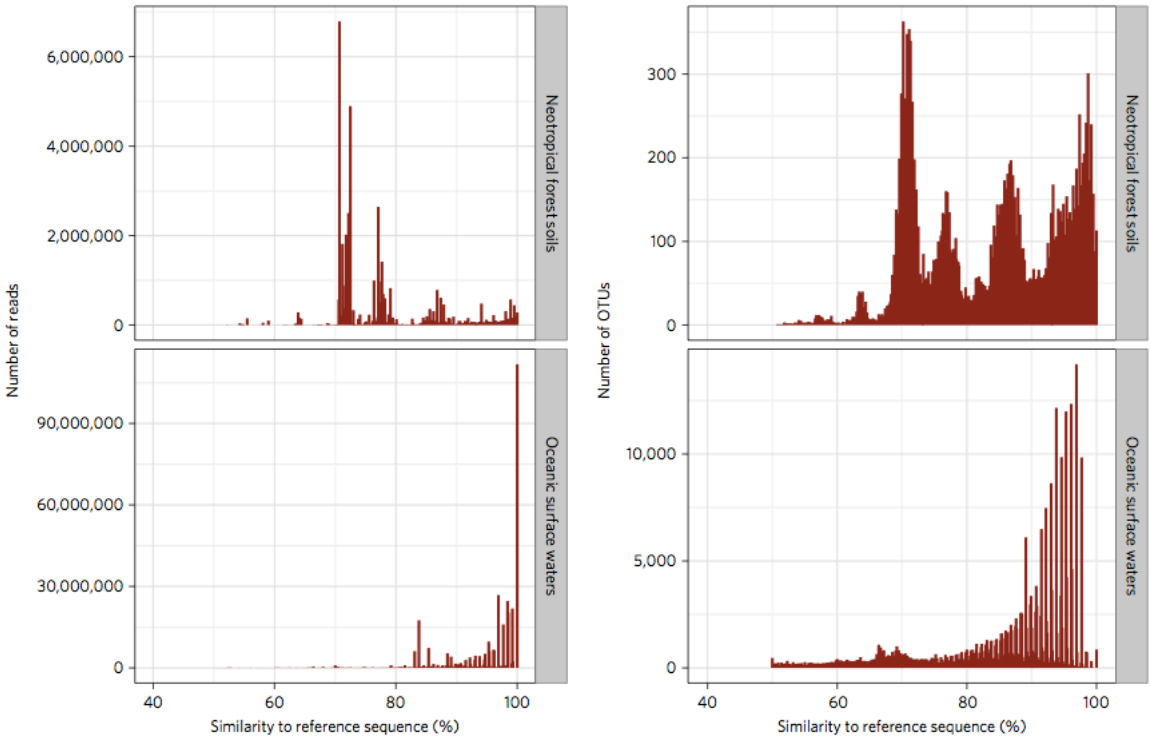


Figure 20 - Similarity of environmental protists to the taxonomic reference database (Mahé et al., 2017). In contrast to marine data, most of the reads and OTUs from the Neotropical rainforest soils (based on V4 metabarcodes) were <80% similar to references in the PR2 database. Only 8.1% of soil reads had a similarity ≥95%, whereas 68.1% of the marine reads from the Tara Oceans (based on V9 metabarcodes) study of the world's open oceans had a similarity ≥95%. The most abundant reads and OTUs in soils display only about 70% similarity with reference databases (n.b. by analogy, these highly abundant sequences belongs to the taxonomical microbial *dark matter*).

From the functional point view, despite its unprecedented sampling effort and sequencing depth, the 116 million eukaryotic unigenes catalog from Carradec et al. (2018) is not saturated, and 59.6% of (taxonomically and functionally) unknown unigenes were observed. A clustering of the unigenes, still reveals 3.26 million of highly expressed gene families in the global ocean, which do not correspond to defined protein domains. These new gene families are distributed relatively less globally than the known families, which might suggested that they correspond to genes that are necessary only in some conditions, potentially related to the adaptation of organisms to specific environments. However, the impressive number of genes without functionally characterized homologs in databases points to the large numbers of understudied yet widely distributed lineages inhabiting marine ecosystems and highlights the need to develop methods for revealing their roles in the ecosystems.

Producing more reference genomes and transcriptomes (Keeling et al 2014), improving culture methods and increasing the number of experimental model systems (Waller et al 2018), will of course help to reduce the proportion of unknown sequences, however these solutions remain expensive and very time-consuming. I am deeply convinced that is essential to develop methods which integrate the massive amount of environmental sequences in order to go beyond the biased view to the strain concept in phytoplankton ecology (Lakeman et al., 2009). In the following sections of this memoire, I will expose which bionformatics methods were and are currently developed and used min my team in order to mine and exploit the microbial (meta-)omic *dark matter* (Figure 21).

Figure 21 - The microbial omic *dark matter*. The "unknowns" of the microbial world (e.g., unknown genes, genomes, functions, organisms, processes, and communities associated with uncultured microbes) are often popularized under the catch-phrase "microbial *dark matter*" (Rinke et al., 2013). (A) When a new genome or transcriptome is sequenced, the functionally unannotated Open Reading Frames (ORFs) could be referred as genomic functional *dark matter* (DM). Here the figure from (Worden and Allen, 2010) highlights the proportion of functional DM in sequenced eukaryotic genomes. The compiled data for euKaryotic Orthologous Groups (KOGs) was obtained from the JGI. (**Total haploid gene count is probably incorrect, potentially skewing percentages). In most cases here, approximately 30% of the ORFs have no known function and another 10% are poorly characterized. Functional knowledge is strongest for core ORFs — usually central metabolic processes, and weakest for 'accessory ORFs'. (B) Classification of the "unknowns" sequences from meta-omic analyses (Bernard et al., 2018. Environmental sequences can be classified based on their taxonomical annotation (horizontal line) and their functional annotation (vertical column). The cells in purple (functional DM or taxonomical DM) and black (functional and taxonomical DM) correspond to categories that are not readily explained based on current biological knowledge. (C) Classification of a newly produced set of environmental sequences according to Antonio Fernandez-Guerra (Max Planck Institute, Bremen): the knowns are sequences which have at least a match in reference databases, the known unknowns are sequences which have a match with previous unknown sequences from other environmental project (e.g. the human gut and the ocean share unknown sequences (Wyman et al., 2018), the unknown unknowns correspond to sequences which have never been seen before.

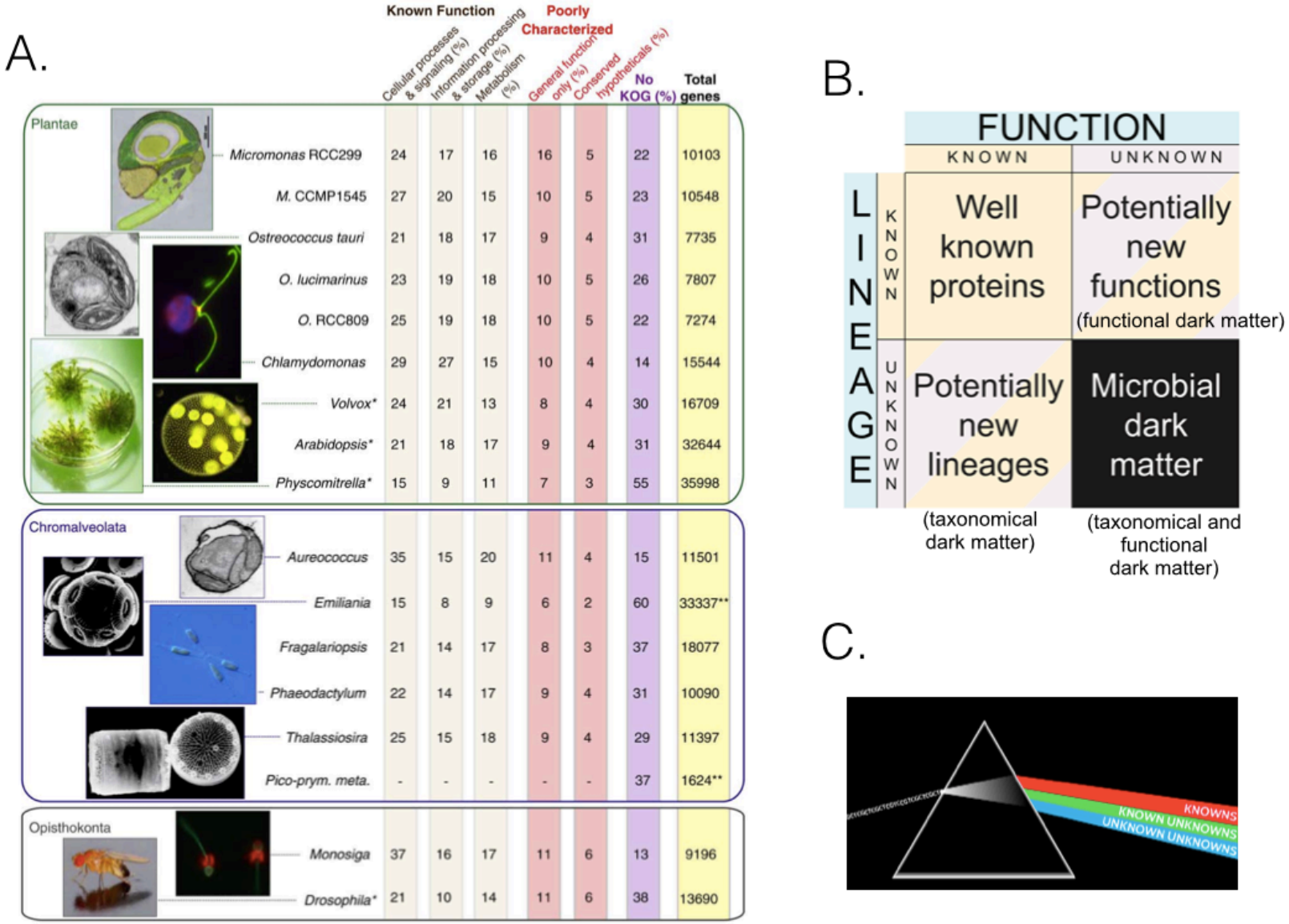


Figure 21 – The microbial omic *dark matter*.

4. Investigating the microbial omic *dark matter* of protists using networks

4.1. Introduction to Biological Networks

Networks are central to the understanding of biological systems. They are used to model cells development, communication and connectivity, to study the immune system or the brain, or even in genetics to build linkage maps among genes, and between genes and phenotypes. They are the basis for our understanding of transcriptional circuits and molecular signaling and as the structure of food webs and ecosystems. The recent and rapid development of social media added a new dimension to biological networks, notably through their charted and online quantification (Ideker and Nussinov, 2017). In the field of computational biology, the last decade has thus seen networks become a major mode of analysis. Networks provide both new data and a conceptual framework for computation. Networks are being generated in ever increasing sizes due to the advanced, the improving and expanding of experimental techniques. The accrual availability of large network data sets drives inherently the creation of bioinformatics methods to analyze these data to extract biological insights. On the other hand, networks correspond to a theoretical model for representing biological structure : i.e. a graph and the flow of information through this structure. In addition, since graphical models are core representations that arise in science (e.g. engineering or physics), they greatly unify the development of computer algorithms and their application across domains.

The ever-increasing amounts of data to be processed in the omic domain makes its comparative and exhaustive analysis increasingly challenging. Network analysis methods are powerful tools which can efficiently process these very large volumes. During the last several years, I used two types of biological network to explore the taxonomical and functional diversity of protists through the analysis of omics and meta-omics datatasets: (1) sequence similarity networks and (2) co-occurence networks.

4.2. Sequence Similarity Networks

With the advent of HTS 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 (Bittner et al., 2010; Forster et al., 2015; Lopez et al., 2015). SSN are indeed useful to visualize and explore genomic diversity to a very large scale, but also to study relationships between and within protein families (e.g. (Alvarez-Ponce et al., 2013; Atkinson et al., 2009; Bittner et al., 2010; Corel et al., 2016; Forster et al., 2015; Lopez et al., 2015; Méheust et al., 2016). Sequence similarity networks model a set of homologous sequences in the form of a graph in which the sequences correspond to the nodes and edges represent similarity relations between sequences (Figure 22, Figure 23).

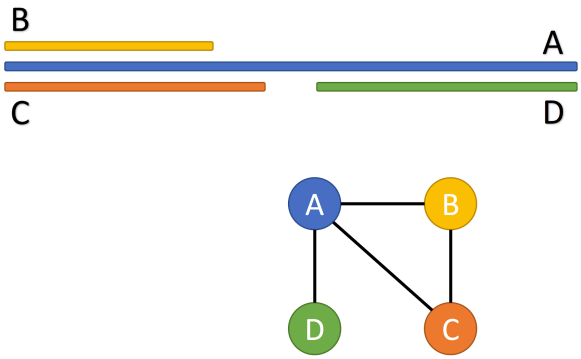


Figure 22 – Example of a sequence similarity network (A. Meng's Phd manuscript). In the upper part is represented a global alignment of the sequences A, B, C and D, and in the lower part, its corresponding SSN. The nodes correspond to the sequences (A, B, C & D) and the edges represent the alignment between each pair of sequences.

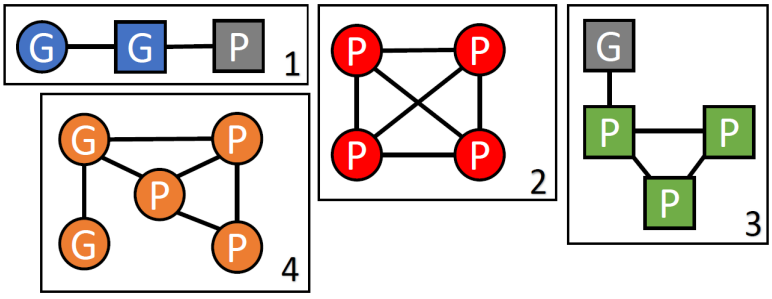


Figure 23 - Formalisation of four connected components (1, 2, 3 and 4) (A. Meng's Phd manuscript). The letters P and G correspond to two fictive taxa. The color of the nodes corresponds to fictive functional annotation (four distinct one here: blue, red, green and orange) or to the absence of functional annotation (grey color). The shape of the nodes (squared or round) corresponds to a fictive trait (e.g. a round corresponds to a sequence obtained from a species producing a toxin whereas the square correspond to a sequence obtained from a species which is not producing this toxin). This figure illustrate how (1) functionally unknown sequences can be linked to functionally known one, (2) the genomic bases of traits can be exhaustively investigated (while including functional *dark matter*).

To build SSN, similarity is computed between each pair of sequences (e.g. DNA, RNA, protein) from a dataset (Figure 24). Then the similarity value are screened (i.e. a rule to define a threshold is defined, e.g. 80% of sequence overlap and 80% of similarity or cf. Figure 24). Groups of related nodes or subgraphs or connected components (CCs) are thus defined, representing sub-communities of the network. The nodes and the edges of the SSN can be labelled. For instance, when available, nodes can be labelled with information on taxonomy, gene functions, organismal trait (Figure 23). Each of the CCs can be treated independently, which greatly speeds up calculations and provides the opportunity to explore issues specific to certain sequence populations (for instance CCs linked to a trait).

In the framework of a functional genomic study, 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). These methods facilitate the exploration of genomes or transcriptomes or proteomes composed of several thousands to millions of sequences. For example, in 2009 Atkinson et al. explore 3 super-families of proteins from the comparison of 773, 621 and 1330 sequences in the form of SSN. In 2016, Méheust et al. studied the cross-linked evolution in eukaryotic lines from SSN comparing 2,192,940 protein sequences.

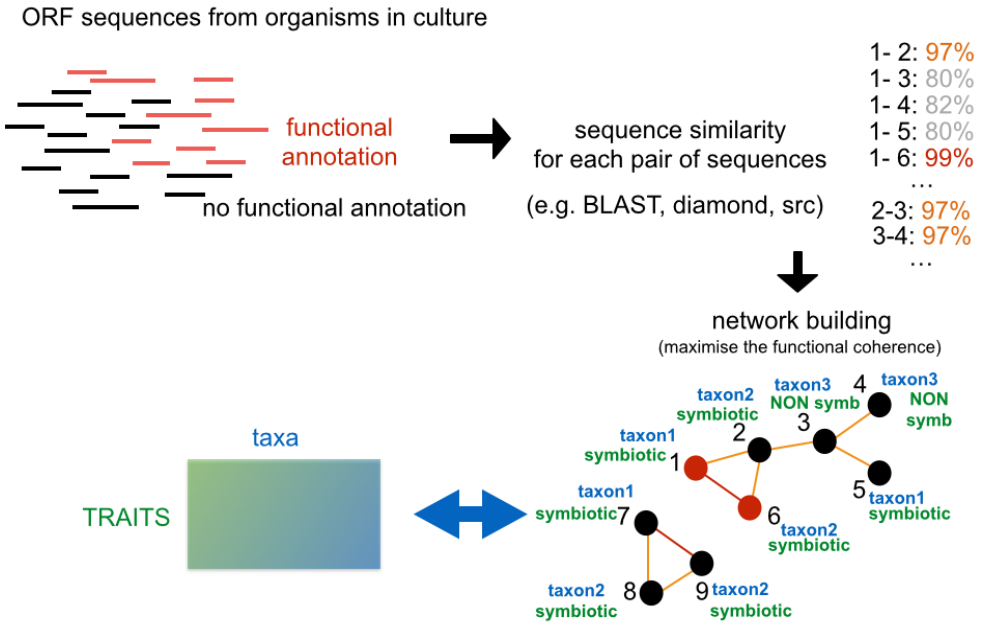


Figure 24 - SSN building and mining. To build a SSN, similarity is computed between each pair of sequences (e.g. here ORFs sequences). The similarity value are screened, for instance here, links were keep in the sequences overlap was at minimum of 80% and if the resulting subgraphs (CCs) were showing homogeneous functional annotation). Here the information of nodes were highlighted : taxonomic origin and information of their corresponding symbiotic trait.

Selected article # Article 1

Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome-based sequence similarity network

Arnaud Meng, Erwan Corre, Ian Probert, Andres Gutierrez-Rodriguez, Raffaele Siano, Anita Annamale, Adriana Alberti, Corinne Da Silva, Patrick Wincker, Stephane Le Crom, Fabrice Not †, Lucie Bittner †
(† co-senior authors)

2018, published in Molecular Ecology as an original article

Outline: In the framework of Arnaud Meng's PhD, we explored the functional genomics of dinoflagellates (i.e. protists which have enormous genomes (cf. table 1)) with a specific focus on the genomic bases of the symbiotic and toxic trait. The article *Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome-based sequence similarity network* was published in February 2018, and corresponds to the first article of Arnaud's PhD. The study was first only focused on four new transcriptomes produced by Dr. Fabrice Not, but in interaction with Arnaud, Fabrice and Erwan Corre, we conceived a larger study in order to involve all MMETSP dinoflagellate transcriptomes, which enable us to explore more exhaustively the functional diversity of this hyperdiverse lineage. This publication involves the *de novo* assembly transcriptomics pipeline from non model organisms (https://github.com/upmcgenomics/dntap) , which Arnaud conceived during his master 2 internship and his first PhD year, which I co-supervised with Dr. Stéphane Le Crom (n.b. this pipeline was used in (Botebol et al., 2017), in which Arnaud, Stéphane and I were co-authors). The bioinformatic improvements made by Anita Annamale during her master 1 internship in our team (cf. C.V.) were also implemented in the pipeline (https://github.com/upmcgenomics/PREMSEQ). To explore simultaneously the largest number of transcriptomes, I supervised Arnaud in the SSN analysis. The nodes of the SSN were labeled with information (e.g. taxonomy, traits of the organisms, and the functional annotation of the sequence). This approach constitutes the most comprehensive picture to date of the genomic potential of dinoflagellates and enabled to identify a core-predicted proteome composed of 252 connected components (CCs) of putative conserved protein domains (pCDs). Of these, 206 were novel and 16 lacked any functional annotation in public databases. Finally, the SSN enabled to explore the data in several dimensions and notably to extract CC specific to traits (as in Figure 24).

Meng et al. 2018, in Molecular Ecology, illustrates how SSNs are useful to perform large-scale comparative omics analyses, while exploiting the functionally unknown sequences. In this article, sequences were obtained from cultivated, known organisms, from which traits were defined. Consequently, we pointed out clusters of sequences or CCs, which constitute groups of homologous sequences (i.e. when CCs form cliques, they correspond to gene families) with additional potentially distant homologs (i.e. when CCs also involve chain in their structure). We highlighted CCs which are putative markers or proxies of organismal traits. These markers need to be classified to extract a top list of markers: i.e. 45 207 toxic CCs were observed, but the most shared CC (5 CCs, involving 2 totally functionally unknowns CCs) constitutes a reasonable top list of sequence (here ORFs) to further study in wet lab (e.g. via qPCRs) and which can be searched in environmental data.

Ophélie Da Silva, who did a M1 internship in our team (2 months of internship in June 2017, cf. section student supervision in CV), developed a script for our team (https://github.com/upmcgenomics/HomDistrib) which takes as input a fasta file of sequences of interest, and then searches their homologs in the *Tara* Oceans metagenomic and metatranscriptomic dataset. The **Figure 25** represents one of the output from this tool.

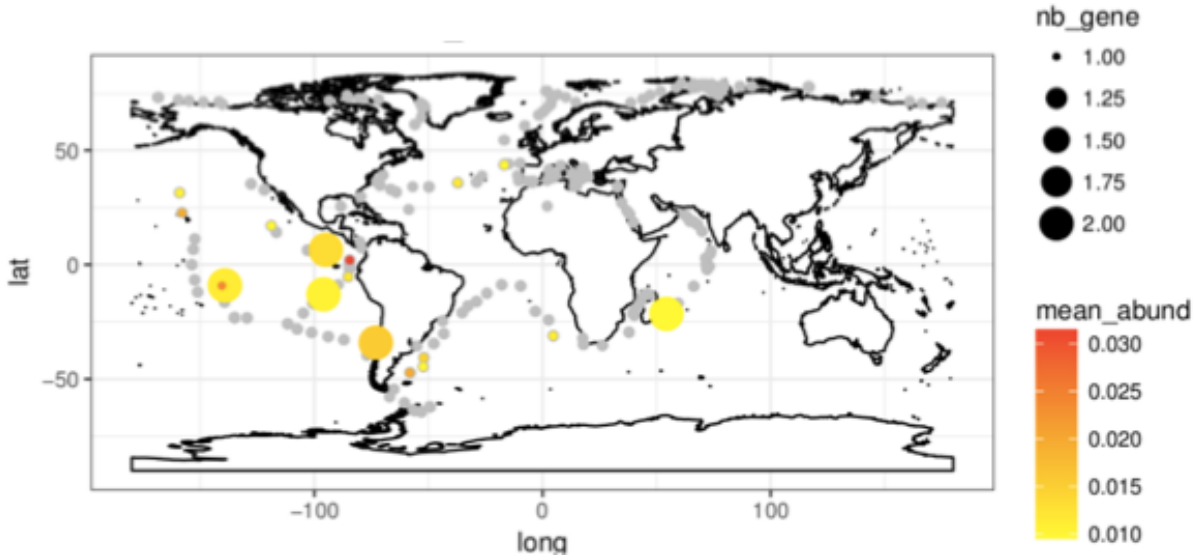


Figure 25 – Abundance of homologous sequences from a symbiotic CC only composed of functionally unknown sequences, found in the metatranscriptomics dataset (fraction 0.8-5 µm) of the *Tara* Ocean expedition.

The (homologous) sequences which are linked to a trait can be found in the environment, and their corresponding abundance matrix can be analysed in lights of abiotic and biotic parameters. One of the next step can be the analysis of the structure of the variants and the highlight of ecological niches. I have implemented this whole reasoning and analysis in

a tutorial course, which I have taught for the first time in October 2018 to masters students in microbiology at the Observatoire Océanologique de Banyuls-sur-mer.

In the framework of Arnaud Meng's Phd, a specific focus has been done on the symbiotic trait. His co-supervisor, Dr. Fabrice Not, who is a specialist of protistan symbioses, generated transcriptomes of several Rhizarian specimens, which are living in symbiosis with photosynthetic protists (presumably Dinoflagellates). Our goal was to de novo assembled these holobiont transcriptomes, and then to search for transcript markers of symbiosis by compiling in the same SSN transcripts from dinoflagellates (Meng et al. 2018 Molecular Ecology), and transcripts from some non-symbiotic Rhizaria (**Figure 26**).

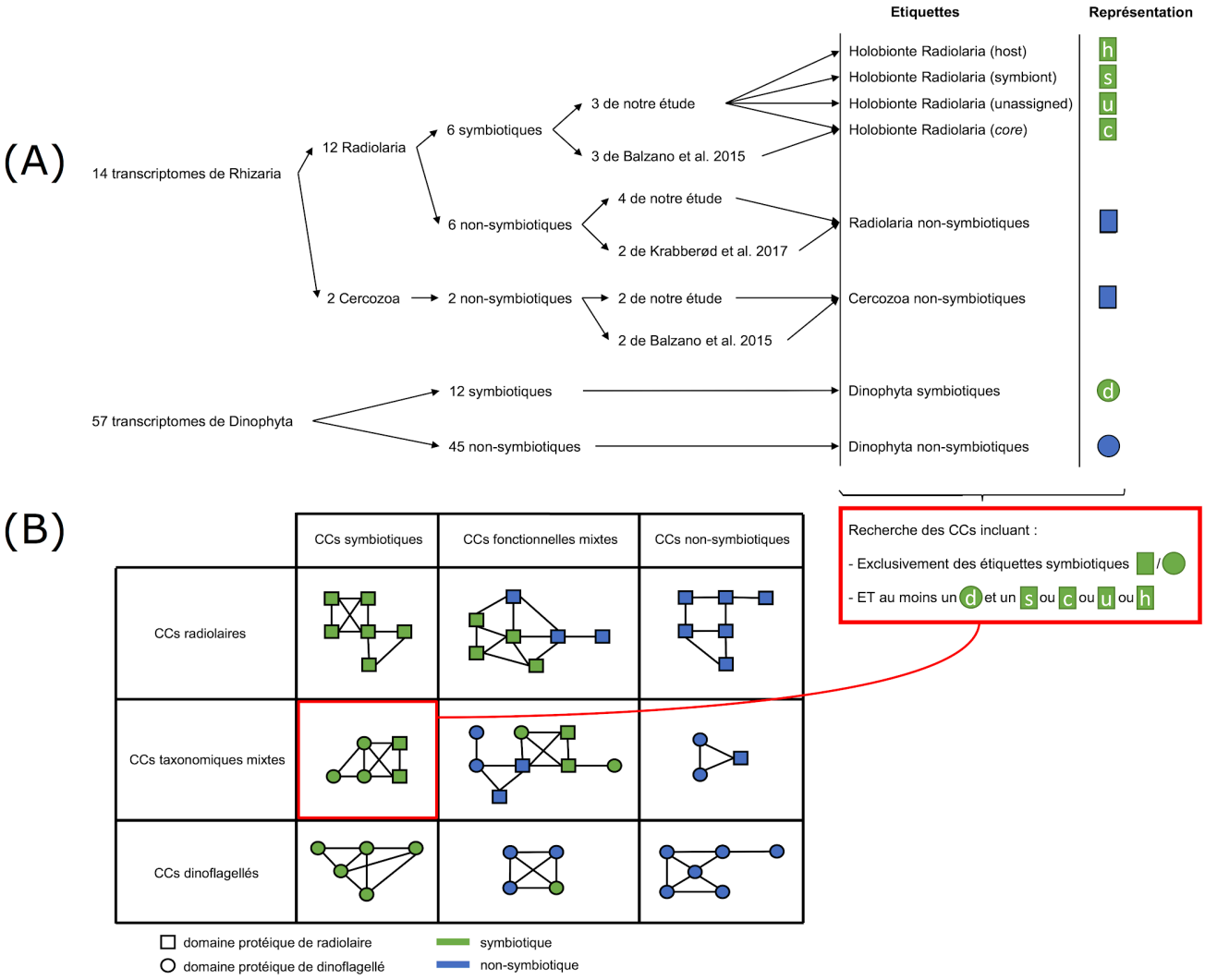


Figure 26 - Diagram of the dataset analysed in Meng et al. in prep (working title: Key functions involved in the establishment and the maintenance of marine plankton symbiosis revealed by a meta-transcriptome approach) (study of holobionts: the host is a Rhizaria, the symbionts are supposed to be dinoflagellates). (A) Datasets and corresponding flags which will be used in the SSN: transcripts from Rhizaria are squared (involving symbiotic (holobiont) and non symbiotic Rhizaria), the one from

dinoflagellates are round (Meng et al., 2018a). The presence of the symbiotic trait is indicated in green, its absence is indicated in blue. (B) different classes of CCs which will be obtained in the SSN. The class of CCs on which the analysis will mainly focus is circled in red (i.e. CCs only composed of transcripts obtained both in holobionts (symbiotic Rhizaria) and in dinoflagellates alone - the labels h/s/u/c refer to the methodology developed in (Meng et al., 2018b)).

In order to minimize the proportion of chimeric transcripts obtained from the de novo assembly of the holobionts transcriptomes (i.e. holobionts correspond to Rhizaria and presumably dinoflagellate symbionts), we debated with Arnaud about a strategy to overcome this bioinformatics challenge. Concurrently, in October 2016, Dr. Pierre Peterlongo invited me to visit the GenScale team at INRIA in Rennes (https://team.inria.fr/genscale/). We discussed and I explained that I was dreaming about a tool which would allow me to calculate similarity between pair of sequences at a very broad scale (e.g. for the *Tara* Oceans catalogs: 40 M of genes (Sunagawa et al., 2015) and 116 M of unigenes (Carradec et al., 2018)) in order to build the most inclusive as possible SSNs. I also exposed them our de novo assembly issue from holobionts transcriptomes. From our discussions, they developed the highly scalable tool Short Read Connector (SRC) (Marchet et al., 2018)° (https://github.com/GATB/short_read_connector). SRC is a k-mer based similarity method which relies on a very lightweight data structure called a quasi-dictionary that enables to work with voluminous sequence sets. SRC enables thus to estimate the similarity between numerous (meta-)omic datasets by extracting their common sequences. Camille Marchet (PhD who was supervised by Pierre Peterlongo) and Arnaud interacted and this collaborative work conducted to a publication in the journal Microbiome in 2018.

Selected article # Article 2

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

Arnaud Meng †, Camille Marchet † († co-first authors), Erwan Corre, Pierre Peterlongo, Adriana Alberti, Corinne Da silva, Patrick Wincker, Eric Pelletier, Ian Provert, Johan Decelle, Stéphane Le Crom, Fabrice Not, Lucie Bittner

2018, published in Microbiome as a research article

Outline: In order to minimize the proportion of chimeric transcripts obtained from the *de novo* assembly of non-model holobiont transcriptomes, we designed an innovative bioinformatic strategy and tested it on marine models as a proof of concept. We considered three holobiont models and sorted their raw reads using Short Read Connector (SRC), a k-mer based similarity method (Marchet et al., 2018)°. Before assembly, we thus defined four distinct categories for each holobiont metatranscriptome: host reads, symbiont reads, shared reads, and unassigned reads. Afterwards, we observed that independent de novo assemblies led to a diminution of the number of chimeras. Moreover, the separation of each partner's transcriptome offered the independent and comparative exploration of their functional diversity in the holobiont. Finally, our strategy allowed to propose new functional annotations for two well- studied holobionts (a Cnidaria-Dinophyta, a Porifera-Bacteria) and a first metatranscriptome from a planktonic Radiolaria-Dinophyta system forming widespread symbiotic association for which our knowledge is considerably limited. In conclusion, in contrast to classical assembly approaches, our strategy generates less chimera and allows biologists to study separately host and symbiont data from a holobiont mixture. The pre-assembly separation of reads using SRC is an effective way to tackle metatranscriptomic challenges and offers bright perpectives to study holobiont systems composed of either well-studied or poorly characterized symbiotic lineages and ultimately expand our knowledge about these associations.

4.3. Co-occurrence or association networks

Organisms do not exist in isolation but form complex ecological interaction webs. Interactions within these ecological webs can have a positive impact (that is, a win), a negative impact (that is, a loss) or no impact on the partners involved. The possible combinations of win, loss and neutral outcomes for two interaction partners allow the classification of various biotic interaction types (Figure 27).

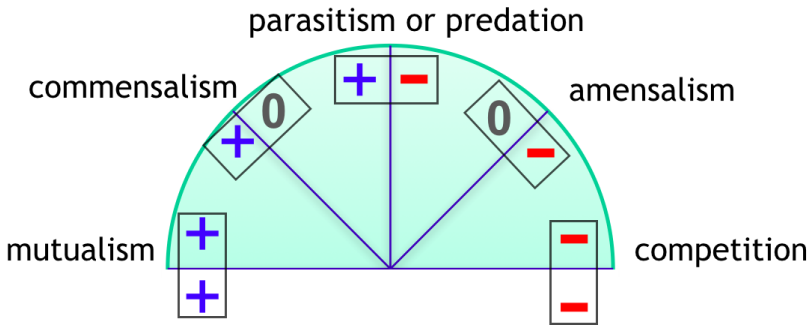


Figure 27 –Summary of ecological interactions between partners of different lineages (Faust and Raes, 2012). All possible pairwise interactions are here summarized. For each interaction partner, there are three possible outcomes: positive (+), negative (–) and neutral (0). For instance, in the parasitism relationship (+ –), the parasite benefits from the relationship (+), whereas the host is harmed (–).

During the last decade, it has been clearly shown that microbial communities structure and function are heavily influenced by abiotic interactions (i.e. environmental conditions) but also by biotic interactions (e.g. microbe–microbe interactions, microbe–host interactions) (e.g. (Chaffron et al., 2010) (Lima-Mendez et al., 2015)°). In this way, understanding the microbial communities as a whole, including the complex interplay among microbial taxa, is becoming a routine in microbial ecology. Network theory, in the form of systems-oriented, graph-theoretical approaches, is an exciting holistic methodology that enhances the understanding of the complex ecological and evolutionary processes of microbial communities. Using network theory, one can model and analyze a microbial communities and all its complex interactions in a single network.

Hypotheses about interactions may be derived from co-occurrence or association networks, built from the analysis of OTUs or genes sampled across different locations, replicates or time points. Co-occurrence networks building techniques can rely on pairwise Pearson on Spearman (e.g. (Arumugam et al., 2011; Barberán et al., 2012)), local similarity analysis (LSA; e.g. (Durno et al., 2013)), compositionality-robust estimation of correlations (e.g. SparCC; (Friedman and Alm, 2012), CCLasso; (Fang et al., 2015)), Gaussian graphical models (e.g. (Kurtz et al., 2015)), sparse regression (e.g. (Faust et al., 2012)), or assessment of co-

occurrence probability with the hypergeometric distribution for presence/absence data (e.g.(Chaffron et al., 2010)). Depending on the techniques, co-occurrence networks are built on presence/absence or abundance data (Figure 28). In the resulting networks, nodes correspond to OTUs (or lineages) or more rarely functions, and the edges represent the significant relationships between them (e.g. either only positive, or positive and negative).

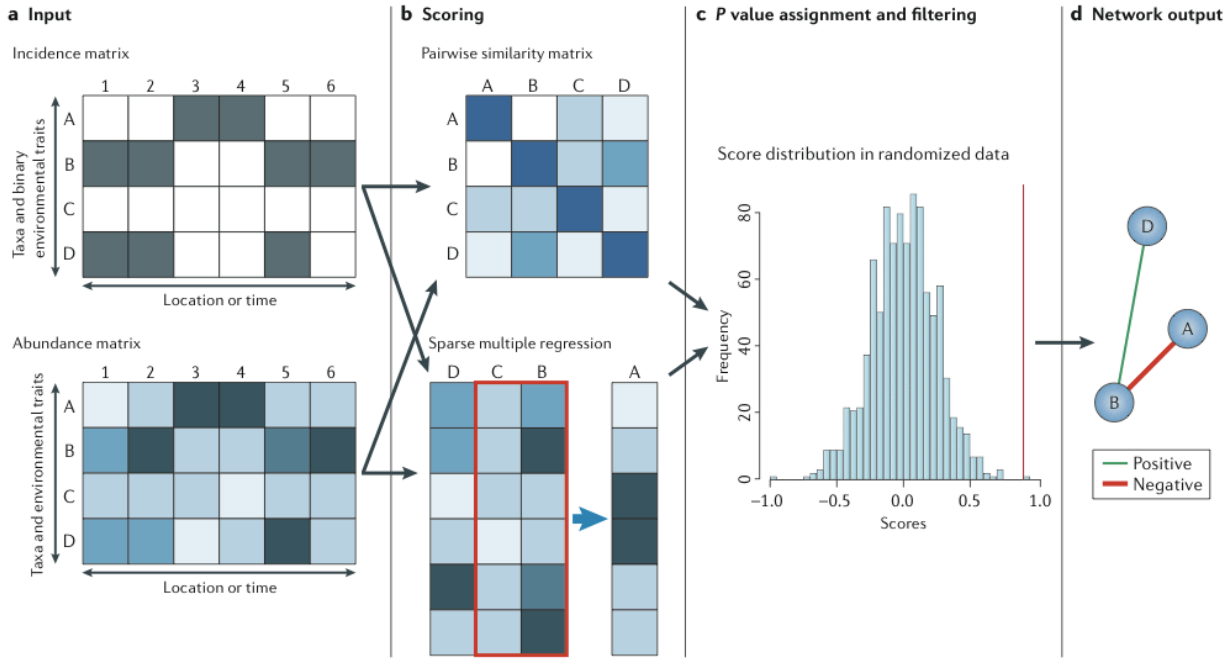


Figure 28 - Principle of association network building based on similarity- and regression-based network inference (Faust and Raes, 2012). The goal of network inference is here to identify combinations of microorganisms that show significant co-presence or mutual exclusion and to combine them into a network. (a) Network inference starts from an incidence (presence/absence) or an abundance matrix, both of which store observations across different samples, locations or time points. (b) Pairwise scores between taxa are then computed using a suitable similarity or distance measure (e.g. Pearson, Spearman, hypergeometric distribution, Jaccard index). In contrast to similarity-based approaches, multiple regression can detect relationships that involve more than two taxa. To reduce over-fitting, sparse multiple regression is usually carried out — that is, the source taxa subset that best predicts the target taxon's abundance is selected. In addition, the regression model is cross-validated: that is, after regression coefficients have been identified with a training data set, the model's prediction accuracy is quantified on a test data set. (c) In the next step, a random score distribution is generated by repeating the scoring step a large number of times (i.e. 1,000 times or more). The random score distribution computes the P value (that is, the probability of obtaining a score by chance that is equal to or better than the observed score) to measure the significance of the predicted relationship. The P value is usually adjusted for multiple testing with procedures such as Bonferroni or Benjamini–Hochberg (Benjamini & Hochberg 1995). (d) Taxon pairs with P values below the threshold are visualized as a network, where nodes represent taxa and edges represent the significant relationships between them.

The network properties — such as modularity, degree, closeness, betweenness distribution, assortativity, or the average path length, nestedness — are often calculated and ecologically interpreted (n.b. the same metrics can be calculated with SSNs, e.g. (Forster et al.,

2015)°. Through the use of incidence or abundance matrices obtained from NGS, researchers can go beyond the knowledge based actors from classical interaction networks and food webs, and can now exhaustevely describe the potential interactions within the microbial communities. The matrices correspond to sequences (corresponding to e.g. OTUs/lineages and / or to clusters of genes / functions) which might involved known and unknown actors, so the networks might involved many dark matter nodes. By mining the numerous interactions from the association network and the properties of the nodes and edges, one can retrieve known interactions and one can predict new ones (e.g. (Lima-Mendez et al., 2015)°, (Mordret et al., 2016), (Vincent et al., 2018)°) as well as suggesting new hub actors (e.g. (Qin et al., 2010) (Guidi et al., 2016)°). Network association analyses can be used to determine drivers in environmental ecology (e.g. (Steele et al., 2011) (Lima-Mendez et al., 2015)°) or contribution to habitat niches or disease (e.g. (Arumugam et al., 2011; Chaffron et al., 2010)). Since approximately a decade, studies are multiplying thanks to the growing amount of meta-omic datasets (n.b . the increasing number of samples are improving the statistical power of the methods), and microbial association networks have been inferred for a range of communities from soil (e.g. (Barberán et al., 2012; Mandakovic et al., 2018), ocean (e.g. (Guidi et al., 2016; Lima-Mendez et al., 2015)°) and human body communities (e.g. (Qin et al., 2010)). It is however interesting to note that interpreting these networks is not always straightforward, and the biological implications of network properties are unclear. Only few articles review thus far the factors that can result in spurious predictions for that kind of analysis. A major problem is that inference are now made from observational data, whereas the few simulations, which have been published so far, show that network properties are affected by tool choice and environmental factors (Röttjers and Faust, 2018; Weiss et al., 2016). For example, hub species might not be consistent across tools, and environmental heterogeneity induces modularity. In the coming years, it is likely that the field of co-occurrence network inferences will be enriched and strengthened by the recommendations which will arise from more simulations, tools benchmark and cross biome comparisons.

During the last five years, I co-authored 3 articles involving co-occurrence networks analyses (Lima-Mendez et al., 2015 ; Guidi et al., 2016 ; Vincent et al., 2018)°. In Guidi et al. (2016), we (Dr Lionel Guidi, Dr Samuel Chaffron, Dr Damien Eveillard and I) revisited the issue of the Biological Carbon Pump on the global ocean - which is a 'classical' subject for biogeochemists and oceanographers - in lights of meta-omics datasets (cf. annexe 2). From the omics abundance matrices of OTUs and functions generated in 2015 by the first large-scale Tara Oceans articles (Brum et al., 2015; Sunagawa et al., 2015; Vargas et al., 2015), we

built co-occurrence networks using the WGCNA method (Aylward et al., 2015; Langfelder and Horvath, 2008) and extracted modules (or sub-communities or sub-graphs, i.e. here corresponding to co-abundant sequences across the samples) linked to carbon export. The entire community of planktonic actors involved in this process was highlighted (involving eukaryotes, prokaryotes and viruses) as well as their hub nodes (**Figure 29**).

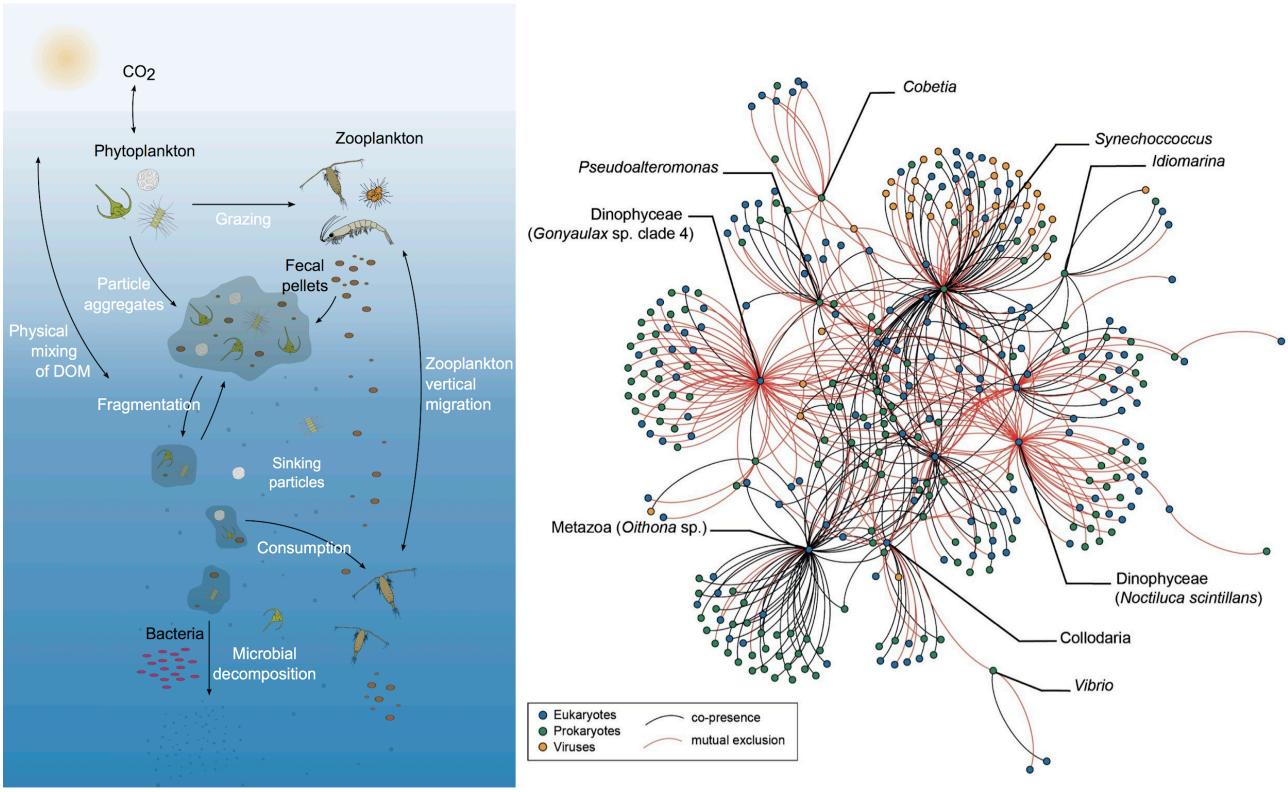


Figure 29 – Classical and omics view of the Biological Carbon Pump. On the left, 'classical' vision of the actors and mechanisms involved in the biological carbon pump in the ocean (figure from(Benoiston et al., 2017)). On the right, interactome of planktonic communities involved in carbon export (network nodes correspond to organisms, and edges to copresence (black) or mutual exclusion (red) relationships) (fig. from Guidi et al., 2016). This integrated network was built from the selection of the VIP nodes in the eukaryotic, prokaryotic and viral subnetworks related to carbon export at 150 m. Co-occurrences between all lineages of interest were extracted, if present, from a previously established global co-occurrence network (Lima-Mendez et al., 2015). The resulting graph is composed of 329 nodes, 467 edges, with a diameter of 7, and average weighted degree of 4.6. The key nodes (whose taxonomic affiliation is reported) correspond here to the hubs of the network.

From the study of the orthologous genes (OGs) of the pico-planktonic fractions, two modules (FNET1 and FNET2) were highlighted as associated with the carbon export process (**Figure 30**). The relevance of these OGs to predict carbon export was confirmed by PLS regression. FNET1 and FNET2 predict 41% and 48% of carbon export variability respectively, with a minimal number of functions (123 and 54 OGs with a VIP score >1 corresponding to the best OGs predictors, for FNET1 and FNET2, respectively). FNET1 involves functions linked to

photosynthesis and growth and FNET2 involves functions linked the formation and degradation of marine aggregates (**Figure 30**). Furthermore, 77% and 58% of the best OGs predictors in FNET1 and FNET2, respectively, are functionally uncharacterized, pointing to the strong need for future molecular work to explore these functions and the tremendous potential of co-occurrence network analyses to reveal the relevance from *dark matter* sequences in biogeochemical processes.

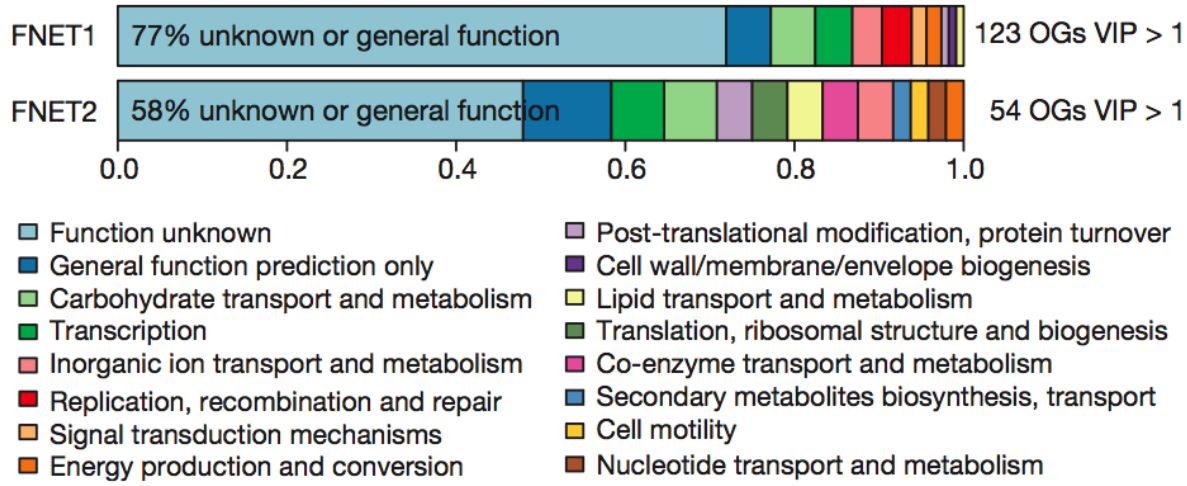


Figure 30 – Key bacterial functional categories associated with carbon export at 150 m (Guidi et al., 2016). A bacterial functional network was built based on orthologous group/gene (OG) relative abundances using the WGCNA methodology (see Methods) and correlated to classical oceanographic parameters. Two functional subnetworks (FNET1 (n = 220) and FNET2 (n = 441)) are significantly associated with carbon export. Higher functional categories are depicted for functions with a VIP score >1 in both subnetworks. One module (FNET1) contains many functions specific to the Synechococcus accessory photosynthetic apparatus (e.g. function related to phycobilisomes, phycocyanin and phycoerythrin), as well as functions related to carbohydrates, inorganic ion transport and metabolism, and transcription, suggesting overall a subnetwork of functions dedicated to photosynthesis and growth. The second module (FNET2) contains several functions encoded by genes taxonomically assigned to Candidatus pelagibacter and Prochlorococcus, known as occupying similar oceanic regions as Synechococcus, but overall most of its relative abundance (74%) is taxonomically unclassified. Top VIP scoring functions in FNET2 are membrane proteins and ABC-type sugar transporters, as well as functions involved in carbohydrate breakdown such as a chitinase . These features highlight the potential roles of bacteria in the formation and degradation of marine aggregates.

This reassessment of the study of the Biological Carbon Pump (BCP) is currently being pursued by Anne-Sophie Benoiston (cf. students supersing section in the CV), a PhD co-supervised by Dr. Lionel Guidi and myself. She is repeating the omic-based co-occurrence network analyses with an extended dataset (more samples have been sequenced since 2010, and e.g. new oceanic regions are now available such as the Arctic zone) while using another tool (i.e. using SPIEC-EASI, (Kurtz et al., 2015)). She is comparing modules related to three quantifiable processes or states of the BCP (i.e. net primary production, carbon export,

remineralisation). She is exploring if the community of lineages and functions are driven by the same VIP actors, and as common actors can be found, she is examining the potential changes in network structure from one BCP state to an other by using a graph alignment method (Malod-Dognin and Pržulj, 2015; Mandakovic et al., 2018). Anne-Sophie is currently writing an article highlighting her new results, which will be submitted to the ISME Journal in the coming months.

Finally, Marie Soret, who did a M2 internship in our team (6 months of internship in 2018, cf. section student supervision in CV), tested and compared different machine learning algorithms in order to highlight the most efficient one(s) to identify the best predictors (i.e. here omic sequences) of the BCP states. Our goal was to be able to predict from a sample its contribution to one of the BCP state just based on omic data. The use of machine learning algorithms in order to exploit the growing amount of omics and / microbial datasets is a hot topic (e.g. (Cordier et al., 2018; Farrell et al., 2018; Perkel, 2018; Sauzède et al., 2017; Tackmann et al., 2018)) and the identification of microbial biomarkers and their use for classification tasks have promising applications in the field of microbial ecology and ocean monitoring (Armbrust, 2014; Bohan et al., 2017; Ottesen, 2016). The results obtained by Marie will thus be developed in a manuscript, which will be submitted at the latest at spring 2019.

5. Perspectives: next research and developments - ongoing and future collaborations

In the next coming years, I will keep on working at the interface between environmental genomics, bioinformatics, ecology and adaptation, I will keep on learning modelisation from my collaborators, and if possible, in the very near future, I wish to reconnect more concretely with evolutionary questions. Thanks to my very nice current collaborations and certainly thanks to additional new ones, I intend to continue and develop the three main research axes detailed in the following perspectives.

Perspective 1: exploring the microbial omic *dark matter* at the global scale

In the coming years, I will keep on submitting projects and applying for grants aiming at studying the microbial omic dark matter (nb. I have submitted until now 5 grants (e.g. ANR, ERC starter, CNRS call, Sorbonne Université call) but none of them was successful, even if some of them reached the second round). I thus intend to keep on developing strategies to exploit already available but under-studied large meta-omics datasets , with a specific focus on protists lineages, by using both sequence similarity networks and co-occurrence networks (an example of overarching analysis strategy can be found in annexe 3, project submitted at the ANR2017 call). The global analyses of the largest number of transcriptomes and metatranscriptomes of eukaryotes involving their dark matter, will help to study and to better understand the evolutionary and functional adaptations of protists in their environment. I have already conducted tests on the eukaryotic (uni)genes catalogs (Carradec et al 2018), and thanks to the tool SRC developed by Dr. Pierre Peterlongo and Dr. Camille Marchet (Marchet et al, 2018°), a SSN was built with the 116 million of unigenes. In collaboration with the Dr. Eric Pelletier (Genoscope, CEA, Evry, France), we intend to continue this research, and we notably wrote a PhD subject in the framework of an ITN project lead by Dr. Sakina-Dorothée Ayata (ITN dedicated to the study of the plankton functions and services in marine ecosystems), which will be submitted in January 2019. From this work, a most wanted list of conserved microbial protein families with no known domains will be obtained (as Wyman et al. 2018 have done for prokaryotic fractions). Interactions with Dr. Antonio Fernandez-Guerra (currently working at the Max Planck Institute, Bremen), who is conducting research on prokaryotic fractions with similar questions, will be set up. Finally, one of our other major objectives will be to detect clusters of

dark matter sequences which will play a major role at the global oceanic scale or in certain niches, and which can be used as proxies to monitor marine ecosystems (Armbrust 2014, Bohan et al 2017, cf. annexe 3 and 4).

Perspective 2: from omics to modelisation via the use of traits

From known organisms transcriptomes and genomes, sequences can be suggested as biomarkers of organism traits (cf. methodology developed in Meng et al 2018, Mol ecol), and homologs of these sequences can be found in the environment. From meta-omics datasets, sequences can also be suggested as predictors of a biogeochemical process (Guidi et al 2016°, and on going work from Anne-Sophie Benoiston detailed in section 4.3., pp85-86).

A validation work of a top list of these markers should be considered. Dr. Raffaele Siano (researcher at Ifremer, Brest, France), who co-authored Meng et al 2018, is considering to test in vivo markers of toxicity on dinoflagellates cultures. A SSN study performed by Quentin Letourneur (M1 internship during spring 2016, cf. student supervision section in CV) on Ostreococcus strains revealed a list of transcript specific from low iron adaptation conditions. Our collaborator Dr. François-Yves Bouget (DR CNRS, Observatoire Océanologique de Banyuls-sur-mer) intended to confirm in vivo the expression of these markers. My collaboration with Dr. Fabrice Not will continue to explore the genomic bases of microbial symbioses, notably through the study of Rhizarian holobiont transcriptomes and the writing of Dr. Arnaud Meng's last PhD article. The expertise of Dr. Fabrice Not is also very useful to guide Emile Faure (PhD grant from october 2017 to septembre 2020, co-supervised by Dr. Sakina-Dorothée Ayata and my-self, cf. student supervision section in CV) to explore the impact of mixotrophic organisms (on notably the symbiotic one) on biogeochemical cycles. Emile studied the environmental diversity and structure of mixotrophic protist lineages during his master 2 internship and his first PhD year based on metabarcoding data. In the coming months, he will extend his research to metagenomics species (MGS or Co-Abundance gene Groups (CAGs) (Nielsen et al. 2014)) and to metagenome-assembled genomes (MAGs, Parks et al. 2017), which will enable to build a more exhaustive and accurate picture of mixotrophy in the global ocean, while also integrating taxonomical and/or functional unknown sequences. SSNs and gene-based predictive models (Burns et al. 2018) will be used in order to predict mixotrophic markers. These research focussed on the genomic bases of mixotrophy will be part of emerging collaborations, which I am starting to initiate at the national scale (cf. annexe 4, project MixOmics submitted at the

EC2CO call in September 2018), and which will be continue with European partners (I will also submit a Working Group project at the next EuroMarine call in 2019). The goal of these projects and collaborations will be to better characterize the role of mixotrophy in marine ecosystems and marine biogeochemistry, particularly its impact on the carbon cycle, from the individual/species to the community/ecosystem scale. To do this, bridges between omics/bioinformatics, experimentation and modeling will be created, in order to maximize the interactions between fields. In this way, mixotrophic markers defined by in silico analyses will be then (1) tested by experimentalists and (2) will be used by modelers in order to build species-centered and also global / large scale biogeochemical models (collaborators at the LOCEAN in Paris (Dr. Olivier Aumont), at the LOPS (Dr. Thomas Gorgues) and at the PELAGOS (Dr. Marc Sourisseau) in Brest). I wouldn't have been able to develop this research axis (i.e. the impacts of planktonic diversity and trophic mode / strategy on oceanic biogeochemical cycles by integrating high-throughput sequencing data into marine ecosystems models) without the input and discussions from my collaborator Dr. Sakina-Dorothée Ayata (cf. section research projects in CV).

Perspective 3: from microbial omic *dark matter* to evolutionary questions

The outcomes from perspective 1 and 2 will contribute to a more inclusive and accurate picture of microbial diversity and functioning within ocean ecosystems, and will ensure solid bases for (i) understanding the adaptation processes of microbial eukaryotes in their environments ; and (ii) improving modeling studies of ecosystem carbon dynamics or fluxes in general. Such studies aim to renew and to become references for future microbial environmental surveys and for "ocean ecosystemics". Furthermore, the exhaustive omic study of organisms (here mainly protists) in their real environment, without a priori, i.e. without the constraint of a metazoa-centric or bacteria-centric vision, will help to build their own conceptual adaptive and evolutionary framework (and I am very concerned about this point since my PhD, cf. the foreword section). In this vein, during the coming years, I intend to turn a few ideas and discussions into research projects tackling different evolutionary aspects of protists: e.g. the mechanisms of Dinoflagellates genome evolution with Dr. Laure Guillou (DR CNRS, Station Biologique de Roscoff), the quantification of recombination events in the environment with Dr. Olivier Jaillon (Genoscope, CEA, Evry, France).

Final words

As this memoire hopes to demonstrate, we (environmental genomicists) are living a tremendous time: the omics data stream is still exponentially increasing and there is a critical need of new and original methods to explore them and to address ecological and evolutionary relevant questions. New thinking frameworks are also possible, which can emerge from multidisciplinary approaches and synergies created between molecular biology, bioinformatics, microbiology, plankton ecology, marine biogeochemistry, oceanography, numerical ecology, and system biology. Interactions, discussions, ideas sharing and transfer, training and transmission will be the keys from fulfilling environmental geneticists, which makes - in my opinion - the future looks always more exciting and pleasant.

Every day, searching, teaching and mentoring is a sincere pleasure. I even more realized it when I listened to the very inspiring Dr. Louis Legendre, a few years ago in Villefranche-sur-mer, who really put words on my every day feelings.

> *"Discoveries are not made by committees and do not result from accretion of knowledge. Discoveries are products of the imagination of creative researchers."*
>
> *"The process leading to discovery requires a pertinent question, which strongly involves intuition, the ripeness of time and creative imagination; the latter combines intuition, the scientific method and pleasure"*
>
> Louis Legendre, in *Scientific Research and Discovery : Process, Consequences and Practice*, 2004

Glossary

16S rDNA: genes encoding the RNA of the small ribosomal subunit (SSU) found in all Eubacteria and Archaea

18S rDNA: genes encoding the RNA of the small ribosomal subunit (SSU) found in eukaryotes. Many copies are found per genome. They are highly expressed and their nucleotide structure combine well-conserved and variable regions. It is the most widely used eukaryotic phylogenetic marker. Because of these characteristics, and for pragmatic reasons, 18S rDNA has been used as a marker to identify and barcode eukaryotes at the species or genus level (with some exceptions).

28S rDNA: genes encoding the RNA of the large ribosomal subunit (LSU) found in eukaryotes.

454: common term for the Roche GS platforms that use bead emulsion methods. It is a next-generation sequencing / NGS / HTS method, classified in the second generation. Reads measures on average 350–450 bp. Homopolymer errors are characteristics from this technology.

Accessory genome: (or accessory genome, also: flexible, dispensable genome) refers to genes not present in all strains of a species. These include genes present in two or more strains or even genes unique to a single strain only, for example, genes for strain specific adaptation such as antibiotic resistance.

Alveolata: a widespread group of unicellular eukaryotes that have adopted diverse life strategies such as predation, photoautotrophy, and intracellular parasitism [29]. They include some environmen- tally relevant groups such as the Syndiniales, the Dinoflagellata, and the ciliates (Ciliophora), as well as the Apicomplexa group that contains notorious parasites such as *Plasmodium* sp. (the agent of malaria), *Toxoplasma* sp. (the agent of toxoplasmosis), and *Cryptosporidium* sp.

Amoebozoa: this group consists of amoeboid organisms, most of them possessing a relatively simple life cycle and limited morpholo- gical features, as well as a few flagellated organisms. They are common free-living protists inhabiting marine, freshwater, and terrestrial environments. Some well-known amoebozoans include the causative agent of amoebiasis (*Entamoeba histolytica*) and *Dictyostelium* sp., a model organism used in the study of the origin of multicellularity.

Animal: cf. the definition of Metazoa. Often refers to an heterotrophic cell / organism.

Archaeplastida: also known as 'the green lineage' or Viridiplantae, this group comprises the green algae and the land plants. The Archaeplastida is one of the major groups of oxygenic photosynthetic eukaryotes [31]. Green algae are diverse and ubiquitous in aquatic habitats. The land plants are probably the most dominant primary producers on terrestrial ecosystems. Both green algae and land plants have historically played a central role in the global ecosystem.

Assortativity: measure of the preferential connection between a set of nodes of interest in a graph or in other words, assortativity quantifies to what extent sequences with the same label (for example, with the same taxonomy) connect with each other rather than with differently labelled sequences.

Average shortest path length (AL): AL is calculated as the average number of steps in the shortest paths between each node to each other node in a network. Networks with a small AL are also known as small-world networks. Microbial association networks have a mostly small AL. A small AL has been interpreted to increase the speed of the network's response to perturbations.

Betweenness: node centrality in a graph, which measures the extent to which a node lies on paths between other nodes. Nodes with high betweenness may have considerable influence within a network by virtue of their control over information passing between others. They are also the ones whose removal from the network will most disrupt communications between other nodes because they lie on the largest number of paths taken by messages.

Biological Carbon Pump (BCP): process by which CO_2 is transformed to organic carbon via

photosynthesis, exported through sinking particles, and finally sequestered in the deep ocean.

Carbon export : component of the BCP, carbon which is export from the surface layer to the deep ocean (i.e. sinking particules and aggregates (e.g. dead organic matter of faecal material) which reaches the sea floor where it is sequestred for long geological times).

Clique : in a graph/network, subsets of nodes, all adjacent to each other, also called complete subgraphs. In a SSN, a clique corresponds to sequences which can all aligned to each other (and thus potentially to a gene family).

Closeness : measure of the centrality/peripherality of a node in a network or in other words measure of the mean distance from a node to other nodes

Connected component (CC): connected component of an undirected graph is a maximal set of nodes such that each pair of nodes is connected by a path.

Contig(s) : a set of reads that are related to one another by overlap of their sequences

Core genome: set of genes that are present in all compared genomes (e.g. at the species level, the core genome represents the genes present in all strains of this species).

Culturing bias: cultured microbial strains do not necessarily represent, and usually are not, the dominant members of the environment from which they were isolated. This bias affects bacteria, viruses, and protists. The culturing bias can be the result of a lack of continuous culturing efforts, or inadequate isolation and/or culturing strategies – or because, for whatever reason, some species in the environment may be refractory to isolation and culturing.

***dark matter*:** microbial *dark matter* refers to the 99% of microorganisms that have never been cultivated in the laboratory. Advances in high throughput sequencing allow now their exhaustive exploration in the environment. However as current inference tools mostly rely on taxonomically or functionally identified sequences, a significant part of the currently available meta-omic data is ignored and our current inferences might be biased. When studying omics, one can considered taxonomical *dark matter*, functional *dark matter* or taxonomical and functional *dark matter* (cf. Figure 21).

Degree (node) : centrality measure that counts how many neighbors a node has.

Ecosystem : biological community of interacting entities and their physical environment.

Environmental genomic (*sensu largo*) or environmental omic: cf. definition of meta-omics.

Eukaryotes: are defined as organisms whose cells have a nucleus enclosed within membranes.

Excavata: the group Excavata was proposed based of shared morphological characters [32], and was later confirmed through phylogenomic analyses [33]. Most members of this group are heterotrophic organisms, among them some well-known human parasites such as Trichomonas vaginalis (the agent of trichomoniasis) and Giardia lamblia (the agent of giardiasis), as well as animal parasites such as Leishmania sp. (the agent of leishmaniasis) as well as Trypanosoma brucei, and Trypanosoma cruzi (the agents of sleeping sickness and Chagas disease respectively).

Genomes OnLine Database (GOLD): an online resource for comprehensive access to information regarding genome and metagenome sequencing projects, and their associated metadata (http://www.genomesonline.org/) (Pagani et al., 2012).

Graph : a set of nodes and edges.

Heterotrophy: nutritional mode that involves the use of preformed organic matter for the acquisition of carbon and energy.

Hub nodes : scale-free networks have many nodes with few links and a few highly connected nodes that are termed hubs. They are therefore supposed to be robust towards random node removal but sensitive to the removal of hub nodes. The hub nodes can been linked to the ecological concept of the keystone species.

Illumina: company producing the Hi-Seq and MiSeq platforms, which uses bridge amplification. This calssical Illumina sequencing technology belong to the has a

Incertae sedis: Latin for 'of uncertain placement', a term used to indicate those organisms or lineages with unclear taxonomical position.

k-mer : refers to all the possible substrings of length k that are contained in a string

Large-subunit (LSU) ribosomal DNA (rDNA) : cf. 28S rDNA definition

Meta-omics: (meta refers to the environement) generic term to designate meta-genomics, meta-transcriptomics, meta-proteomics, meta-bolomics, etc.

Metabarcoding or Marker gene surveys: high-throughput environmental sequencing utilizing homologous genetic loci (e.g. 16S rDNA for prokaryotes, 18S rDNA for eukaryotes) amplified via theoretically universal and conserved primer sets.

Metagenomes: collections of all the DNA present in communities of microorganisms, representing all the genetic potential of the communities. The metagenome of a community can be used to reconstruct the genomes of the individual species comprising that community, thus assigning specific metabolic roles to those taxa.

Metagenomics: high-throughput, random sequencing of genomic DNA from environmental isolates.

Metatranscriptomes: Collections of all the transcriptomes (all RNA transcripts) present in communities of microorganisms; a metatranscriptome of a community is derived from RNA extraction and purification, reverse transcription of RNA to cDNA and sequencing of the resulting cDNA.

Metatranscriptomics: high-throughput sequencing of expressed gene transcripts (mRNA) from environmental isolates.

Metazoa: multicellular, eukaryotic organisms (animals) that have differentiated cells and tissues.

Microbe(s): microscopic organism or or micro-organisms is a living thing that is too small to be seen with the naked eye (involving Prokaryotes, Viruses/Giruses, Protists).

Microbial Dark matter: cf. dark matter definition and (Rinke et al., 2013)

Microbiome: *sensu stricto* refers to microbial functions from an environment, *sensu largo* refers to microbial species and functions from an environment.

Microbiote (plural microbiota or microbiotes) : refers to microbial species from an environment.

Mitags : SSU rDNA fragments derived from Illumina-sequenced environmental metagenomes (Logares et al., 2014).

Mixotrophy : the physiological feature of an organism whose cells (at least some cells, for multicellular organisms) use both photosynthesis and external organic matter as a source of carbon and/or non-carbon elements. It is now considered that the majority of protists are mixotrophs (more details in annexe 4).

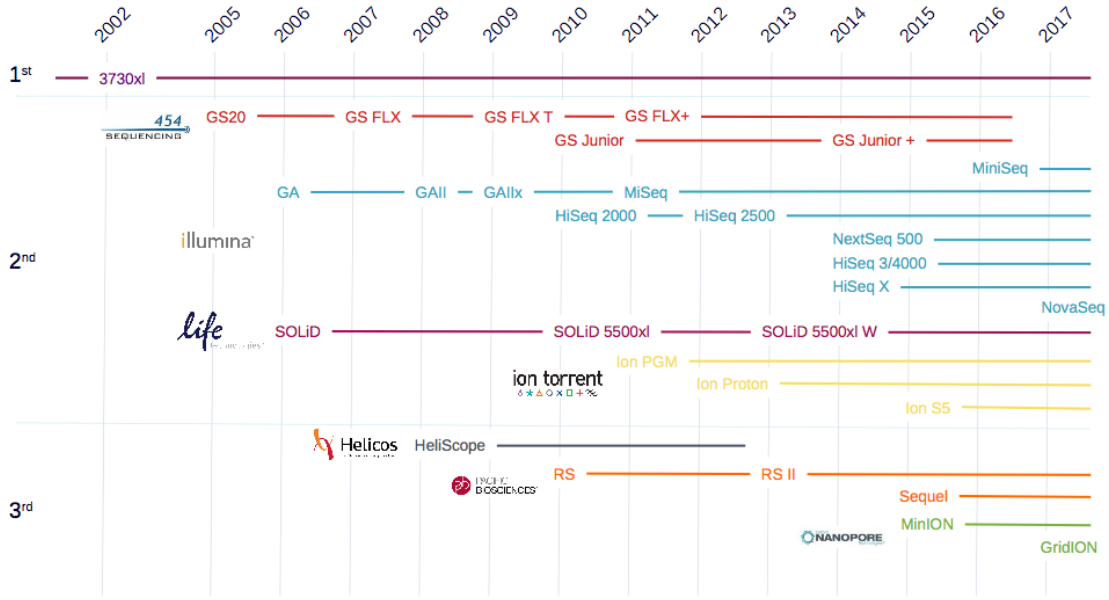
Modularity: networks can be divided into clusters either manually or by using a network cluster algorithm. More modular networks have a higher number of within-cluster edges than between-cluster edges compared to random expectation.

Module (in the framework association / co-occurrence network analyses) : clusters of highly interconnected genes. In an unsigned co-expression network, modules correspond to clusters of genes with high absolute correlations. In a signed network, modules correspond to positively correlated genes.

Nestedness: nested neighborhood structure of the nodes in a network. A network exhibits nestedness if the neighborhood of a node is contained in the neighborhoods of the nodeswith higher degrees.

Next Generation Sequencing : sequencing technologies following the first generation (i.e. first generation correspond for instance to ABI 3730xl DNA Analyzer sequencer). Cf. below chronology of

sequencing technologies (figure shared by S. Le Crom).



Node degree distribution: the distribution of the number of neighbor each node has in a network. In random interaction networks, the node degree distribution follows a Poisson distribution. However, for most biological networks, the degree distribution is better described by a power law distribution ('as seen in scale-free' networks). Although the node degree distributions of microbial assocatioan networks are not always fit by a power law, they are clearly far from being random.

Omics: refers to genomics, transcriptomics, proteomics, metabolomics etc.

Operational taxonomic unit(s) (OTU(s)): an operational definition which is supposed to be a proxy for a species or group of species. In microbial ecology, and in particular protist ecology, this operational definition is generally based in a percentage similarity threshold of the 18S rDNA (e.g., OTU_{97} refers to a cluster of sequences with >97% similarity that are inferred to represent a single taxonomic unit).

Opisthokonta: the opisthokonts include two of the best-studied kingdoms of life: the Metazoa (animals) and the Fungi. Recent phylogenetic and phylogenomic analyses have shown that the Opisthokonta also include several unicellular lineages. These include the Choanoflagellata (the closest unicellular relatives of the animals) and the Ichthyospora (that include several fish parasites that impact negatively on aquaculture).

Pangenome: is the entire gene set of compared datasets (e.g. the entire gene set of all strains of a species). It includes genes present in the core genome and genes present only in some datasets (variable or accessory genome).

Phototrophy: A nutritional mode that involves the use of light for the production of organic carbon and the acquisition of energy.

Phytoplankton: Planktonic protists that use phototrophy as their nutritional mode. The term has ecological importance but no phylogenetic correspondance because the behaviour occurs across many lineages of protists.

Plankton: aggregate of passively floating or drifting organisms occurring in a body of water (e.g. ocean, lake). By contrast to the planktonic organisms, the benthic organisms are the organisms living on the bottom of the seafloor (i.e. not in the water column). Plankton corresponds to a wide range of organisms (from viruses to fish larvae), is usely classified accoding to their size (e.g. pico-, nano-, micro-, meso-) and / or to their trophic mode (e.g. phytoplankton, zooplankton).

Group	*Approximative size range*
Megaplankton	> 20 cm
Macroplankton	2→20 cm
Mesoplankton	0.2→20 mm
Microplankton	20→200 µm
Nanoplankton	2→20 µm
Picoplankton	0.2→2 µm
Femtoplankton	< 0.2 µm

Plants: or land Plants or Embryophyta (cf. (Adl et al., 2012))

Primary production: the photosynthetic production of organic carbon, carried out by a wide variety of protists, macroalgae and plants.

Prokaryote(s): microscopic single-celled organism that lack a membrane-bound nucleus, mitochondria or any other membrane-bound organelles. Bacteria and Archaea are prokaryotes.

Protist(s): term currently commonly used to refer to all eukaryotic lineages that are neither plants, nor animals, nor fungi (Pawlowski et al., 2012)

Protozoa: Protists that are not photosynthetic, but are instead dependent on the ingestion of preformed organic matter (usually prey) for their nutrition. This older term is still in use; 'heterotrophic protists' is used synonymously.

Pyrosequencing: general term referring to light-based high-throughput sequencing techniques (e.g. 454).

rDNA: is the DNA-part that encodes for rRNA

Read(s): sequence(s) obtained by HTS / NGS

rRNA: is the RNA component of the ribosome

Remineralization: component of the BCP, remineralization corresponds to the transformation of organic carbon into CO_2 by the process of respiration.

Rhizaria: diverse group of protists, mostly heterotrophic unicellular eukaryotes including both amoeboid and flagellate forms. Two iconic protist groups, Haeckel's Radiolaria and the Foraminifera, are members of the Rhizaria. Foraminifera have been very useful in paleoclimatology and paleoceanography due to their external shell that can be detected in the fossil record.

SAR (Stramenopila - Alveolata, and Rhizaria): three protistan groups that have been historically studied separately. Phylogenetic analyses, however, have shown that those three groups share a common ancestor, forming a supergroup known as SAR (e.g. Burki et al., 2016).

Scale-free network: is a network whose degree distribution follows a power law, at least asymptotically.

Shortest path : shortest distance between a pair of nodes (in a graph)

Single amplified genomes (SAGs): the products of single cell whole-genome amplification that can be further analyzed in similar ways to DNA extracts from pure cultures.

Single cell genomics (SCG): a method to amplify and sequence the genome of a single cell. The method consists of an integrated pipeline that starts with the collection and preservation of environmental samples, followed by physical separation, lysis, and whole-genome amplification from individual cells. This is followed by sequencing of the resulting material.

Small-subunit (SSU) ribosomal DNA (rDNA): cf. 18S rDNA or 16S rDNA definition

Stramenopila: also known as heterokonts, the stramenopiles include a wide range of ubiquitous phototrophic and heterotrophic organisms. Most are unicellular flagellates but there are also some multicellular organisms, such as the giant kelps. Other relevant members of the Stramenopila are the diatoms (algae contained within a silica cell wall), the chrysophytes (abundant in freshwater environments), the MAST (marine stramenopile) groups (the most abundant microbial predators of the ocean), and plant parasites such as the Peronosporomycetes.

Taxon (plural taxa): group of one or more populations of an organism or organisms

Trait: phenotypic characteristic of a (micro)organism. Functional traits are morphological, biochemical, physiological, structural, phenological or behavioural characteristics of organisms that influence performance or fitness of the organisms.

Abbreviations :

BCP : Biological Carbon Pump
bp : base pair
CC(s): connected component(s)
DNA : Deoxyribonucleic acid
Gb : gigabase or 10^9 base
HTS : high-throughput sequencing
LSU : large ribosomal subunit (cf. 28S
Mb : megabase or 10^6 base
NGS : next generation sequencing
NPP : net primary production
nm: nanometer or 10^{-9} meter
ORF(s) : Open Reading Frame(s)
OTU(s) : Operational taxonomic unit
PCR : Polymerase chain reaction (PCR) amplification
RNA : Ribonucleic acid
SSN : Sequence Similarity Network (cf. Atkinson et al., 2009)
SSU : small ribosomal subunit (cf. definition 18S rDNA and 16S rDNA)
Tb : terabase or 10^{12} base
µm : micrometer
VIP : variable importance in projection (cf. Guidi et al., 2016)

n.b. Metric prefix

10^{12}	tera	T
10^{9}	giga	G
10^{6}	mega	M
10^{-3}	milli	m
10^{-6}	micro	µ
10^{-9}	nano	n
10^{-12}	pico	p

Bibliographical references

Adl, S.M., Bass, D., Lane, C.E., Lukeš, J., Schoch, C.L., Smirnov, A., Agatha, S., Berney, C., Brown, M.W., Burki, F., Cárdenas, P., Čepička, I., Chistyakova, L., Campo, J. del, Dunthorn, M., Edvardsen, B., Eglit, Y., Guillou, L., Hampl, V., Heiss, A.A., Hoppenrath, M., James, T.Y., Karpov, S., Kim, E., Kolisko, M., Kudryavtsev, A., Lahr, D.J.G., Lara, E., Gall, L.L., Lynn, D.H., Mann, D.G., Molera, R.M. i, Mitchell, E.A.D., Morrow, C., Park, J.S., Pawlowski, J.W., Powell, M.J., Richter, D.J., Rueckert, S., Shadwick, L., Shimano, S., Spiegel, F.W., Cortes, G.T. i, Youssef, N., Zlatogursky, V., Zhang, Q., 2018. Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. J. Eukaryot. Microbiol. 0. https://doi.org/10.1111/jeu.12691

Adl, S.M., Simpson, A.G., Lane, C.E., Lukeš, J., Bass, D., Bowser, S.S., Brown, M., Burki, F., Dunthorn, M., Hampl, V., Heiss, A., Hoppenrath, M., Lara, E., leGall, L., Lynn, D.H., McManus, H., Mitchell, E.A.D., Mozley-Stanridge, S.E., Parfrey, L.W., Pawlowski, J., Rueckert, S., Shadwick, L., Schoch, C., Smirnov, A., Spiegel, F.W., 2012. The revised classification of eukaryotes. J. Eukaryot. Microbiol. 59, 429–493. https://doi.org/10.1111/j.1550-7408.2012.00644.x

Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I., Albini, G., Aury, J.-M., Belser, C., Bertrand, A., Cruaud, C., Da Silva, C., Dossat, C., Gavory, F., Gas, S., Guy, J., Haquelle, M., Jacoby, E., Jaillon, O., Lemainque, A., Pelletier, E., Samson, G., Wessner, M., Genoscope Technical Team, Bazire, P., Beluche, O., Bertrand, L., Besnard-Gonnet, M., Bordelais, I., Boutard, M., Dubois, M., Dumont, C., Ettedgui, E., Fernandez, P., Garcia, E., Aiach, N.G., Guerin, T., Hamon, C., Brun, E., Lebled, S., Lenoble, P., Louesse, C., Mahieu, E., Mairey, B., Martins, N., Megret, C., Milani, C., Muanga, J., Orvain, C., Payen, E., Perroud, P., Petit, E., Robert, D., Ronsin, M., Vacherie, B., Acinas, S.G., Royo-Llonch, M., Cornejo-Castillo, F.M., Logares, R., Fernández-Gómez, B., Bowler, C., Cochrane, G., Amid, C., Hoopen, P.T., Vargas, C.D., Grimsley, N., Desgranges, E., Kandels-Lewis, S., Ogata, H., Poulton, N., Sieracki, M.E., Stepanauskas, R., Sullivan, M.B., Brum, J.R., Duhaime, M.B., Poulos, B.T., Hurwitz, B.L., Coordinators, T.O.C., Acinas, S.G., Bork, P., Boss, E., Bowler, C., Vargas, C.D., Follows, M., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Karp-Boss, L., Karsenti, E., Not, F., Ogata, H., Pesant, S., Raes, J., Sardet, C., Sieracki, M.E., Speich, S., Stemmann, L., Sullivan, M.B., Sunagawa, S., Wincker, P., Pesant, S., Karsenti, E., Wincker, P., 2017. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. Sci. Data 4, 170093. https://doi.org/10.1038/sdata.2017.93

Alvarez-Ponce, D., Lopez, P., Bapteste, E., McInerney, J.O., 2013. Gene similarity networks provide tools for understanding eukaryote origins and evolution. Proc. Natl. Acad. Sci. 110, E1594–E1603. https://doi.org/10.1073/pnas.1211371110

Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W., Huse, S.M., 2009. A Method for Studying Protistan Diversity Using Massively Parallel Sequencing of V9 Hypervariable Regions of Small-Subunit Ribosomal RNA Genes. PLOS ONE 4, e6372. https://doi.org/10.1371/journal.pone.0006372

Antoine-Lorquin, A., Mahé, F., Dunthorn, M., Belleannée, C., 2016. Detection of mutated primers and impact on targeted metagenomics results. RCAM16 Recent Comput. Adv. Metagenomics 21.

Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815. https://doi.org/10.1038/35048692

Armbrust, E.V., 2014. Taking the pulse of ocean microbes. Science 345, 134–135. https://doi.org/10.1126/science.1256578

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., Brzezinski, M.A., Chaal, B.K., Chiovitti, A., Davis, A.K., Demarest, M.S., Detter, J.C., Glavina, T., Goodstein, D., Hadi, M.Z., Hellsten, U., Hildebrand, M., Jenkins, B.D., Jurka, J., Kapitonov, V.V., Kröger, N., Lau, W.W.Y., Lane, T.W., Larimer, F.W., Lippmeier, J.C., Lucas, S., Medina, M., Montsant, A., Obornik, M., Parker, M.S., Palenik, B., Pazour, G.J., Richardson, P.M., Rynearson, T.A., Saito, M.A., Schwartz, D.C., Thamatrakoln, K., Valentin, K., Vardi, A., Wilkerson, F.P., Rokhsar, D.S., 2004. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science 306, 79–86. https://doi.org/10.1126/science.1101156

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H.B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E.G., Wang, J., Guarner, F., Pedersen, O., de Vos, W.M., Brunak, S., Doré, J., MetaHIT Consortium (additional Members), Antolín, M., Artiguenave, F., Blottiere, H.M., Almeida, M., Brechot, C., Cara, C., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Foerstner, K.U., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Huber, W., van Hylckama-Vlieg, J., Jamet, A., Juste, C., Kaci, G., Knol, J., Kristiansen, K., Lakhdari, O., Layec, S., Le Roux, K., Maguin, E., Mérieux, A., Melo Minardi, R., M'rini, C., Muller, J., Oozeer, R., Parkhill, J., Renault, P., Rescigno, M., Sanchez, N., Sunagawa, S., Torrejon, A., Turner, K., Vandemeulebrouck, G., Varela, E., Winogradsky, Y., Zeller, G., Weissenbach, J., Ehrlich, S.D., Bork, P., 2011. Enterotypes of the human gut microbiome. Nature 473, 174–180. https://doi.org/10.1038/nature09944

Atkinson, H.J., Morris, J.H., Ferrin, T.E., Babbitt, P.C., 2009. Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. PLoS ONE 4. https://doi.org/10.1371/journal.pone.0004345

Aylward, F.O., Eppley, J.M., Smith, J.M., Chavez, F.P., Scholin, C.A., DeLong, E.F., 2015. Microbial community transcriptional networks are conserved in three domains at ocean basin scales. Proc. Natl. Acad. Sci. U. S. A. 112, 5443–5448. https://doi.org/10.1073/pnas.1502883112

Baldauf, S.L., 2003. The Deep Roots of Eukaryotes. Science 300, 1703–1706. https://doi.org/10.1126/science.1085544

Barberán, A., Bates, S.T., Casamayor, E.O., Fierer, N., 2012. Using network analysis to explore co-occurrence patterns in soil microbial communities. ISME J. 6, 343–351. https://doi.org/10.1038/ismej.2011.119

Bates, S.T., Clemente, J.C., Flores, G.E., Walters, W.A., Parfrey, L.W., Knight, R., Fierer, N., 2013. Global biogeography of highly diverse protistan communities in soil. ISME J. 7, 652–659. https://doi.org/10.1038/ismej.2012.147

Benoiston, A.-S., Ibarbalz, F.M., Bittner, L., Guidi, L., Jahn, O., Dutkiewicz, S., Bowler, C., 2017. The evolution of diatoms and their biogeochemical functions. Phil Trans R Soc B 372, 20160397. https://doi.org/10.1098/rstb.2016.0397

Bernard, G., Pathmanathan, J.S., Lannes, R., Lopez, P., Bapteste, E., 2018. Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery. Genome Biol. Evol. 10, 707–715. https://doi.org/10.1093/gbe/evy031

Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., Lennard, N.J., Caler, E., Hamlin, N.E., Haas, B., Böhme, U., Hannick, L., Aslett, M.A., Shallom, J., Marcello, L., Hou, L., Wickstead, B., Alsmark, U.C.M., Arrowsmith, C., Atkin, R.J., Barron, A.J., Bringaud, F., Brooks, K., Carrington, M., Cherevach, I., Chillingworth, T.-J., Churcher, C., Clark, L.N., Corton, C.H., Cronin, A., Davies, R.M., Doggett, J., Djikeng, A., Feldblyum, T., Field, M.C., Fraser, A., Goodhead, I., Hance, Z., Harper, D., Harris, B.R., Hauser, H., Hostetler, J.,

Ivens, A., Jagels, K., Johnson, D., Johnson, J., Jones, K., Kerhornou, A.X., Koo, H., Larke, N., Landfear, S., Larkin, C., Leech, V., Line, A., Lord, A., MacLeod, A., Mooney, P.J., Moule, S., Martin, D.M.A., Morgan, G.W., Mungall, K., Norbertczak, H., Ormond, D., Pai, G., Peacock, C.S., Peterson, J., Quail, M.A., Rabbinowitsch, E., Rajandream, M.-A., Reitter, C., Salzberg, S.L., Sanders, M., Schobel, S., Sharp, S., Simmonds, M., Simpson, A.J., Tallon, L., Turner, C.M.R., Tait, A., Tivey, A.R., Aken, S.V., Walker, D., Wanless, D., Wang, S., White, B., White, O., Whitehead, S., Woodward, J., Wortman, J., Adams, M.D., Embley, T.M., Gull, K., Ullu, E., Barry, J.D., Fairlamb, A.H., Opperdoes, F., Barrell, B.G., Donelson, J.E., Hall, N., Fraser, C.M., Melville, S.E., El-Sayed, N.M., 2005. The Genome of the African Trypanosome Trypanosoma brucei. Science 309, 416–422. https://doi.org/10.1126/science.1112642

Bik, H.M., Sung, W., Ley, P.D., Baldwin, J.G., Sharma, J., Rocha-Olivares, A., Thomas, W.K., 2012. Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. Mol. Ecol. 21, 1048–1059. https://doi.org/10.1111/j.1365-294X.2011.05297.x

Bittner, L., Gobet, A., Audic, S., Romac, S., Egge, E.S., Santini, S., Ogata, H., Probert, I., Edvardsen, B., de Vargas, C., 2013. Diversity patterns of uncultured Haptophytes unravelled by pyrosequencing in Naples Bay. Mol. Ecol. 22, 87–101. https://doi.org/10.1111/mec.12108

Bittner, L., Halary, S., Payri, C., Cruaud, C., de Reviers, B., Lopez, P., Bapteste, E., 2010. Some considerations for analyzing biodiversity using integrative metagenomics and gene networks. Biol. Direct 5, 47. https://doi.org/10.1186/1745-6150-5-47

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., Shao, Y., 1997. The complete genome sequence of Escherichia coli K-12. Science 277, 1453–1462.

Blazewicz, S.J., Barnard, R.L., Daly, R.A., Firestone, M.K., 2013. Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. ISME J. 7, 2061–2068. https://doi.org/10.1038/ismej.2013.102

Bohan, D.A., Vacher, C., Tamaddoni-Nezhad, A., Raybould, A., Dumbrell, A.J., Woodward, G., 2017. Next-Generation Global Biomonitoring: Large-scale, Automated Reconstruction of Ecological Networks. Trends Ecol. Evol. 32, 477–487. https://doi.org/10.1016/j.tree.2017.03.001

Bontell, I.L., Hall, N., Ashelford, K.E., Dubey, J., Boyle, J.P., Lindh, J., Smith, J.E., 2009. Whole genome sequencing of a natural recombinant Toxoplasma gondiistrain reveals chromosome sorting and local allelic variants. Genome Biol. 10, R53. https://doi.org/10.1186/gb-2009-10-5-r53

Bork, P., Bowler, C., Vargas, C. de, Gorsky, G., Karsenti, E., Wincker, P., 2015. Tara Oceans studies plankton at planetary scale. Science 348, 873–873. https://doi.org/10.1126/science.aac5605

Botebol, H., Lelandais, G., Six, C., Lesuisse, E., Meng, A., Bittner, L., Lecrom, S., Sutak, R., Lozano, J.-C., Schatt, P., Vergé, V., Blain, S., Bouget, F.-Y., 2017. Acclimation of a low iron adapted Ostreococcus strain to iron limitation through cell biomass lowering. Sci. Rep. 7, 327. https://doi.org/10.1038/s41598-017-00216-6

Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R.P., Rayko, E., Salamov, A., Vandepoele, K., Beszteri, B., Gruber, A., Heijde, M., Katinka, M., Mock, T., Valentin, K., Verret, F., Berges, J.A., Brownlee, C., Cadoret, J.-P., Chiovitti, A., Choi, C.J., Coesel, S., De Martino, A., Detter, J.C., Durkin, C., Falciatore, A., Fournet, J., Haruta, M., Huysman, M.J.J., Jenkins, B.D., Jiroutova, K., Jorgensen, R.E., Joubert, Y., Kaplan, A., Kröger, N., Kroth, P.G., La Roche, J., Lindquist, E., Lommer, M., Martin–Jézéquel, V., Lopez, P.J., Lucas, S., Mangogna, M., McGinnis, K., Medlin, L.K., Montsant, A., Secq, M.-P.O., Napoli, C., Obornik, M., Parker, M.S., Petit, J.-L., Porcel, B.M., Poulsen, N., Robison, M., Rychlewski, L., Rynearson, T.A., Schmutz, J., Shapiro, H., Siaut, M., Stanley, M., Sussman, M.R., Taylor, A.R., Vardi, A., von Dassow, P., Vyverman, W., Willis, A., Wyrwicz, L.S., Rokhsar, D.S., Weissenbach, J., Armbrust, E.V., Green, B.R., Van de Peer, Y., Grigoriev, I.V., 2008. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature 456, 239–244. https://doi.org/10.1038/nature07410

Brown, M.W., Heiss, A.A., Kamikawa, R., Inagaki, Y., Yabuki, A., Tice, A.K., Shiratori, T., Ishida, K.-I., Hashimoto, T., Simpson, A.G.B., Roger, A.J., 2018. Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group. Genome Biol. Evol. 10, 427–433. https://doi.org/10.1093/gbe/evy014

Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., Chaffron, S., Cruaud, C., Vargas, C. de, Gasol, J.M., Gorsky, G., Gregory, A.C., Guidi, L., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Poulos, B.T., Schwenck, S.M., Speich, S., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Coordinators, T.O., Bork, P., Bowler, C., Sunagawa, S., Wincker, P., Karsenti, E., Sullivan, M.B., 2015. Patterns and ecological drivers of ocean viral communities. Science 348, 1261498. https://doi.org/10.1126/science.1261498

Burki, F., 2014. The Eukaryotic Tree of Life from a Global Phylogenomic Perspective. Cold Spring Harb. Perspect. Biol. 6, a016147. https://doi.org/10.1101/cshperspect.a016147

Burki, F., Kaplan, M., Tikhonenkov, D.V., Zlatogursky, V., Minh, B.Q., Radaykina, L.V., Smirnov, A., Mylnikov, A.P., Keeling, P.J., 2016. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. Proc R Soc B 283, 20152802. https://doi.org/10.1098/rspb.2015.2802

Burki, F., Keeling, P.J., 2014. Rhizaria. Curr. Biol. 24, R103–R107. https://doi.org/10.1016/j.cub.2013.12.025

Callahan, B.J., McMurdie, P.J., Holmes, S.P., 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 11, 2639–2643. https://doi.org/10.1038/ismej.2017.119

Caron, D.A., Alexander, H., Allen, A.E., Archibald, J.M., Armbrust, E.V., Bachy, C., Bell, C.J., Bharti, A., Dyhrman, S.T., Guida, S.M., Heidelberg, K.B., Kaye, J.Z., Metzner, J., Smith, S.R., Worden, A.Z., 2017. Probing the evolution, ecology and physiology of marine protists using transcriptomics. Nat. Rev. Microbiol. 15, 6–20. https://doi.org/10.1038/nrmicro.2016.160

Caron, D.A., Worden, A.Z., Countway, P.D., Demir, E., Heidelberg, K.B., 2009. Protists are microbes too: a perspective. ISME J. 3, 4–12. https://doi.org/10.1038/ismej.2008.101

Carradec, Q., Pelletier, E., Silva, C.D., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M.-A., Méheust, R., Poulain, J., Romac, S., Richter, D.J., Yoshikawa, G., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Jaillon, O., Aury, J.-M., Karsenti, E., Sullivan, M.B., Sunagawa, S., Bork, P., Not, F., Hingamp, P., Raes, J., Guidi, L., Ogata, H., Vargas, C., Iudicone, D., Bowler, C., Wincker, P., 2018. A global ocean atlas of eukaryotic genes. Nat. Commun. 9, 373. https://doi.org/10.1038/s41467-017-02342-1

Chaffron, S., Rehrauer, H., Pernthaler, J., Mering, C. von, 2010. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res. 20, 947–959. https://doi.org/10.1101/gr.104521.109

Cheng, S., Karkar, S., Bapteste, E., Yee, N., Falkowski, P., Bhattacharya, D., 2014. Sequence similarity network reveals the imprints of major diversification events in the evolution of microbial life. Front. Ecol. Evol. 2. https://doi.org/10.3389/fevo.2014.00072

Cock, J.M., Sterck, L., Rouzé, P., Scornet, D., Allen, A.E., Amoutzias, G., Anthouard, V., Artiguenave, F., Aury, J.-M., Badger, J.H., Beszteri, B., Billiau, K., Bonnet, E., Bothwell, J.H., Bowler, C., Boyen, C., Brownlee, C., Carrano, C.J., Charrier, B., Cho, G.Y., Coelho, S.M., Collén, J., Corre, E., Da Silva, C., Delage, L., Delaroque, N., Dittami, S.M., Doulbeau, S., Elias, M., Farnham, G., Gachon, C.M.M., Gschloessl, B., Heesch, S., Jabbari, K., Jubin, C., Kawai, H., Kimura, K., Kloareg, B., Küpper, F.C., Lang, D., Le Bail, A., Leblanc, C., Lerouge, P., Lohr, M.,

Lopez, P.J., Martens, C., Maumus, F., Michel, G., Miranda-Saavedra, D., Morales, J., Moreau, H., Motomura, T., Nagasato, C., Napoli, C.A., Nelson, D.R., Nyvall-Collén, P., Peters, A.F., Pommier, C., Potin, P., Poulain, J., Quesneville, H., Read, B., Rensing, S.A., Ritter, A., Rousvoal, S., Samanta, M., Samson, G., Schroeder, D.C., Ségurens, B., Strittmatter, M., Tonon, T., Tregear, J.W., Valentin, K., von Dassow, P., Yamagishi, T., Van de Peer, Y., Wincker, P., 2010. The Ectocarpus genome and the independent evolution of multicellularity in brown algae. Nature 465, 617–621. https://doi.org/10.1038/nature09016

Cordier, T., Forster, D., Dufresne, Y., Martins, C.I., Stoeck, T., Pawlowski, J., 2018. Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. Mol. Ecol. Resour. https://doi.org/10.1111/1755-0998.12926

Corel, E., Lopez, P., Méheust, R., Bapteste, E., 2016. Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. Trends Microbiol. 24, 224–237. https://doi.org/10.1016/j.tim.2015.12.003

Corradi, N., Pombert, J.-F., Farinelli, L., Didier, E.S., Keeling, P.J., 2010. The complete sequence of the smallest known nuclear genome from the microsporidian Encephalitozoon intestinalis. Nat. Commun. 1, 77. https://doi.org/10.1038/ncomms1082

Countway, P.D., Gast, R.J., Dennett, M.R., Savai, P., Rose, J.M., Caron, D.A., 2007. Distinct protistan assemblages characterize the euphotic zone and deep sea (2500 m) of the western North Atlantic (Sargasso Sea and Gulf Stream). Environ. Microbiol. 9, 1219–1232. https://doi.org/10.1111/j.1462-2920.2007.01243.x

Csárdi G., Nepusz T., 2006. The igraph software package for complex network research. InterJournal Complex Systems;1695.

Cuvelier, M.L., Allen, A.E., Monier, A., McCrow, J.P., Messie, M., Tringe, S.G., Woyke, T., Welsh, R.M., Ishoey, T., Lee, J.-H., Binder, B.J., DuPont, C.L., Latasa, M., Guigand, C., Buck, K.R., Hilton, J., Thiagarajan, M., Caler, E., Read, B., Lasken, R.S., Chavez, F.P., Worden, A.Z., 2010. Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. Proc. Natl. Acad. Sci. 107, 14679–14684. https://doi.org/10.1073/pnas.1001665107

Cuvelier, M.L., Ortiz, A., Kim, E., Moehlig, H., Richardson, D.E., Heidelberg, J.F., Archibald, J.M., Worden, A.Z., 2008. Widespread distribution of a unique marine protistan lineage. Environ. Microbiol. 10, 1621–1634. https://doi.org/10.1111/j.1462-2920.2008.01580.x

De Queiroz, K., 2007. Species Concepts and Species Delimitation. Syst. Biol. 56, 879–886. https://doi.org/10.1080/10635150701701083

Decelle, J., Romac, S., Stern, R.F., Bendif, E.M., Zingone, A., Audic, S., Guiry, M.D., Guillou, L., Tessier, D., Gall, F.L., Gourvil, P., Santos, A.L.D., Probert, I., Vaulot, D., Vargas, C. de, Christen, R., 2015. PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. Mol. Ecol. Resour. 15, 1435–1445. https://doi.org/10.1111/1755-0998.12401

del Campo, J., Sieracki, M.E., Molestina, R., Keeling, P., Massana, R., Ruiz-Trillo, I., 2014. The others: our biased perspective of eukaryotic genomes. Trends Ecol. Evol. 29, 252–259. https://doi.org/10.1016/j.tree.2014.03.006

Delmont, T.O., Eren, A.M., Vineis, J.H., Post, A.F., 2015. Genome reconstructions indicate the partitioning of ecological functions inside a phytoplankton bloom in the Amundsen Sea, Antarctica. Front. Microbiol. 6. https://doi.org/10.3389/fmicb.2015.01090

Delmont, T.O., Quince, C., Shaiber, A., Esen, Ö.C., Lee, S.T., Rappé, M.S., MacLellan, S.L., Lücker, S., Eren, A.M., 2018. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. Nat. Microbiol. 1. https://doi.org/10.1038/s41564-018-0176-9

Derelle, E., Ferraz, C., Rombauts, S., Rouze, P., Worden, A.Z., Robbens, S., Partensky, F., Degroeve, S., Echeynie, S., Cooke, R., Saeys, Y., Wuyts, J., Jabbari, K., Bowler, C., Panaud, O., Piegu, B., Ball, S.G., Ral, J.-P., Bouget, F.-Y., Piganeau, G., De Baets, B., Picard, A., Delseny, M., Demaille, J., Van de Peer, Y., Moreau, H., 2006. Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. Proc. Natl. Acad. Sci. 103, 11647–11652. https://doi.org/10.1073/pnas.0604795103

Díez, B., Pedrós-Alió, C., Massana, R., 2001. Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. Appl. Environ. Microbiol. 67, 2932–2941. https://doi.org/10.1128/AEM.67.7.2932-2941.2001

Dobell C., 1911. The principles of protistology. Arch. Protistenkd. 23, 269–310.

Dunthorn, M., Klier, J., Bunge, J., Stoeck, T., 2012. Comparing the Hyper-Variable V4 and V9 Regions of the Small Subunit rDNA for Assessment of Ciliate Environmental Diversity. J. Eukaryot. Microbiol. 59, 185–187. https://doi.org/10.1111/j.1550-7408.2011.00602.x

Durno, W.E., Hanson, N.W., Konwar, K.M., Hallam, S.J., 2013. Expanding the boundaries of local similarity analysis. BMC Genomics 14, S3. https://doi.org/10.1186/1471-2164-14-S1-S3

Edgcomb, V., Orsi, W., Bunge, J., Jeon, S., Christen, R., Leslin, C., Holder, M., Taylor, G.T., Suarez, P., Varela, R., Epstein, S., 2011. Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. ISME J. 5, 1344–1356. https://doi.org/10.1038/ismej.2011.6

Egge, E., Bittner, L., Andersen, T., Audic, S., Vargas, C. de, Edvardsen, B., 2013. 454 Pyrosequencing to Describe Microbial Eukaryotic Community Composition, Diversity and Relative Abundance: A Test for Marine Haptophytes. PLOS ONE 8, e74371. https://doi.org/10.1371/journal.pone.0074371

Egge, E.S., Johannessen, T.V., Andersen, T., Eikrem, W., Bittner, L., Larsen, A., Sandaa, R.-A., Edvardsen, B., 2015. Seasonal diversity and dynamics of haptophytes in the Skagerrak, Norway, explored by high-throughput sequencing. Mol. Ecol. 24, 3026–3042. https://doi.org/10.1111/mec.13160

Eichinger, L., Pachebat, J.A., Glöckner, G., Rajandream, M.-A., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., Tunggal, B., Kummerfeld, S., Madera, M., Konfortov, B.A., Rivero, F., Bankier, A.T., Lehmann, R., Hamlin, N., Davies, R., Gaudet, P., Fey, P., Pilcher, K., Chen, G., Saunders, D., Sodergren, E., Davis, P., Kerhornou, A., Nie, X., Hall, N., Anjard, C., Hemphill, L., Bason, N., Farbrother, P., Desany, B., Just, E., Morio, T., Rost, R., Churcher, C., Cooper, J., Haydock, S., van Driessche, N., Cronin, A., Goodhead, I., Muzny, D., Mourier, T., Pain, A., Lu, M., Harper, D., Lindsay, R., Hauser, H., James, K., Quiles, M., Babu, M.M., Saito, T., Buchrieser, C., Wardroper, A., Felder, M., Thangavelu, M., Johnson, D., Knights, A., Loulseged, H., Mungall, K., Oliver, K., Price, C., Quail, M.A., Urushihara, H., Hernandez, J., Rabbinowitsch, E., Steffen, D., Sanders, M., Ma, J., Kohara, Y., Sharp, S., Simmonds, M., Spiegler, S., Tivey, A., Sugano, S., White, B., Walker, D., Woodward, J., Winckler, T., Tanaka, Y., Shaulsky, G., Schleicher, M., Weinstock, G., Rosenthal, A., Cox, E.C., Chisholm, R.L., Gibbs, R., Loomis, W.F., Platzer, M., Kay, R.R., Williams, J., Dear, P.H., Noegel, A.A., Barrell, B., Kuspa, A., 2005. The genome of the social amoeba Dictyostelium discoideum. Nature 435, 43–57. https://doi.org/10.1038/nature03481

Ereshefsky, M., 2000. The Poverty of the Linnaean Hierarchy: A Philosophical Study of Biological Taxonomy. Cambridge University Press.

Falkowski, P., 2012. Ocean Science: The power of plankton. Nature 483, S17–S20. https://doi.org/10.1038/483S17a

Falkowski, P.G., Fenchel, T., Delong, E.F., 2008. The Microbial Engines That Drive Earth's Biogeochemical Cycles. Science 320, 1034–1039. https://doi.org/10.1126/science.1153213

Fang, H., Huang, C., Zhao, H., Deng, M., 2015. CCLasso: correlation inference for compositional data through Lasso. Bioinformatics 31, 3172–3180. https://doi.org/10.1093/bioinformatics/btv349

Farrell, F., Soyer, O.S., Quince, C., 2018. Machine learning based prediction of functional capabilities in metagenomically assembled microbial genomes. bioRxiv 307157. https://doi.org/10.1101/307157

Faust, K., Raes, J., 2012. Microbial interactions: from networks to models. Nat. Rev. Microbiol. 10, 538–550. https://doi.org/10.1038/nrmicro2832

Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., Huttenhower, C., 2012. Microbial Co-occurrence Relationships in the Human Microbiome. PLoS Comput. Biol. 8, e1002606. https://doi.org/10.1371/journal.pcbi.1002606

Field, null, Behrenfeld, null, Randerson, null, Falkowski, null, 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. Science 281, 237–240.

Fiore-Donno, A.M., Weinert, J., Wubet, T., Bonkowski, M., 2016. Metacommunity analysis of amoeboid protists in grassland soils. Sci. Rep. 6, 19068. https://doi.org/10.1038/srep19068

Forster, D., Bittner, L., Karkar, S., Dunthorn, M., Romac, S., Audic, S., Lopez, P., Stoeck, T., Bapteste, E., 2015. Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. BMC Biol. 13. https://doi.org/10.1186/s12915-015-0125-5

Forster, D., Dunthorn, M., Mahé, F., Dolan, J.R., Audic, S., Bass, D., Bittner, L., Boutte, C., Christen, R., Claverie, J.-M., Decelle, J., Edvardsen, B., Egge, E., Eikrem, W., Gobet, A., Kooistra, W.H.C.F., Logares, R., Massana, R., Montresor, M., Not, F., Ogata, H., Pawlowski, J., Pernice, M.C., Romac, S., Shalchian-Tabrizi, K., Simon, N., Richards, T.A., Santini, S., Sarno, D., Siano, R., Vaulot, D., Wincker, P., Zingone, A., de Vargas, C., Stoeck, T., 2016a. Benthic protists: the under-charted majority. FEMS Microbiol. Ecol. 92. https://doi.org/10.1093/femsec/fiw120

Forster, D., Dunthorn, M., Stoeck, T., Mahé, F., 2016b. Comparison of three clustering approaches for detecting novel environmental microbial diversity. PeerJ 4. https://doi.org/10.7717/peerj.1692

Friedman, J., Alm, E.J., 2012. Inferring Correlation Networks from Genomic Survey Data. PLoS Comput. Biol. 8, e1002687. https://doi.org/10.1371/journal.pcbi.1002687

Friz, C.T., 1968. The biochemical composition of the free-living Amoebae Chaos chaos, amoeba dubia and Amoeba proteus. Comp. Biochem. Physiol. 26, 81–90. https://doi.org/10.1016/0010-406X(68)90314-9

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., Shallom, S.J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M.A., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., McFadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D.J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M., Barrell, B., 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419. https://doi.org/10.1038/nature01097

Gerlt, J.A., Babbitt, P.C., Jacobson, M.P., Almo, S.C., 2012. Divergent Evolution in Enolase Superfamily: Strategies for Assigning Functions. J. Biol. Chem. 287, 29–34. https://doi.org/10.1074/jbc.R111.240945

Gilbert, J.A., Meyer, F., Antonopoulos, D., Balaji, P., Brown, C. Titus, Brown, Christopher T., Desai, N., Eisen, J.A., Evers, D., Field, D., Feng, W., Huson, D., Jansson, J., Knight, R., Knight, J., Kolker, E., Konstantindis, K., Kostka, J., Kyrpides, N., Mackelprang, R., McHardy, A., Quince, C., Raes, J., Sczyrba, A., Shade, A., Stevens, R., 2010. Meeting Report: The Terabase Metagenomics Workshop and the Vision of an Earth Microbiome Project. Stand. Genomic Sci. 3, 243. https://doi.org/10.4056/sigs.1433550

Godhe, A., Asplund, M.E., Härnström, K., Saravanan, V., Tyagi, A., Karunasagar, I., 2008. Quantification of Diatom and Dinoflagellate Biomasses in Coastal Marine Seawater Samples by Real-Time PCR. Appl Env. Microbiol 74, 7174–7182. https://doi.org/10.1128/AEM.01298-08

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S.G., 1996. Life with 6000 genes. Science 274, 546, 563–567.

Grant, J.R., Katz, L.A., 2014. Building a Phylogenomic Pipeline for the Eukaryotic Tree of Life – Addressing Deep Phylogenies with Genome-Scale Data. PLOS Curr. Tree Life. https://doi.org/10.1371/currents.tol.c24b6054aebf3602748ac042ccc8f2e9

Groisillier, A., Massana, R., Valentin, K., Vaulot, D., Guillou, L., 2006. Genetic diversity and habitats of two enigmatic marine alveolate lineages. Aquat. Microb. Ecol. 42, 277–291. https://doi.org/10.3354/ame042277

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., Darzi, Y., Audic, S., Berline, L., Brum, J.R., Coelho, L.P., Espinoza, J.C.I., Malviya, S., Sunagawa, S., Dimier, C., Kandels-Lewis, S., Picheral, M., Poulain, J., Searson, S., Tara Oceans Consortium Coordinators, Stemmann, L., Not, F., Hingamp, P., Speich, S., Follows, M., Karp-Boss, L., Boss, E., Ogata, H., Pesant, S., Weissenbach, J., Wincker, P., Acinas, S.G., Bork, P., de Vargas, C., Iudicone, D., Sullivan, M.B., Raes, J., Karsenti, E., Bowler, C., Gorsky, G., 2016. Plankton networks driving carbon export in the oligotrophic ocean. Nature 532, 465–470. https://doi.org/10.1038/nature16942

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J.R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W.H.C.F., Lara, E., Le Bescot, N., Logares, R., Mahé, F., Massana, R., Montresor, M., Morard, R., Not, F., Pawlowski, J., Probert, I., Sauvadet, A.-L., Siano, R., Stoeck, T., Vaulot, D., Zimmermann, P., Christen, R., 2013. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Res. 41, D597–D604. https://doi.org/10.1093/nar/gks1160

Guillou, L., Viprey, M., Chambouvet, A., Welsh, R.M., Kirkham, A.R., Massana, R., Scanlan, D.J., Worden, A.Z., 2008. Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). Environ. Microbiol. 10, 3349–3365. https://doi.org/10.1111/j.1462-2920.2008.01731.x

Haas, B.J., Kamoun, S., Zody, M.C., Jiang, R.H.Y., Handsaker, R.E., Cano, L.M., Grabherr, M., Kodira, C.D., Raffaele, S., Torto-Alalibo, T., Bozkurt, T.O., Ah-Fong, A.M.V., Alvarado, L., Anderson, V.L., Armstrong, M.R., Avrova, A., Baxter, L., Beynon, J., Boevink, P.C., Bollmann, S.R., Bos, J.I.B., Bulone, V., Cai, G., Cakir, C., Carrington, J.C., Chawner, M., Conti, L., Costanzo, S., Ewan, R., Fahlgren, N., Fischbach, M.A., Fugelstad, J., Gilroy, E.M., Gnerre, S., Green, P.J., Grenville-Briggs, L.J., Griffith, J., Grünwald, N.J., Horn, K., Horner, N.R., Hu, C.-H., Huitema, E., Jeong, D.-H., Jones, A.M.E., Jones, J.D.G., Jones, R.W., Karlsson, E.K., Kunjeti, S.G., Lamour, K., Liu, Z., Ma, L., MacLean, D., Chibucos, M.C., McDonald, H., McWalters, J., Meijer, H.J.G., Morgan, W., Morris, P.F., Munro, C.A., O'Neill, K., Ospina-Giraldo, M., Pinzón, A., Pritchard, L., Ramsahoye, B., Ren, Q., Restrepo, S., Roy, S., Sadanandom, A., Savidor, A., Schornack, S., Schwartz, D.C., Schumann, U.D., Schwessinger, B., Seyer, L., Sharpe, T., Silvar, C., Song, J., Studholme, D.J., Sykes, S., Thines, M., van de Vondervoort, P.J.I., Phuntumart, V., Wawra, S., Weide, R., Win, J., Young, C., Zhou, S., Fry, W., Meyers, B.C., van West, P., Ristaino, J., Govers, F., Birch, P.R.J., Whisson, S.C., Judelson, H.S., Nusbaum, C., 2009. Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. Nature 461, 393–398. https://doi.org/10.1038/nature08358

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N., Regev, A., 2013. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat. Protoc. 8. https://doi.org/10.1038/nprot.2013.084

Haeckel, E. (1834-1919) A. du texte, 1899. Kunstformen der Natur / von Prof. Dr. Ernst Haeckel [WWW Document]. Gallica. URL https://gallica.bnf.fr/ark:/12148/btv1b525055842 (accessed 10.26.18).

He, L., Liu, F., Karuppiah, V., Ren, Y., Li, Z., 2014. Comparisons of the Fungal and Protistan Communities among Different Marine Sponge Holobionts by Pyrosequencing. Microb. Ecol. 67, 951–961. https://doi.org/10.1007/s00248-014-0393-6

Hidalgo, O., Pellicer, J., Christenhusz, M., Schneider, H., Leitch, A.R., Leitch, I.J., 2017. Is There an Upper Limit to Genome Size? Trends Plant Sci. 22, 567–573. https://doi.org/10.1016/j.tplants.2017.04.005

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D.A., Finstad, K.M., Amundson, R., Thomas, B.C., Banfield, J.F., 2016. A new view of the tree of life. Nat. Microbiol. 1, 16048. https://doi.org/10.1038/nmicrobiol.2016.48

Hurwitz, B.L., Youens-Clark, K., and Walls, R.L. 2017. iMicrobe. Available at: http://imicrobe.us/

Ideker, T., Nussinov, R., 2017. Network approaches and applications in biology. PLOS Comput. Biol. 13, e1005771. https://doi.org/10.1371/journal.pcbi.1005771

Jazayeri, S.M., Melgarejo-Muñoz, L.M., Romero, H.M., 2014. RNA-SEQ: A GLANCE AT TECHNOLOGIES AND METHODOLOGIES. Acta Biológica Colomb. 20. https://doi.org/10.15446/abc.v20n2.43639

Johnson, L.K., Alexander, H., Brown, C.T., 2018. Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. https://doi.org/10.1101/323576

Jones, H.L.J., Leadbeater, B.S.C., Green, J.C., 1993. Mixotrophy in marine species of Chrysochromulina (Prymnesiophyceae): ingestion and digestion of a small green flagellate. J. Mar. Biol. Assoc. U. K. 73, 283–296. https://doi.org/10.1017/S0025315400032859

Karsenti, E., Acinas, S.G., Bork, P., Bowler, C., Vargas, C.D., Raes, J., Sullivan, M., Arendt, D., Benzoni, F., Claverie, J.-M., Follows, M., Gorsky, G., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Krzic, U., Not, F., Ogata, H., Pesant, S., Reynaud, E.G., Sardet, C., Sieracki, M.E., Speich, S., Velayoudon, D., Weissenbach, J., Wincker, P., Consortium, the T.O., 2011. A Holistic Approach to Marine Eco-Systems Biology. PLOS Biol. 9, e1001177. https://doi.org/10.1371/journal.pbio.1001177

Keeling, P.J., 2014. Phylogenetic Place of Microsporidia in the Tree of Eukaryotes, in: Microsporidia. Wiley-Blackwell, pp. 195–202. https://doi.org/10.1002/9781118395264.ch5

Keeling, P.J., 2007. Ostreococcus tauri: seeing through the genes to the genome. Trends Genet. 23, 151–154. https://doi.org/10.1016/j.tig.2007.02.008

Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., Beszteri, B., Bidle, K.D., Cameron, C.T., Campbell, L., Caron, D.A., Cattolico, R.A., Collier, J.L., Coyne, K., Davy, S.K., Deschamps, P., Dyhrman, S.T., Edvardsen, B., Gates, R.D., Gobler, C.J., Greenwood, S.J., Guida, S.M., Jacobi, J.L., Jakobsen, K.S., James, E.R., Jenkins, B., John, U., Johnson, M.D., Juhl, A.R., Kamp, A., Katz, L.A., Kiene, R., Kudryavtsev, A., Leander, B.S., Lin, S., Lovejoy, C., Lynn, D., Marchetti, A., McManus, G., Nedelcu, A.M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M.A., Murray, S., Nadathur, G., Nagai, S., Ngam, P.B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M.C., Rengefors, K., Romano, G., Rumpho, M.E., Rynearson, T., Schilling, K.B., Schroeder, D.C., Simpson, A.G.B., Slamovits, C.H., Smith, D.R., Smith, G.J., Smith, S.R., Sosik, H.M., Stief, P., Theriot, E., Twary, S.N., Umale, P.E., Vaulot, D., Wawrik, B., Wheeler, G.L., Wilson, W.H., Xu, Y., Zingone, A., Worden, A.Z., 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. PLOS Biol. 12, e1001889. https://doi.org/10.1371/journal.pbio.1001889

Keeling, P.J., Campo, J. del, 2017. Marine Protists Are Not Just Big Bacteria. Curr. Biol. 27, R541–R549. https://doi.org/10.1016/j.cub.2017.03.075

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J.A., Hugenholtz, P., van der Lelie, D., Meyer, F., Stevens, R., Bailey, M.J., Gordon, J.I., Kowalchuk, G.A., Gilbert, J.A., 2012. Unlocking the potential of metagenomics through replicated experimental design. Nat. Biotechnol. 30, 513–520. https://doi.org/10.1038/nbt.2235

Koch, P., Platzer, M., Downie, B.R., 2014. RepARK—de novo creation of repeat libraries from whole-genome NGS reads. Nucleic Acids Res. 42, e80. https://doi.org/10.1093/nar/gku210

Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., Lehmann, K., Fernandez-Guerra, A., Jeanthon, C., Rahav, E., Ullrich, M., Wichels, A., Gerdts, G., Polymenakou, P., Kotoulas, G., Siam, R., Abdallah, R.Z., Sonnenschein, E.C., Cariou, T., O'Gara, F., Jackson, S., Orlic, S., Steinke, M., Busch, J., Duarte, B., Caçador, I., Canning-Clode, J., Bobrova, O., Marteinsson, V., Reynisson, E., Loureiro, C.M., Luna, G.M., Quero, G.M., Löscher, C.R., Kremp, A., DeLorenzo, M.E., Øvreås, L., Tolman, J., LaRoche, J., Penna, A., Frischer, M., Davis, T., Katherine, B., Meyer, C.P., Ramos, S., Magalhães, C., Jude-Lemeilleur, F., Aguirre-Macedo, M.L., Wang, S., Poulton, N., Jones, S., Collin, R., Fuhrman, J.A., Conan, P., Alonso, C., Stambler, N., Goodwin, K., Yakimov, M.M., Baltar, F., Bodrossy, L., Van De Kamp, J., Frampton, D.M., Ostrowski, M., Van Ruth, P., Malthouse, P., Claus, S., Deneudt, K., Mortelmans, J., Pitois, S., Wallom, D., Salter, I., Costa, R., Schroeder, D.C., Kandil, M.M., Amaral, V., Biancalana, F., Santana, R., Pedrotti, M.L., Yoshida, T., Ogata, H., Ingleton, T., Munnik, K., Rodriguez-Ezpeleta, N., Berteaux-Lecellier, V., Wecker, P., Cancio, I., Vaulot, D., Bienhold, C., Ghazal, H., Chaouni, B., Essayeh, S., Ettamimi, S., Zaid, E.H., Boukhatem, N., Bouali, A., Chahboune, R., Barrijal, S., Timinouni, M., El Otmani, F., Bennani, M., Mea, M., Todorova, N., Karamfilov, V., ten Hoopen, P., Cochrane, G., L'Haridon, S., Bizsel, K.C., Vezzi, A., Lauro, F.M., Martin, P., Jensen, R.M., Hinks, J., Gebbels, S., Rosselli, R., De Pascale, F., Schiavon, R., dos Santos, A., Villar, E., Pesant, S., Cataletto, B., Malfatti, F., Edirisinghe, R., Silveira, J.A.H., Barbier, M., Turk, V., Tinta, T., Fuller, W.J., Salihoglu, I., Serakinci, N., Ergoren, M.C., Bresnan, E., Iriberri, J., Nyhus, P.A.F., Bente, E., Karlsen, H.E., Golyshin, P.N., Gasol, J.M., Moncheva, S., Dzhembekova, N., Johnson, Z., Sinigalliano, C.D., Gidley, M.L., Zingone, A., Danovaro, R., Tsiamis, G., Clark, M.S., Costa, A.C., El Bour, M., Martins, A.M., Collins, R.E., Ducluzeau, A.-L., Martinez, J., Costello, M.J., Amaral-Zettler, L.A., Gilbert, J.A., Davies, N., Field, D., Glöckner, F.O., 2015. The ocean sampling day consortium. GigaScience 4. https://doi.org/10.1186/s13742-015-0066-5

Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., Bonneau, R.A., 2015. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLOS Comput. Biol. 11, e1004226. https://doi.org/10.1371/journal.pcbi.1004226

Lakeman, M.B., von Dassow, P., Cattolico, R.A., 2009. The strain concept in phytoplankton ecology. Harmful Algae, This issue contains the special section on "Strains" 8, 746–758. https://doi.org/10.1016/j.hal.2008.11.011

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E.,

Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J., Szustakowki, J., International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921. https://doi.org/10.1038/35057062

Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559. https://doi.org/10.1186/1471-2105-9-559

Lanier, W., Moustafa, A., Bhattacharya, D., Comeron, J.M., 2008. EST Analysis of Ostreococcus lucimarinus, the Most Compact Eukaryotic Genome, Shows an Excess of Introns in Highly Expressed Genes. PLoS ONE 3. https://doi.org/10.1371/journal.pone.0002171

Lau, Y.-L., Lee, W.-C., Gudimella, R., Zhang, G., Ching, X.-T., Razali, R., Aziz, F., Anwar, A., Fong, M.-Y., 2016. Deciphering the Draft Genome of Toxoplasma gondii RH Strain. PLOS ONE 11, e0157901. https://doi.org/10.1371/journal.pone.0157901

Leinonen, R., Sugawara, H., Shumway, M., International Nucleotide Sequence Database Collaboration, 2011. The sequence read archive. Nucleic Acids Res. 39, D19-21. https://doi.org/10.1093/nar/gkq1019

Lie, A.A.Y., Liu, Z., Terrado, R., Tatters, A.O., Heidelberg, K.B., Caron, D.A., 2018. A tale of two mixotrophic chrysophytes: Insights into the metabolisms of two Ochromonas species (Chrysophyceae) through a comparison of gene expression. PLoS ONE 13. https://doi.org/10.1371/journal.pone.0192439

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J.C., Roux, S., Vincent, F., Bittner, L., Darzi, Y., Wang, J., Audic, S., Berline, L., Bontempi, G., Cabello, A.M., Coppola, L., Cornejo-Castillo, F.M., d'Ovidio, F., Meester, L.D., Ferrera, I., Garet-Delmas, M.-J., Guidi, L., Lara, E., Pesant, S., Royo-Llonch, M., Salazar, G., Sánchez, P., Sebastian, M., Souffreau, C., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Coordinators, T.O., Gorsky, G., Not, F., Ogata, H., Speich, S., Stemmann, L., Weissenbach, J., Wincker, P., Acinas, S.G., Sunagawa, S., Bork, P., Sullivan, M.B., Karsenti, E., Bowler, C., Vargas, C. de, Raes, J., 2015. Determinants of community structure in the global plankton interactome. Science 348, 1262073. https://doi.org/10.1126/science.1262073

Lin, D., Yin, X., Wang, X., Zhou, P., Guo, F.-B., 2013. Re-Annotation of Protein-Coding Genes in the Genome of Saccharomyces cerevisiae Based on Support Vector Machines. PLoS ONE 8. https://doi.org/10.1371/journal.pone.0064477

Liu, H., Probert, I., Uitz, J., Claustre, H., Aris-Brosou, S., Frada, M., Not, F., Vargas, C. de, 2009. Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans. Proc. Natl. Acad. Sci. 106, 12803–12808. https://doi.org/10.1073/pnas.0905841106

Loftus, B., Anderson, I., Davies, R., Alsmark, U.C.M., Samuelson, J., Amedeo, P., Roncaglia, P., Berriman, M., Hirt, R.P., Mann, B.J., Nozaki, T., Suh, B., Pop, M., Duchene, M., Ackers, J., Tannich, E., Leippe, M., Hofer, M., Bruchhaus, I., Willhoeft, U., Bhattacharya, A., Chillingworth, T., Churcher, C., Hance, Z., Harris, B., Harris, D., Jagels, K., Moule, S., Mungall, K., Ormond, D., Squares, R., Whitehead, S., Quail, M.A., Rabbinowitsch, E., Norbertczak, H., Price, C., Wang, Z., Guillén, N., Gilchrist, C., Stroup, S.E., Bhattacharya, S., Lohia, A., Foster, P.G., Sicheritz-Ponten, T., Weber, C., Singh, U., Mukherjee, C., El-Sayed, N.M., Petri Jr, W.A., Clark, C.G., Embley, T.M., Barrell, B., Fraser, C.M., Hall, N., 2005. The genome of the protist parasite Entamoeba histolytica. Nature 433, 865–868. https://doi.org/10.1038/nature03291

Logares, R., Audic, S., Bass, D., Bittner, L., Boutte, C., Christen, R., Claverie, J.-M., Decelle, J., Dolan, J.R., Dunthorn, M., Edvardsen, B., Gobet, A., Kooistra, W.H.C.F., Mahé, F., Not, F., Ogata, H., Pawlowski, J., Pernice, M.C., Romac, S., Shalchian-Tabrizi, K., Simon, N., Stoeck, T., Santini, S., Siano, R., Wincker, P., Zingone, A., Richards, T.A., de Vargas, C., Massana, R., 2014a. Patterns of Rare and Abundant Marine Microbial Eukaryotes. Curr. Biol. 24, 813–821. https://doi.org/10.1016/j.cub.2014.02.050

Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F.M., Ferrera, I., Sarmento, H., Hingamp, P., Ogata, H., Vargas, C. de, Lima-Mendez, G., Raes, J., Poulain, J., Jaillon, O., Wincker, P., Kandels-Lewis, S., Karsenti, E., Bork, P., Acinas, S.G., 2014b. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. Environ. Microbiol. 16, 2659–2671. https://doi.org/10.1111/1462-2920.12250

Lopez, P., Halary, S., Bapteste, E., 2015. Highly divergent ancient gene families in metagenomic samples are compatible with additional divisions of life. Biol. Direct 10. https://doi.org/10.1186/s13062-015-0092-3

López-García, P., Philippe, H., Gail, F., Moreira, D., 2003. Autochthonous eukaryotic diversity in hydrothermal sediment and experimental microcolonizers at the Mid-Atlantic Ridge. Proc. Natl. Acad. Sci. 100, 697–702. https://doi.org/10.1073/pnas.0235779100

López-García, P., Rodríguez-Valera, F., Pedrós-Alió, C., Moreira, D., 2001. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. Nature 409, 603–607. https://doi.org/10.1038/35054537

Mahé, F., de Vargas, C., Bass, D., Czech, L., Stamatakis, A., Lara, E., Singer, D., Mayor, J., Bunge, J., Sernaker, S., Siemensmeyer, T., Trautmann, I., Romac, S., Berney, C., Kozlov, A., Mitchell, E.A.D., Seppey, C.V.W., Egge, E., Lentendu, G., Wirth, R., Trueba, G., Dunthorn, M., 2017. Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. Nat. Ecol. Evol. 1, 0091. https://doi.org/10.1038/s41559-017-0091

Mahé, F., Mayor, J., Bunge, J., Chi, J., Siemensmeyer, T., Stoeck, T., Wahl, B., Paprotka, T., Filker, S., Dunthorn, M., 2015. Comparing High-throughput Platforms for Sequencing the V4 Region of SSU-rDNA in Environmental Microbial Eukaryotic Diversity Surveys. J. Eukaryot. Microbiol. 62, 338–345. https://doi.org/10.1111/jeu.12187

Mahé, F., Rognes, T., Quince, C., Vargas, C. de, Dunthorn, M., 2014. Swarm: robust and fast clustering method for amplicon-based studies. PeerJ 2, e593. https://doi.org/10.7717/peerj.593

Malod-Dognin, N., Pržulj, N., 2015. L-GRAAL: Lagrangian graphlet-based network aligner. Bioinformatics 31, 2182–2189. https://doi.org/10.1093/bioinformatics/btv130

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., Wincker, P., Iudicone, D., Vargas, C. de, Bittner, L., Zingone, A., Bowler, C., 2016. Insights into global diatom distribution and diversity in the world's ocean. Proc. Natl. Acad. Sci. 113, E1516–E1525. https://doi.org/10.1073/pnas.1509523113

Mandakovic, D., Rojas, C., Maldonado, J., Latorre, M., Travisany, D., Delage, E., Bihouée, A., Jean, G., Díaz, F.P., Fernández-Gómez, B., Cabrera, P., Gaete, A., Latorre, C., Gutiérrez, R.A., Maass, A., Cambiazo, V., Navarrete, S.A., Eveillard, D., González, M., 2018. Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. Sci. Rep. 8, 5875.

https://doi.org/10.1038/s41598-018-23931-0
Mangot, J.-F., Domaizon, I., Taib, N., Marouni, N., Duffaud, E., Bronner, G., Debroas, D., 2013. Short-term dynamics of diversity patterns: evidence of continual reassembly within lacustrine small eukaryotes. Environ. Microbiol. 15, 1745–1758. https://doi.org/10.1111/1462-2920.12065
Marchet, C., Lecompte, L., Limasset, A., Bittner, L., Peterlongo, P., 2018. A resource-frugal probabilistic dictionary and applications in bioinformatics. Discrete Appl. Math. https://doi.org/10.1016/j.dam.2018.03.035
Massana, R., Castresana, J., Balagué, V., Guillou, L., Romari, K., Groisillier, A., Valentin, K., Pedrós-Alió, C., 2004. Phylogenetic and Ecological Analysis of Novel Marine Stramenopiles. Appl. Environ. Microbiol. 70, 3528–3534. https://doi.org/10.1128/AEM.70.6.3528-3534.2004
Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., Chambouvet, A., Christen, R., Claverie, J.-M., Decelle, J., Dolan, J.R., Dunthorn, M., Edvardsen, B., Forn, I., Forster, D., Guillou, L., Jaillon, O., Kooistra, W.H.C.F., Logares, R., Mahé, F., Not, F., Ogata, H., Pawlowski, J., Pernice, M.C., Probert, I., Romac, S., Richards, T., Santini, S., Shalchian-Tabrizi, K., Siano, R., Simon, N., Stoeck, T., Vaulot, D., Zingone, A., de Vargas, C., 2015. Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. Environ. Microbiol. 17, 4035–4049. https://doi.org/10.1111/1462-2920.12955
Massana, R., Guillou, L., Díez, B., Pedrós-Alió, C., 2002. Unveiling the Organisms behind Novel Eukaryotic Ribosomal DNA Sequences from the Ocean. Appl. Environ. Microbiol. 68, 4554–4558. https://doi.org/10.1128/AEM.68.9.4554-4558.2002
Matasci, N., Hung, L.-H., Yan, Z., Carpenter, E.J., Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Ayyampalayam, S., Barker, M., Burleigh, J.G., Gitzendanner, M.A., Wafula, E., Der, J.P., dePamphilis, C.W., Roure, B., Philippe, H., Ruhfel, B.R., Miles, N.W., Graham, S.W., Mathews, S., Surek, B., Melkonian, M., Soltis, D.E., Soltis, P.S., Rothfels, C., Pokorny, L., Shaw, J.A., DeGironimo, L., Stevenson, D.W., Villarreal, J.C., Chen, T., Kutchan, T.M., Rolf, M., Baucom, R.S., Deyholos, M.K., Samudrala, R., Tian, Z., Wu, X., Sun, X., Zhang, Y., Wang, J., Leebens-Mack, J., Wong, G.K.-S., 2014. Data access for the 1,000 Plants (1KP) project. GigaScience 3. https://doi.org/10.1186/2047-217X-3-17
Méheust, R., Zelzion, E., Bhattacharya, D., Lopez, P., Bapteste, E., 2016. Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. Proc. Natl. Acad. Sci. 113, 3579–3584. https://doi.org/10.1073/pnas.1517551113
Meng, A., Corre, E., Probert, I., Gutierrez-Rodriguez, A., Siano, R., Annamale, A., Alberti, A., Da Silva, C., Wincker, P., Le Crom, S., Not, F., Bittner, L., 2018a. Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome-based sequence similarity network. Mol. Ecol. https://doi.org/10.1111/mec.14579
Meng, A., Marchet, C., Corre, E., Peterlongo, P., Alberti, A., Da Silva, C., Wincker, P., Pelletier, E., Probert, I., Decelle, J., Le Crom, S., Not, F., Bittner, L., 2018b. A de novo approach to disentangle partner identity and function in holobiont systems. Microbiome 6, 105. https://doi.org/10.1186/s40168-018-0481-9
Mirsky, A.E., Ris, H., 1951. THE DESOXYRIBONUCLEIC ACID CONTENT OF ANIMAL CELLS AND ITS EVOLUTIONARY SIGNIFICANCE. J. Gen. Physiol. 34, 451–462.
Mitchell, A., Bucchini, F., Cochrane, G., Denise, H., ten Hoopen, P., Fraser, M., Pesseat, S., Potter, S., Scheremetjew, M., Sterk, P., Finn, R.D., 2016. EBI metagenomics in 2016--an expanding and evolving resource for the analysis and archiving of metagenomic data. Nucleic Acids Res. 44, D595-603. https://doi.org/10.1093/nar/gkv1195
Moon-van der Staay, S.Y., De Wachter, R., Vaulot, D., 2001. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. Nature 409, 607–610. https://doi.org/10.1038/35054541
Mordret, S., Romac, S., Henry, N., Colin, S., Carmichael, M., Berney, C., Audic, S., Richter, D.J., Pochon, X., de Vargas, C., Decelle, J., 2016. The symbiotic life of Symbiodinium in the open ocean within a new species of calcifying ciliate (Tiarina sp.). ISME J. 10, 1424–1436. https://doi.org/10.1038/ismej.2015.211
Murray, S.A., Suggett, D.J., Doblin, M.A., Kohli, G.S., Seymour, J.R., Fabris, M., Ralph, P.J., 2016. Unravelling the functional genetics of dinoflagellates: a review of approaches and opportunities. Perspect. Phycol. 3, 37–52. https://doi.org/10.1127/pip/2016/0039
Nielsen, H.B., Almeida, M., Juncker, Agnieszka Sierakowska, Rasmussen, S., Li, J., Sunagawa, S., Plichta, D.R., Gautier, L., Pedersen, A.G., Chatelier, E.L., Pelletier, E., Bonde, I., Nielsen, T., Manichanh, C., Arumugam, M., Batto, J.-M., Santos, M.B.Q. dos, Blom, N., Borruel, N., Burgdorf, K.S., Boumezbeur, F., Casellas, F., Doré, J., Dworzynski, P., Guarner, F., Hansen, T., Hildebrand, F., Kaas, R.S., Kennedy, S., Kristiansen, K., Kultima, J.R., Léonard, P., Levenez, F., Lund, O., Moumen, B., Paslier, D.L., Pons, N., Pedersen, O., Prifti, E., Qin, J., Raes, J., Sørensen, S., Tap, J., Tims, S., Ussery, D.W., Yamada, T., Consortium, M., Nielsen, H.B., Almeida, M., Juncker, Agnieszka S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D.R., Gautier, L., Pedersen, A.G., Chatelier, E.L., Pelletier, E., Bonde, I., Nielsen, T., Manichanh, C., Arumugam, M., Batto, J.-M., Santos, M.B.Q. dos, Blom, N., Borruel, N., Burgdorf, K.S., Boumezbeur, F., Casellas, F., Doré, J., Dworzynski, P., Guarner, F., Hansen, T., Hildebrand, F., Kaas, R.S., Kennedy, S., Kristiansen, K., Kultima, J.R., Leonard, P., Levenez, F., Lund, O., Moumen, B., Paslier, D.L., Pons, N., Pedersen, O., Prifti, E., Qin, J., Raes, J., Sørensen, S., Tap, J., Tims, S., Ussery, D.W., Yamada, T., Renault, P., Sicheritz-Ponten, T., Bork, P., Wang, J., Brunak, S., Ehrlich, S.D., Jamet, A., Mérieux, A., Cultrone, A., Torrejon, A., Quinquis, B., Brechot, C., Delorme, C., M'Rini, C., Vos, W.M. de, Maguin, E., Varela, E., Guedon, E., Gwen, F., Haimet, F., Artiguenave, F., Vandemeulebrouck, G., Denariaz, G., Khaci, G., Blottière, H., Knol, J., Weissenbach, J., Vlieg, J.E.T. van H., Torben, J., Parkhill, J., Turner, K., Guchte, M. van de, Antolin, M., Rescigno, M., Kleerebezem, M., Derrien, M., Galleron, N., Sanchez, N., Grarup, N., Veiga, P., Oozeer, R., Dervyn, R., Layec, S., Bruls, T., Winogradski, Y., G, Z.E., Renault, P., Sicheritz-Ponten, T., Bork, P., Wang, J., Brunak, S., Ehrlich, S.D., 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. 32, 822–828. https://doi.org/10.1038/nbt.2939
Not, F., Campo, J. del, Balagué, V., Vargas, C. de, Massana, R., 2009. New Insights into the Diversity of Marine Picoeukaryotes. PLOS ONE 4, e7143. https://doi.org/10.1371/journal.pone.0007143
Not, F., Gausling, R., Azam, F., Heidelberg, J.F., Worden, A.Z., 2007a. Vertical distribution of picoeukaryotic diversity in the Sargasso Sea. Environ. Microbiol. 9, 1233–1252. https://doi.org/10.1111/j.1462-2920.2007.01247.x
Not, F., Valentin, K., Romari, K., Lovejoy, C., Massana, R., Töbe, K., Vaulot, D., Medlin, L.K., 2007b. Picobiliphytes: A Marine Picoplanktonic Algal Group with Unknown Affinities to Other Eukaryotes. Science 315, 253–255. https://doi.org/10.1126/science.1136264
Ottesen, E.A., 2016. Probing the living ocean with ecogenomic sensors. Curr. Opin. Microbiol., Environmental microbiology * Special Section: Megaviromes 31, 132–139. https://doi.org/10.1016/j.mib.2016.03.012
Ottesen, E.A., Marin, R., Preston, C.M., Young, C.R., Ryan, J.P., Scholin, C.A., DeLong, E.F., 2011. Metatranscriptomic analysis of autonomously collected and preserved marine bacterioplankton. ISME J. 5, 1881–1895. https://doi.org/10.1038/ismej.2011.70
Pagani, I., Liolios, K., Jansson, J., Chen, I.-M.A., Smirnova, T., Nosrat, B., Markowitz, V.M., Kyrpides, N.C., 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 40, D571–D579. https://doi.org/10.1093/nar/gkr1100
Pais, M., Win, J., Yoshida, K., Etherington, G.J., Cano, L.M., Raffaele, S., Banfield, M.J., Jones, A., Kamoun, S., Saunders, D.G., 2013. From pathogen genomes to host plant processes: the power of plant parasitic oomycetes. Genome Biol. 14, 211. https://doi.org/10.1186/gb-2013-14-6-211

Parfrey, L.W., Lahr, D.J.G., Katz, L.A., 2008. The dynamic nature of eukaryotic genomes. Mol. Biol. Evol. 25, 787–794. https://doi.org/10.1093/molbev/msn032

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., Tyson, G.W., 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. 2, 1533–1542. https://doi.org/10.1038/s41564-017-0012-7

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., Bowser, S.S., Cepicka, I., Decelle, J., Dunthorn, M., Fiore-Donno, A.M., Gile, G.H., Holzmann, M., Jahn, R., Jirků, M., Keeling, P.J., Kostka, M., Kudryavtsev, A., Lara, E., Lukeš, J., Mann, D.G., Mitchell, E.A.D., Nitsche, F., Romeralo, M., Saunders, G.W., Simpson, A.G.B., Smirnov, A.V., Spouge, J.L., Stern, R.F., Stoeck, T., Zimmermann, J., Schindel, D., Vargas, C. de, 2012. CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. PLOS Biol. 10, e1001419. https://doi.org/10.1371/journal.pbio.1001419

Pedrós-Alió, C., 2006. Marine microbial diversity: can it be determined? Trends Microbiol. 14, 257–263. https://doi.org/10.1016/j.tim.2006.04.007

Perkel, J.M., 2018. Machine learning gets to grips with plankton challenge. Nature 561, 567. https://doi.org/10.1038/d41586-018-06792-5

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., Iudicone, D., Karsenti, E., Speich, S., Troublé, R., Dimier, C., Searson, S., Tara Oceans Consortium Coordinators, Acinas, S.G., Bork, P., Boss, E., Bowler, C., Vargas, C.D., Follows, M., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Karp-Boss, L., Karsenti, E., Krzic, U., Not, F., Ogata, H., Pesant, S., Raes, J., Reynaud, E.G., Sardet, C., Sieracki, M., Speich, S., Stemmann, L., Sullivan, M.B., Sunagawa, S., Velayoudon, D., Weissenbach, J., Wincker, P., 2015. Open science resources for the discovery and analysis of Tara Oceans data. Sci. Data 2, 150023. https://doi.org/10.1038/sdata.2015.23

Pettersson, E., Lundeberg, J., Ahmadian, A., 2009. Generations of sequencing technologies. Genomics 93, 105–111. https://doi.org/10.1016/j.ygeno.2008.10.003

Pichard, S.L., Campbell, L., Paul, J.H., 1997. Diversity of the Ribulose Bisphosphate Carboxylase/Oxygenase Form I Gene (rbcL) in Natural Phytoplankton Communities 63, 7.

Piganeau, G., Desdevises, Y., Derelle, E., Moreau, H., 2008. Picoeukaryotic sequences in the Sargasso Sea metagenome. Genome Biol. 9, R5. https://doi.org/10.1186/gb-2008-9-1-r5

Piganeau, G., Eyre-Walker, A., Grimsley, N., Moreau, H., 2011. How and Why DNA Barcodes Underestimate the Diversity of Microbial Eukaryotes. PLOS ONE 6, e16342. https://doi.org/10.1371/journal.pone.0016342

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, Shaochuan, Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, Shengting, Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, Songgang, Qin, N., Yang, H., Wang, Jian, Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., MetaHIT Consortium, Antolin, M., Artiguenave, F., Blottiere, H., Borruel, N., Bruls, T., Casellas, F., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Forte, M., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Jamet, A., Juste, C., Kaci, G., Kleerebezem, M., Knol, J., Kristensen, M., Layec, S., Le Roux, K., Leclerc, M., Maguin, E., Melo Minardi, R., Oozeer, R., Rescigno, M., Sanchez, N., Tims, S., Torrejon, T., Varela, E., de Vos, W., Winogradsky, Y., Zoetendal, E., Bork, P., Ehrlich, S.D., Wang, Jun, 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59–65. https://doi.org/10.1038/nature08821

Rappé, M.S., Suzuki, M.T., Vergin, K.L., Giovannoni, S.J., 1998. Phylogenetic Diversity of Ultraplankton Plastid Small-Subunit rRNA Genes Recovered in Environmental Nucleic Acid Samples from the Pacific and Atlantic Coasts of the United States. Appl. Environ. Microbiol. 64, 294–303.

Read, B.A., Kegel, J., Klute, M.J., Kuo, A., Lefebvre, S.C., Maumus, F., Mayer, C., Miller, J., Monier, A., Salamov, A., Young, J., Aguilar, M., Claverie, J.-M., Frickenhaus, S., Gonzalez, K., Herman, E.K., Lin, Y.-C., Napier, J., Ogata, H., Sarno, A.F., Shmutz, J., Schroeder, D., Vargas, C. de, Verret, F., Dassow, P. von, Valentin, K., Peer, Y.V. de, Wheeler, G., Consortium, E. huxleyi A., Allen, A.E., Bidle, K., Borodovsky, M., Bowler, C., Brownlee, C., Cock, J.M., Elias, M., Gladyshev, V.N., Groth, M., Guda, C., Hadaegh, A., Iglesias-Rodriguez, M.D., Jenkins, J., Jones, B.M., Lawson, T., Leese, F., Lindquist, E., Lobanov, A., Lomsadze, A., Malik, S.-B., Marsh, M.E., Mackinder, L., Mock, T., Mueller-Roeber, B., Pagarete, A., Parker, M., Probert, I., Quesneville, H., Raines, C., Rensing, S.A., Riaño-Pachón, D.M., Richier, S., Rokitta, S., Shiraiwa, Y., Soanes, D.M., Giezen, M. van der, Wahlund, T.M., Williams, B., Wilson, W., Wolfe, G., Wurch, L.L., Dacks, J.B., Delwiche, C.F., Dyhrman, S.T., Glöckner, G., John, U., Richards, T., Worden, A.Z., Zhang, X., Grigoriev, I.V., 2013. Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature 499, 209–213. https://doi.org/10.1038/nature12221

Reuter, J.A., Spacek, D., Snyder, M.P., 2015. High-Throughput Sequencing Technologies. Mol. Cell 58, 586–597. https://doi.org/10.1016/j.molcel.2015.05.004

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., Dodsworth, J.A., Hedlund, B.P., Tsiamis, G., Sievert, S.M., Liu, W.-T., Eisen, J.A., Hallam, S.J., Kyrpides, N.C., Stepanauskas, R., Rubin, E.M., Hugenholtz, P., Woyke, T., 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499, 431–437. https://doi.org/10.1038/nature12352

Rodríguez-Martínez, R., Labrenz, M., Campo, J.D., Forn, I., Jürgens, K., Massana, R., 2009. Distribution of the uncultured protist MAST-4 in the Indian Ocean, Drake Passage and Mediterranean Sea assessed by real-time quantitative PCR. Environ. Microbiol. 11, 397–408. https://doi.org/10.1111/j.1462-2920.2008.01779.x

Rogato, A., Richard, H., Sarazin, A., Voss, B., Cheminant Navarro, S., Champeimont, R., Navarro, L., Carbone, A., Hess, W.R., Falciatore, A., 2014. The diversity of small non-coding RNAs in the diatom Phaeodactylum tricornutum. BMC Genomics 15. https://doi.org/10.1186/1471-2164-15-698

Romari, K., Vaulot, D., 2004. Composition and temporal variability of picoeukaryote communities at a coastal site of the English Channel from 18S rDNA sequences. Limnol. Oceanogr. 49, 784–798. https://doi.org/10.4319/lo.2004.49.3.0784

Röttjers, L., Faust, K., 2018. From hairballs to hypotheses–biological insights from microbial networks. FEMS Microbiol. Rev. 42, 761–780. https://doi.org/10.1093/femsre/fuy030

Roux, S., Brum, J.R., Dutilh, B.E., Sunagawa, S., Duhaime, M.B., Loy, A., Poulos, B.T., Solonenko, N., Lara, E., Poulain, J., Pesant, S., Kandels-Lewis, S., Dimier, C., Picheral, M., Searson, S., Cruaud, C., Alberti, A., Duarte, C.M., Gasol, J.M., Vaqué, D., Tara Oceans Coordinators, Bork, P., Acinas, S.G., Wincker, P., Sullivan, M.B., 2016. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature 537, 689–693. https://doi.org/10.1038/nature19366

Sauzède, R., Bittig, H.C., Claustre, H., Pasqueron de Fommervault, O., Gattuso, J.-P., Legendre, L., Johnson, K.S., 2017. Estimates of Water-Column Nutrient Concentrations and Carbonate System Parameters in the Global Ocean: A Novel Approach Based on Neural Networks. Front. Mar. Sci. 4. https://doi.org/10.3389/fmars.2017.00128

Sboner, A., Mu, X.J., Greenbaum, D., Auerbach, R.K., Gerstein, M.B., 2011. The real cost of sequencing: higher than you think! Genome Biol. 12,

125. https://doi.org/10.1186/gb-2011-12-8-125
Scamardella, J.M., 1999. Not plants or animals: a brief history of the origin of Kingdoms Protozoa, Protista and Protoctista. Int. Microbiol. Off. J. Span. Soc. Microbiol. 2, 207–216.
Seenivasan, R., Sausen, N., Medlin, L.K., Melkonian, M., 2013. Picomonas judraskeda Gen. Et Sp. Nov.: The First Identified Member of the Picozoa Phylum Nov., a Widespread Group of Picoeukaryotes, Formerly Known as 'Picobiliphytes.' PLoS ONE 8, e59565. https://doi.org/10.1371/journal.pone.0059565
Ser-Giacomi, E., Zinger, L., Malviya, S., Vargas, C.D., Karsenti, E., Bowler, C., Monte, S.D., 2018. Ubiquitous abundance distribution of non-dominant plankton across the global ocean. Nat. Ecol. Evol. 2, 1243. https://doi.org/10.1038/s41559-018-0587-2
Shi, X.L., Marie, D., Jardillier, L., Scanlan, D.J., Vaulot, D., 2009. Groups without Cultured Representatives Dominate Eukaryotic Picophytoplankton in the Oligotrophic South East Pacific Ocean. PLoS ONE 4. https://doi.org/10.1371/journal.pone.0007657
Sibbald, S.J., Archibald, J.M., 2017. More protist genomes needed. Nat. Ecol. Evol. https://doi.org/10.1038/s41559-017-0145
Simpson, A.G.B., Roger, A.J., 2004. The real "kingdoms" of eukaryotes. Curr. Biol. CB 14, R693-696. https://doi.org/10.1016/j.cub.2004.08.038
Siva, N., 2008. 1000 Genomes project. Nat. Biotechnol. 26, 256. https://doi.org/10.1038/nbt0308-256b
Sogin, M.L., 1991. Early evolution and the origin of eukaryotes. Curr. Opin. Genet. Dev. 1, 457–463.
Speijer, D., Lukeš, J., Eliáš, M., 2015. Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. Proc. Natl. Acad. Sci. U. S. A. 112, 8827–8834. https://doi.org/10.1073/pnas.1501725112
Steele, J.A., Countway, P.D., Xia, L., Vigil, P.D., Beman, J.M., Kim, D.Y., Chow, C.-E.T., Sachdeva, R., Jones, A.C., Schwalbach, M.S., Rose, J.M., Hewson, I., Patel, A., Sun, F., Caron, D.A., Fuhrman, J.A., 2011. Marine bacterial, archaeal and protistan association networks reveal ecological linkages. ISME J. 5, 1414–1425. https://doi.org/10.1038/ismej.2011.24
Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M.D.M., Breiner, H.-W., Richards, T.A., 2010. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol. Ecol. 19, 21–31. https://doi.org/10.1111/j.1365-294X.2009.04480.x
Stoeck, T., Zuendorf, A., Breiner, H.-W., Behnke, A., 2007. A molecular approach to identify active microbes in environmental eukaryote clone libraries. Microb. Ecol. 53, 328–339. https://doi.org/10.1007/s00248-006-9166-1
Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., Cornejo-Castillo, F.M., Costea, P.I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J.M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B.T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Coordinators, T.O., Bowler, C., Vargas, C. de, Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M.B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S.G., Bork, P., 2015. Structure and function of the global ocean microbiome. Science 348, 1261359. https://doi.org/10.1126/science.1261359
Swart, E.C., Bracht, J.R., Magrini, V., Minx, P., Chen, X., Zhou, Y., Khurana, J.S., Goldman, A.D., Nowacki, M., Schotanus, K., Jung, S., Fulton, R.S., Ly, A., McGrath, S., Haub, K., Wiggins, J.L., Storton, D., Matese, J.C., Parsons, L., Chang, W.-J., Bowen, M.S., Stover, N.A., Jones, T.A., Eddy, S.R., Herrick, G.A., Doak, T.G., Wilson, R.K., Mardis, E.R., Landweber, L.F., 2013. The Oxytricha trifallax Macronuclear Genome: A Complex Eukaryotic Genome with 16,000 Tiny Chromosomes. PLOS Biol. 11, e1001473. https://doi.org/10.1371/journal.pbio.1001473
Tackmann, J., Arora, N., Schmidt, T.S.B., Rodrigues, J.F.M., von Mering, C., 2018. Ecologically informed microbial biomarkers and accurate classification of mixed and unmixed samples in an extensive cross-study of human body sites. Microbiome 6, 192. https://doi.org/10.1186/s40168-018-0565-6
Tekle, Y.I., Parfrey, L.W., Katz, L.A., 2009. Molecular Data are Transforming Hypotheses on the Origin and Diversification of Eukaryotes. Bioscience 59, 471–481. https://doi.org/10.1525/bio.2009.59.6.5
The 1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. Nature 526, 68–74. https://doi.org/10.1038/nature15393
The MetaSUB International Consortium, 2016. The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report. Microbiome 4. https://doi.org/10.1186/s40168-016-0168-z
Thompson, L.R., Sanders, J.G., McDonald, D., Amir, A., Ladau, J., Locey, K.J., Prill, R.J., Tripathi, A., Gibbons, S.M., Ackermann, G., Navas-Molina, J.A., Janssen, S., Kopylova, E., Vázquez-Baeza, Y., González, A., Morton, J.T., Mirarab, S., Zech Xu, Z., Jiang, L., Haroon, M.F., Kanbar, J., Zhu, Q., Jin Song, S., Kosciolek, T., Bokulich, N.A., Lefler, J., Brislawn, C.J., Humphrey, G., Owens, S.M., Hampton-Marcell, J., Berg-Lyons, D., McKenzie, V., Fierer, N., Fuhrman, J.A., Clauset, A., Stevens, R.L., Shade, A., Pollard, K.S., Goodwin, K.D., Jansson, J.K., Gilbert, J.A., Knight, R., The Earth Microbiome Project Consortium, Rivera, J.L.A., Al-Moosawi, L., Alverdy, J., Amato, K.R., Andras, J., Angenent, L.T., Antonopoulos, D.A., Apprill, A., Armitage, D., Ballantine, K., Bárta, J., Baum, J.K., Berry, A., Bhatnagar, A., Bhatnagar, M., Biddle, J.F., Bittner, L., Boldgiv, B., Bottos, E., Boyer, D.M., Braun, J., Brazelton, W., Brearley, F.Q., Campbell, A.H., Caporaso, J.G., Cardona, C., Carroll, J., Cary, S.C., Casper, B.B., Charles, T.C., Chu, H., Claar, D.C., Clark, R.G., Clayton, J.B., Clemente, J.C., Cochran, A., Coleman, M.L., Collins, G., Colwell, R.R., Contreras, M., Crary, B.B., Creer, S., Cristol, D.A., Crump, B.C., Cui, D., Daly, S.E., Davalos, L., Dawson, R.D., Defazio, J., Delsuc, F., Dionisi, H.M., Dominguez-Bello, M.G., Dowell, R., Dubinsky, E.A., Dunn, P.O., Ercolini, D., Espinoza, R.E., Ezenwa, V., Fenner, N., Findlay, H.S., Fleming, I.D., Fogliano, V., Forsman, A., Freeman, C., Friedman, E.S., Galindo, G., Garcia, L., Garcia-Amado, M.A., Garshelis, D., Gasser, R.B., Gerdts, G., Gibson, M.K., Gifford, I., Gill, R.T., Giray, T., Gittel, A., Golyshin, P., Gong, D., Grossart, H.-P., Guyton, K., Haig, S.-J., Hale, V., Hall, R.S., Hallam, S.J., Handley, K.M., Hasan, N.A., Haydon, S.R., Hickman, J.E., Hidalgo, G., Hofmockel, K.S., Hooker, J., Hulth, S., Hultman, J., Hyde, E., Ibáñez-Álamo, J.D., Jastrow, J.D., Jex, A.R., Johnson, L.S., Johnston, E.R., Joseph, S., Jurburg, S.D., Jurelevicius, D., Karlsson, A., Karlsson, R., Kauppinen, S., Kellogg, C.T.E., Kennedy, S.J., Kerkhof, L.J., King, G.M., Kling, G.W., Koehler, A.V., Krezalek, M., Kueneman, J., Lamendella, R., Landon, E.M., Lane-deGraaf, K., LaRoche, J., Larsen, P., Laverock, B., Lax, S., Lentino, M., Levin, I.I., Liancourt, P., Liang, W., Linz, A.M., Lipson, D.A., Liu, Y., Lladser, M.E., Lozada, M., Spirito, C.M., MacCormack, W.P., MacRae-Crerar, A., Magris, M., Martín-Platero, A.M., Martín-Vivaldi, M., Martínez, L.M., Martínez-Bueno, M., Marzinelli, E.M., Mason, O.U., Mayer, G.D., McDevitt-Irwin, J.M., McDonald, J.E., McGuire, K.L., McMahon, K.D., McMinds, R., Medina, M., Mendelson, J.R., Metcalf, J.L., Meyer, F., Michelangeli, F., Miller, K., Mills, D.A., Minich, J., Mocali, S., Moitinho-Silva, L., Moore, A., Morgan-Kiss, R.M., Munroe, P., Myrold, D., Neufeld, J.D., Ni, Y., Nicol, G.W., Nielsen, S., Nissimov, J.I., Niu, K., Nolan, M.J., Noyce, K., O'Brien, S.L., Okamoto, N., Orlando, L., Castellano, Y.O., Osuolale, O., Oswald, W., Parnell, J., Peralta-Sánchez, J.M., Petraitis, P., Pfister, C., Pilon-Smits, E., Piombino, P., Pointing, S.B., Pollock, F.J., Potter, C., Prithiviraj, B., Quince, C., Rani, A., Ranjan, R., Rao, S., Rees, A.P., Richardson, M., Riebesell, U., Robinson, C., Rockne, K.J., Rodriguezl, S.M., Rohwer, F., Roundstone, W., Safran, R.J., Sangwan, N., Sanz, V., Schrenk, M., Schrenzel, M.D., Scott, N.M., Seger, R.L., Seguin-Orlando, A., Seldin, L., Seyler, L.M., Shakhsheer, B., Sheets, G.M., Shen, C., Shi, Y., Shin, H., Shogan, B.D., Shutler, D., Siegel, J., Simmons, S., Sjöling, S., Smith, D.P., Soler, J.J., Sperling, M., Steinberg, P.D., Stephens, B., Stevens, M.A., Taghavi, S., Tai, V., Tait, K., Tan, C.L., Tas¸, N., Taylor, D.L., Thomas, T., Timling, I., Turner, B.L., Urich, T., Ursell, L.K., Lelie, D. van der, Treuren, W.V., Zwieten, L. van, Vargas-Robles, D., Thurber, R.V., Vitaglione, P., Walker, D.A., Walters, W.A., Wang, S., Wang, T., Weaver, T., Webster, N.S., Wehrle, B., Weisenhorn,

P., Weiss, S., Werner, J.J., West, K., Whitehead, A., Whitehead, S.R., Whittingham, L.A., Willerslev, E., Williams, A.E., Wood, S.A., Woodhams, D.C., Yang, Y., Zaneveld, J., Zarraonaindia, I., Zhang, Q., Zhao, H., 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551, 457–463. https://doi.org/10.1038/nature24621

Tomczak, K., Czerwińska, P., Wiznerowicz, M., 2015. Review
The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol. Onkol. 2015, 68–77. https://doi.org/10.5114/wo.2014.47136

Vardi, A., Thamatrakoln, K., Bidle, K.D., Falkowski, P.G., 2008. Diatom genomes come of age. Genome Biol. 9, 245. https://doi.org/10.1186/gb-2008-9-12-245

Vargas, C. de, Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Bescot, N.L., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horák, A., Jaillon, O., Lima-Mendez, G., Lukeš, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Coordinators, T.O., Acinas, S.G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M.E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., Karsenti, E., 2015. Eukaryotic plankton diversity in the sunlit ocean. Science 348, 1261605. https://doi.org/10.1126/science.1261605

Vincent, F.J., Colin, S., Romac, S., Scalco, E., Bittner, L., Garcia, Y., Lopes, R.M., Dolan, J.R., Zingone, A., Vargas, C., Bowler, C., 2018. The epibiotic life of the cosmopolitan diatom Fragilariopsis doliolus on heterotrophic ciliates in the open ocean. ISME J. 1. https://doi.org/10.1038/s41396-017-0029-1

Waller, R.F., Cleves, P.A., Rubio-Brotons, M., Woods, A., Bender, S.J., Edgcomb, V., Gann, E.R., Jones, A.C., Teytelman, L., Dassow, P. von, Wilhelm, S.W., Collier, J.L., 2018. Strength in numbers: collaborative science for new experimental model systems. bioRxiv 308304. https://doi.org/10.1101/308304

Wang, C., Zhang, T., Wang, Y., Katz, L.A., Gao, F., Song, W., 2017. Disentangling sources of variation in SSU rDNA sequences from single cell analyses of ciliates: impact of copy number variation and experimental error. Proc. R. Soc. B Biol. Sci. 284, 20170425. https://doi.org/10.1098/rspb.2017.0425

Weiss, S., Treuren, W.V., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L.C., Xu, Z.Z., Ursell, L., Alm, E.J., Birmingham, A., Cram, J.A., Fuhrman, J.A., Raes, J., Sun, F., Zhou, J., Knight, R., 2016. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME J. 10, 1669–1681. https://doi.org/10.1038/ismej.2015.235

Whittaker, R.H., 1969. New concepts of kingdoms or organisms. Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. Science 163, 150–160.

Wisecaver, J.H., Hackett, J.D., 2011. Dinoflagellate Genome Evolution. Annu. Rev. Microbiol. 65, 369–387. https://doi.org/10.1146/annurev-micro-090110-102841

Worden, A.Z., Allen, A.E., 2010. The voyage of the microbial eukaryote. Curr. Opin. Microbiol. 13, 652–660. https://doi.org/10.1016/j.mib.2010.08.001

Worden, A.Z., Follows, M.J., Giovannoni, S.J., Wilken, S., Zimmerman, A.E., Keeling, P.J., 2015. Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. Science 347, 1257594. https://doi.org/10.1126/science.1257594

Worden, A.Z., Lee, J.-H., Mock, T., Rouzé, P., Simmons, M.P., Aerts, A.L., Allen, A.E., Cuvelier, M.L., Derelle, E., Everett, M.V., Foulon, E., Grimwood, J., Gundlach, H., Henrissat, B., Napoli, C., McDonald, S.M., Parker, M.S., Rombauts, S., Salamov, A., Von Dassow, P., Badger, J.H., Coutinho, P.M., Demir, E., Dubchak, I., Gentemann, C., Eikrem, W., Gready, J.E., John, U., Lanier, W., Lindquist, E.A., Lucas, S., Mayer, K.F.X., Moreau, H., Not, F., Otillar, R., Panaud, O., Pangilinan, J., Paulsen, I., Piegu, B., Poliakov, A., Robbens, S., Schmutz, J., Toulza, E., Wyss, T., Zelensky, A., Zhou, K., Armbrust, E.V., Bhattacharya, D., Goodenough, U.W., Van de Peer, Y., Grigoriev, I.V., 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. Science 324, 268–272. https://doi.org/10.1126/science.1167222

Wyman, S.K., Avila-Herrera, A., Nayfach, S., Pollard, K.S., 2018. A most wanted list of conserved microbial protein families with no known domains. PLOS ONE 13, e0205749. https://doi.org/10.1371/journal.pone.0205749

Zhu, F., Massana, R., Not, F., Marie, D., Vaulot, D., 2005. Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. FEMS Microbiol. Ecol. 52, 79–92. https://doi.org/10.1016/j.femsec.2004.10.006

Acknowledgements

Nothing would have been possible without many mentors …
my mum (biology teacher in high school), many professors in evolution and ecology during my master (Professor Hervé Le Guyader, Professor Dominique Higuet, Dr Nathalie Simon, …), my post-docs supervisors (notably Dr Chris Bowler !), and all scientists who are dreamers and dream makers (such as Dr Eric Karsenti).

Science is team work.
My deepest thanks go to my nearest collaborators Dr Fabrice Not, the dream team (Dr Lionel Guidi, Dr Samuel Chaffron and Dr Damien Eveillard), and of course to Dr Sakina-Dorothée Ayata. Among others, I wish to thank especially: Dr Stéphane Le Crom, Dr Chris Bowler (and his entire team), Dr Erwan Corre, Dr Eric Pelletier, Dr Pierre Peterlongo and Dr Camille Marchet, our interractions were and will be always a pleasure.
Without the energy, the motivation and the desire to learn from the lab students, much less research roads would have been explored … Dr. Arnaud Meng, Anne-Sophie Benoiston, Emile Faure, Ophélie Da Silva, and also Jade Leconte, Anita Annamale, Quentin Letourneur and Marie Sorret, thanks a million !

Finally, I would like to thank my friends and most of all my family … my daily, constant support and source of inspiration.