Marine Microbial Biodiversity, Bioinformatics & Biotechnology

Grant agreement n°287589

Acronym: Micro B3

Start date of project: 01/01/2012, funded for 48 month

Deliverable 6.6

Genomes from single cells

Version: 1.1

Circulated to: Chris Bowler WP6 leader; 13/07/2015

Approved by: Prof. Frank Oliver Glöckner, 20/07/2015

Expected Submission Date: 30/06/2015

Actual Submission Date: 20/07/2015

Lead Party for Deliverable: CNRS, Daniel Vaulot
Mail: vaulot@sb-roscoff.fr **Tel.:+33 6 75 54 20 41**

Public (PU)	X
Restricted to other programme participants (including the Commission Services) (PP)	
Restricted to a group specified by the consortium (including the Commission Services) (RE)	
Confidential, only for members of the consortium (including the Commission Services) (CO)	

The Micro B3 project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 287589 (Joint Call OCEAN.2011-2: Marine microbial diversity – new insights into marine ecosystems functioning and its biotechnological potential).
The Micro B3 project is solely responsible for this publication. It does not represent the opinion of the EU. The EU is not responsible for any use that might be made of data appearing herein.

Summary

Many photosynthetic eukaryotes have escaped all attempts to bring them in culture. It is in particular the case for groups that are abundant within the smallest size classes (picoplankton) such as prasinophytes, prymnesiophytes or chrysophytes. Many uncultured clades have been discovered in these groups through the analysis of specific genes such as nuclear 18S or plastid 16S rRNA from environmental samples. It is often difficult to deduce the physiology and ecological role of these uncultured taxa from their placement in the phylogenetic tree. One strategy to obtain genomic data from specific populations used a combination of flow cytometry cell sorting, whole genome amplification, and pyrosequencing. By doing that, we obtained metagenomic data for six natural pico-phytoplankton populations from the South-Eastern Pacific which is a very contrasted oceanic region harboring some of the most oligotrophic (gyre) as well as some of the most productive (upwelling) waters. For two upwelling samples about 60% of the reads could be mapped to the genome of a Mediterranean *Bathycoccus* strain. The DNA identity between the metagenomes and the cultured genome was very high (96.3%). From these samples we also obtained evidences about novel DNA-RNA viruses potentially associated to prasinophytes. At least two to three different genotypes seemed to be present in each natural sample. For the four other samples originating from more oceanic stations, reads could not be mapped to any reference genome and analysis was therefore much more difficult. The recent availability of transcriptomes from a large number of protists species (MMETSP project) allowed however a more precise annotation. Moreover, these data allowed by serendipity to support hypotheses concerning the hosts of novel DNA-RNA hybrid viruses and about a unique symbiosis between a prymnesiophyte and a cyanobacterium.

Table of Content

Introduction

Small photosynthetic eukaryotes (< 3 µm in size) play an important role in marine ecosystems. In many oceanic regions, they may account for a large fraction of the biomass as well as of the primary production. Until now, only a small number of these organisms have been brought into culture, in particular members of the Chlorophyta (prasinophytes and Mamiellophyceae). Most of our knowledge about key groups and about diversity within these groups is based upon the genetic analysis of the 18S rRNA gene in marine samples. In many cases however, sequences obtained using classical approaches (universal primers applied to filtered samples) are dominated by heterotrophic eukaryote groups such as marine alveolates or stramenopiles. The cloning/sequencing of filtered samples of plastid genes such as 16S rRNA and the cloning/sequencing of the 18S rRNA gene of chlorophyll containing populations sorted by flow cytometry have converged towards establishing that a few groups appear to dominate open-ocean, meso- and oligo-trophic waters: Prymnesiophyceae, Chrysophyceae, Pelagophyceae, and two clades of prasinophytes (VII and IX) (Fuller et al. 2006, Shi et al. 2009, 2011, Lepère et al. 2011). In these waters, many of the dominant clades within these groups have very few or no cultured representatives, and no information is available on their morphology and physiology. In contrast, in temperate coastal as well as arctic pelagic waters, the small photosynthetic eukaryote community is dominated by Mamiellophyceae, a group of small green algae for which the three major genera *Bathycoccus*, *Micromonas*, and *Ostreococcus* are easily isolated in culture. These three genera have also been found to bloom sporadically in open ocean waters. The availability of cultures has allowed sequencing of their genomes (Derelle et al. 2006, Worden et al. 2009), providing key information on the genetic basis of their niche differentiation and fostering further analysis of metabolic pathways, mechanisms of genome evolution and life cycle of these organisms. Obtaining genomic information from uncultivated populations would allow exploring their physiological adaptation and could provide clues for their isolation.

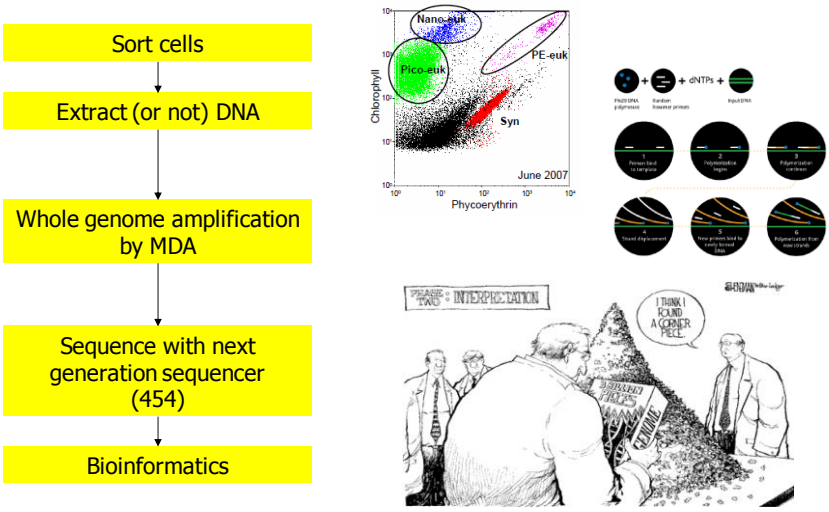
In the last decade, metagenomics approaches have been developed for marine microbial communities. These approaches revealed the existence and ubiquity of processes such as photo-heterotrophy among bacteria. However, until very recently metagenomics have not been applied to eukaryotic populations. The first reason is that filtered samples that are used in general for metagenomic studies are dominated by prokaryotic sequences. Second, eukaryote genomes can be very large with many repeated genes, such that metagenomic data carry little information. One approach, proposed a few years ago for prokaryotes, has been gaining popularity lately (Rinke et al. 2013). It consists in the coupling flow cytometry sorting, which permits to obtain specific cellular groups based on their size and pigment content, with Whole Genome Amplification (WGA), which allows obtaining enough material for genome sequencing with next generation sequencing (NGS) technologies such as 454 or Illumina sequencing. This approach allowed, for example, reconstructing the genome of

UCYN-A, a group of small nitrogen fixing cyanobacteria (Zehr et al. 2008, Tripp et al. 2010). Recent work demonstrated that this strategy is applicable to small eukaryotes, both autotrophic and heterotrophic. This strategy allowed for example retrieving genomic information on a group of uncultivated eukaryotes, the picobiliphytes (Yoon et al. 2011).

We have applied this strategy to six natural picoplankton populations sampled in the South Eastern Pacific in order to assess its feasibility. Although strictly speaking we are not dealing with single cell metagenomes, this is the first step, using "few" cells metagenomes, to highlight potential problems but also show the interest of this approach.

Material and methods

Overall strategy. Samples obtained from the South Pacific Ocean during the 2004 BIOSOPE cruise were concentrated by tangential flow filtration and sorted by flow cytometry on board. DNA was extracted, amplified by Multiple Displacement Amplification and sequenced on 454 platform. Reads were then assembled into contigs and analyzed.



Sampling. Sampling was performed in December 2004 during the oceanographic cruise BIOSOPE that sailed a transect through the eastern South Pacific Ocean on board the research vessel L'Atalante. Seawater was collected using Niskin bottles mounted on a CTD frame at 5 stations (Figure 1 and Table 1). Samples were concentrated by tangential flow filtration using a 100 000 MWCO (Regenerated Cellulose- RC ref VF20C4) Vivaflow 200 cassette.

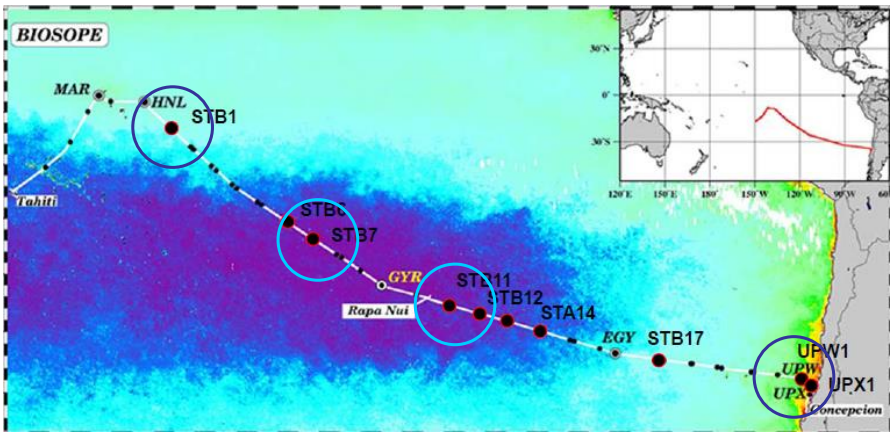


Figure 1: Sampling location for metagenomes from sorted cells collected during the 2004 BIOSOPE cruise

Table 1: List of samples wth conditions and dominant populations.

Station	Samples	Depth	Condition	Dominant picoplankton groups
1	19	Surface	Mesotrophic	Prasinophytes clade VII and IX
7	39	DCM	Oligotrophic	Chrysophyceae/Mamiello/Clade VII
7	41	Surface	Oligotrophic	Prasinophytes clade IX/Chryso
11	60	Surface	Oligotrophic	Chrysophyceae/Clade IX
UPW	142, 149	Mixed layer	Eutrophic	Mamiellophyceae

Flow cytometry analysis and sorting. Concentrated samples were analyzed on board using a FACSAria flow cytometer (Becton Dickinson, San Jose, CA, USA) equipped with a laser emitting at 488 nm and the normal filter setup. The signal was triggered on the red fluorescence from chlorophyll. Photosynthetic picoeukaryotes were discriminated based on side scatter, as well as orange and red fluorescence, and sorted in "purity" mode. Of the order of 100 000 cells per sample were collected into Eppendorf tubes and, after a quick centrifugation, the volume of sorted samples was adjusted to 250 µL by adding filtered seawater. Samples were deep frozen in liquid nitrogen.

DNA extraction and amplification. DNA from the sorted pico-eukaryote population was extracted using DNeasy blood and tissue kit (Qiagen), as recommended by the manufacturer. Multiple displacement amplification (MDA) was performed using the REPLI-g Mini kit (Qiagen) following the manufacturer's protocol with modified buffers as described previously (Lepère et al. 2011). About ten separate reactions were performed and then pooled together for each sample in order to reach the 10 µg of DNA required for 454 sequencing. The amplified products were then purified and concentrated using a Microcon YM- 100 column (Millipore, Molsheim, France).

Sequencing. Sequencing was performed at Genoscope in Paris. About 10 µg of DNA amplified were fragmented by nebulisation. Fragments between 500 and 800 bp were selected and purified by AMPure (Beckman Coulter Genomics). Libraries were prepared following 454 protocol (GS FLX Titanium Library Preparation Kit, Roche Diagnostic, USA).

Libraries were quantified and libraries profiles were evaluated using a 2100 Bioanalyzer (RNA 6000 PicoLabchip kit, Agilent Technologies, USA). Each library was sequenced using 1/2 or one Pico Titer Plate on 454 GSFlx instrument with Titanium chemistry (Roche Diagnostic, USA).

Genome assembly – *de novo*. Assembly was performed using the native Geneious Assembler [http://www.geneious.com/]. First, 454 reads were trimmed at both ends using a probability threshold of p=0.01 and no ambiguity. Trimmed reads were assembled using the default Medium Sensitivity (Allow Gaps = true; Word length = 14; Index word length = 12; Ignore words repeated more than 200 times; Maximum mismatches per reads = 15%; Maximum ambiguity = 4; Maximum gap size = 2).

Genome assembly – *Bathycoccus* genome. Trimmed 454 reads from samples T142 and T149 which were dominated by *Bathycoccus* were assembled against the genome of *Bathycoccus* (downloaded with annotations for CDS [coding sequences], UTR [untranslated regions], introns and exons from the University of Ghent BOGAS web site http://bioinformatics.psb.ugent.be/webtools/bogas/overview/Bathy) using Geneious with the default "Medium Sensitivity".

BLASTN analysis of reads and contigs. We performed a BLASTN search of the raw reads and Geneious contigs against a subset of nr GenBank database. Results from the BLASTN search were analyzed with MEGAN 4.0 [41] with the standard parameters (min score = 35, top-percent = 10%, min support = 5) in order to provide a taxonomic affiliation for each read and contig.

Analysis of Open Reading Frames (ORF). ORF prediction was done on contigs using the RAMMCAP software (Li 2009). The number of ORFs varied from 27,000 to 83,000 depending on the samples. Predicted peptides from the MMETSP project containing transcriptomes of 678 protist culture samples representing a wide taxonomic range (Keeling et al. 2014) were downloaded from iMicrobe (http://data.imicrobe.us/project/view/104) and combined into a single BLAST database. We performed a BLASTP search of predicted metagenome ORFs against this MMETSP database. KEGG and KOG assignation of ORFs was performed on the WGA web site (WebMGA : http://weizhong-lab.ucsd.edu/metagenomic-analysis/). Core gene (CEGMA) analysis (Parra et al. 2007) was performed on the iPlant Discovery Environment (http://www.iplantcollaborative.org/ci/discovery-environment). Principal component analysis (PCA) was performed under R with the FactoMineR library?

Genotypes. For samples T142 and T149 from the coastal upwelling, assembly of small regions of a few selected individual single-copy genes was manually analyzed under Geneious to assess the minimum number of haplotypes present in the two samples. Briefly, reads were ordered based on their similarity/differences to the reference RCC1105. Several reads were considered to correspond to one genotype when they presented similar changes at least two positions.

Coastal metagenomes from sorted picoplankton cells

Our initial effort (Vaulot et al. 2012) targeted two coastal picoplankton samples from the Chile upwelling (samples T142 and T149 corresponding to 5 and 30 m depth, respectively, Figure 1) which were dominated according to 18S rRNA clone libraries with the small Mamiellophyceae *Bathycoccus* (Lepère et al. 2011). Each MDA amplified sample was sequenced on a half run of 454 machine about 670,000 reads, each with an average length ~420 bp. Reads were assembled into contigs using either the Geneious assembler after trimming to remove low quality ends (see Material and Methods).

Bathycoccus metagenomes

In order to determine the taxonomic composition of the assemblage from which the metagenomic sequences originated, we began by searching for SSU rRNA gene within contigs. Although the number of contigs harbouring SSU rRNA genes was slightly higher in sample T149 than T142, their phylogenetic composition was very similar. In both samples, we found nuclear 18S rRNA gene signature for *Bathycoccus* (99.9% identity) but not for other Mamiellophyceae. Although eukaryotes were targeted during flow cytometry sorting, we also found several 16S rRNA genes from bacteria in both samples. All bacterial signatures were affiliated to clades that are typical of the marine environment, in particular the alpha-proteobacteria *Candidatus* Pelagibacter and *Roseobacter*, as well as several Sargasso Sea clades (SAR).

In a second step, we subjected both reads to a global BLASTN search against GenBank and analyzed the results with MEGAN (Huson et al. 2007), which provides, when possible, a taxonomic affiliation for each sequence using the Last Common Ancestor (LCA) algorithm. This search was performed twice first in 2010 and the later in 2013. In 2010 prior to the sequencing of the *Bathycoccus* genome (Moreau et al. 2012) a very small fraction of the reads (6% of total) were affiliated to Mamiellophyceae. However, when the same search was performed in 2013 after the genome of Bathycoccus had been deposited to GenBank more than 63% of the reads presented a hit to *Bathycoccus* (Figure 2). This result illustrates the importance of reference genomes to interpret metagenomics data. Despite the phylogenetic closeness between *Bathycoccus* and other Mamiellophyceae such as *Ostreococcus* for which genomes were available in 2010, no hit was registered for these species.

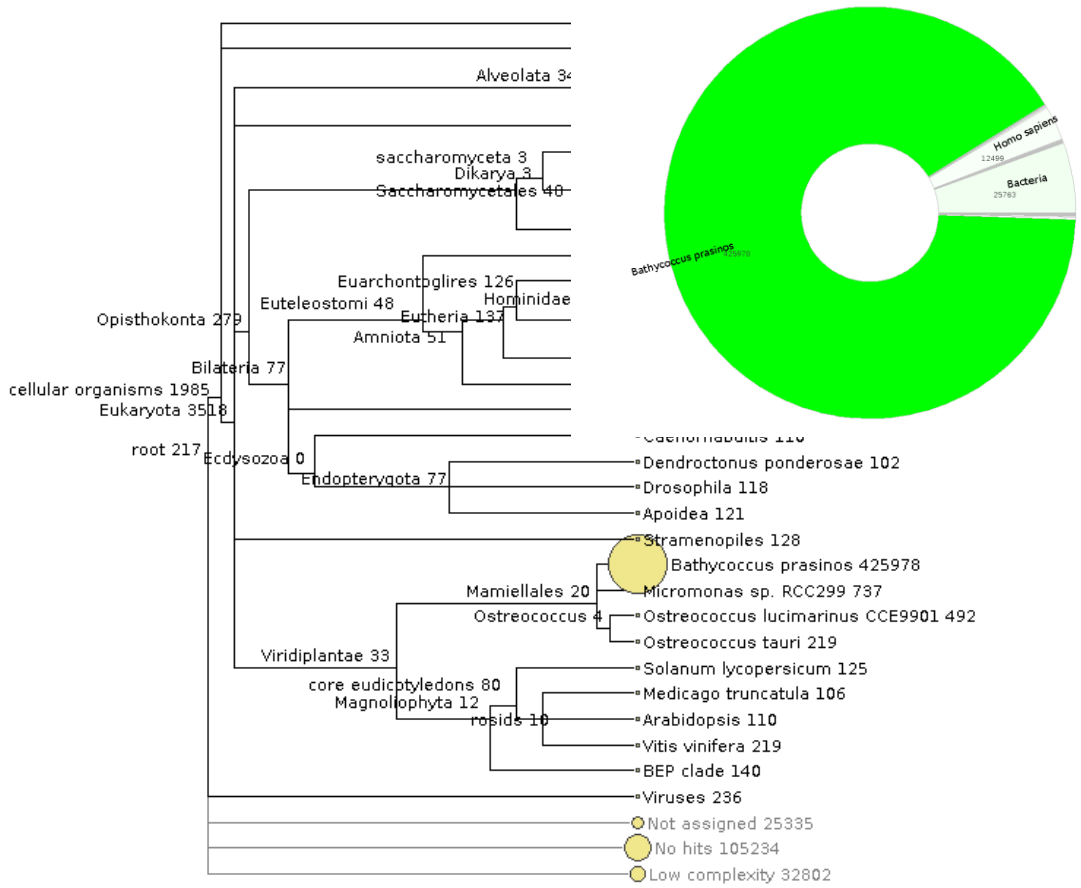


Figure 2: BLASTN analysis of reads of samples T142 against GenBank NR performed in 2013 and visualized by MEGAN

Since BLAST analysis revealed that a large fraction of the reads could be recruited to *Bathycoccus* RCC1105 genome, a direct assembly of reads was performed against the sequenced *Bathycoccus* genome providing detailed information on the genome coverage (

Figure **3**). On average, while coverage depth (i.e., the average number of reads covering a given base from the reference chromosome) was around 10x and similar between both samples, the coverage fraction (i.e., the fraction of the reference genome covered by at least one read) was higher in T149 (88%) than in T142 (62%). Both coverage depth and coverage fraction varied widely among chromosomes, the two indices being somehow related since chromosomes with a higher coverage fraction had also generally a larger coverage depth. The two *Bathycoccus* outlier chromosomes 14 and 19 and the chloroplast genome that have lower GC content had low coverage fraction and low coverage depth. When looking at the detail of coverage depth along each chromosome coverage appeared very unequal and varied between the two samples. Some regions were well-covered in both samples but most other high coverage regions were present in only one of the samples (

Figure **3**).

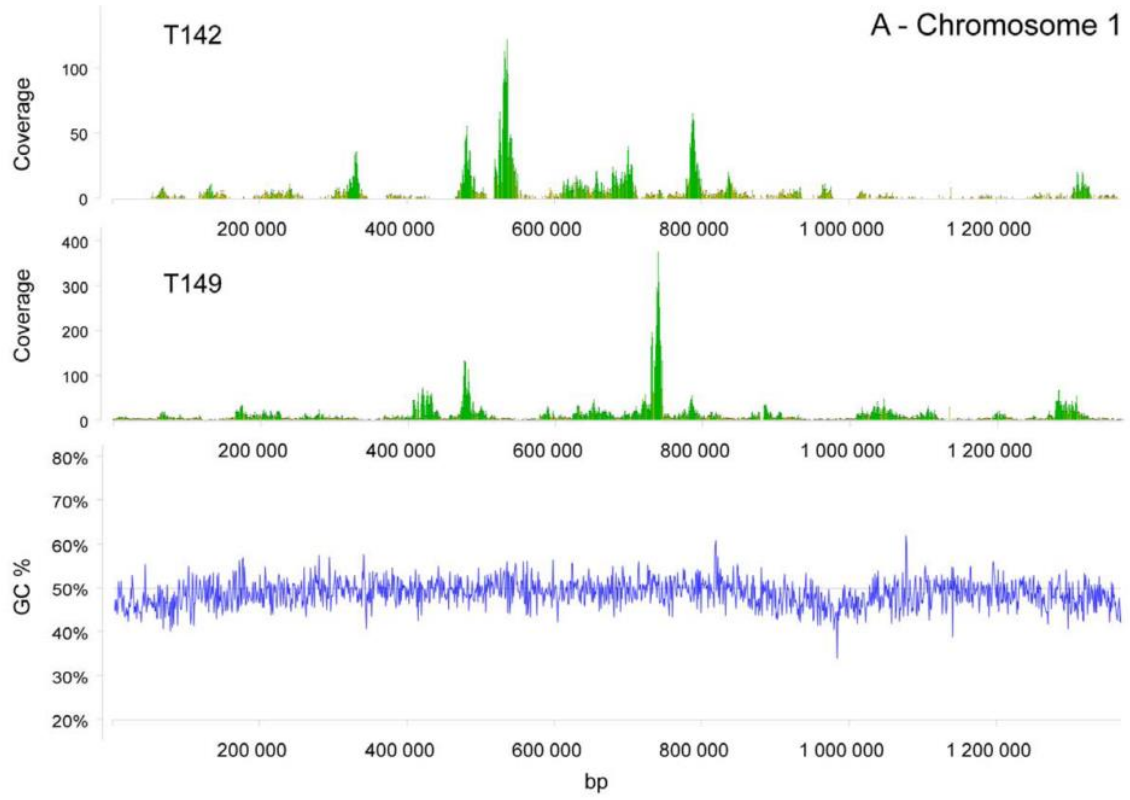


Figure 3: Coverage and GC% for the first Chromosome of *Bathycoccus* for samples T142 and T149 (reprinted from Vaulot et al. 2012)

We analyzed the degree of similarity between the *Bathycoccus* metagenomes and the genome of *Bathycoccus* by aligning the RCC1105 genome sequence and the T142 and T149 consensus sequences for the three largest chromosomes. We removed all regions that had a coverage depth below 10x for both metagenomes and obtained three alignments varying between 78 and 153 kbp. The percentage of identical nucleotides over all positions of three genomes varied between 95.8% and 96.7%. However for non-coding regions, this identity was lower between 87.2 and 89.5%.

Detailed analysis of the reads mapped to specific regions of *Bathycoccus* single-copy genes that have a good coverage allowed estimating the number of different *Bathycoccus* genotypes present in each sample. For example, for gene Bathy02g01050 (gene code according to BOGAS web site – see Materials and Methods) involved in pigment synthesis, sample T142 presented two major sequences accounting for 85 and 15% of the reads, respectively (Figure 4). In the same region of sample T149, the same two genotypes were present but in different proportion, 59 and 41%, respectively. Examination of a few regions

for genes that had a good coverage revealed that the number of different sequences varied from 1 to 3 per samples. In some cases, the same sequences were present in both samples, while in other cases unique sequences were present in each sample.

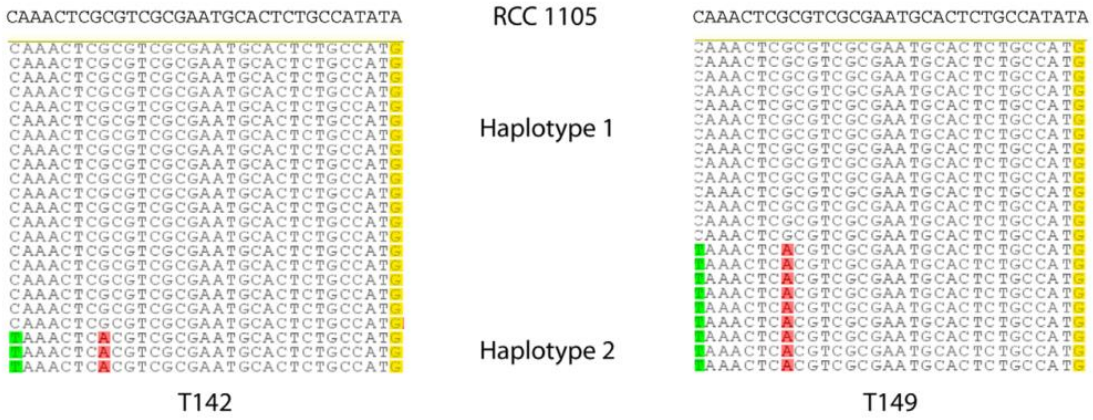


Figure 4: Genotypes observed *B. prasinos* RCC1105 gene Bathy02g01050 (pigment synthesis protein, BOGAS annotation code). At least two different sequences appear to be present in both samples, differing by one and three positions, respectively, from the reference *B. prasinos* RCC 1105 sequence (reprinted from Vaulot et al. 2012)

The *Bathycoccus* Mediterranean culture genome and the two Chile metagenomes that originate from opposite sides of the Earth appear very closely related, with an average nucleotide identity up to 98% at the individual gene level. Conservation between these three genomes is much higher than, for example, between the genomes of *O. tauri* and *O. 'lucimarinus'*, for which average amino-acid identity is only 88% when using a BLASTX-based algorithm to affiliate *O. 'lucimarinus'* sequences (random 1,000 bp sequences) to *O. tauri* genome taken as a reference. In fact, it is possible that all three genomes correspond to the same biological species, since an analysis of the full rRNA operon reveals nearly identical sequences in the ITS2 region (Vaulot et al. 2012) and identical sequences in the ITS2 have been shown to correspond in general to reproductive compatibility. Despite the high level of sequence identity between these three genomes, several genotypes appear to be present within each of the Chile upwelling populations (Figure 4).

Monier et al. (2012) reported the construction of a *Bathycoccus* metagenome from flow cytometry sorted cells collected at the deep chlorophyll maximum (DCM) in the tropical Atlantic. We compared, using a custom BLASTX-based algorithm, the DCM metagenome and the RCC1105 genome. Amino acid identity with the reference *Bathycoccus* RCC1105 genome was much lower (84.4 %) than for our metagenomes (96.0 and 95.9 %, for T142 and T149, respectively), despite the fact that the SSU rRNA sequence of the DCM metagenome is 100% identical to that of RCC1105. This raises the intriguing possibility that the *Bathycoccus* genus could contain different ecotypes, some adapted to coastal waters, the others to pelagic/deep waters, in a manner similar to its sister genera *Ostreococcus* and *Micromonas*.

RNA-DNA viruses

Single-stranded (ss) DNA viruses represent a rapidly expanding, diverse supergroup of important pathogens preying on hosts from all three domains of life. Recently viruses with ssDNA genomes have been repeatedly isolated from diverse environments, including marine ecosystems. Numerous studies on uncultivated viral communities using metagenomic approaches have revealed that genetic diversity of ssDNA viruses is much greater than originally recognized. Recently, a novel chimeric viral (CHIV) ssDNA genome (Diemer and Stedman 2012) was recovered from a hot, acidic Boiling Springs Lake (BSL), USA. It was suggested that the virus has emerged via recombination between a DNA and an RNA virus. Three assemblies of new CHIV genomes (Figure 5) have been recovered from the BIOSOPE upwelling metagenome samples T142 and T149 (Roux et al. 2013). Since these metagenomes were dominated by *Bathycoccus*, this suggests that these hybrid viruses could be hosted by picoplanktonic algae.

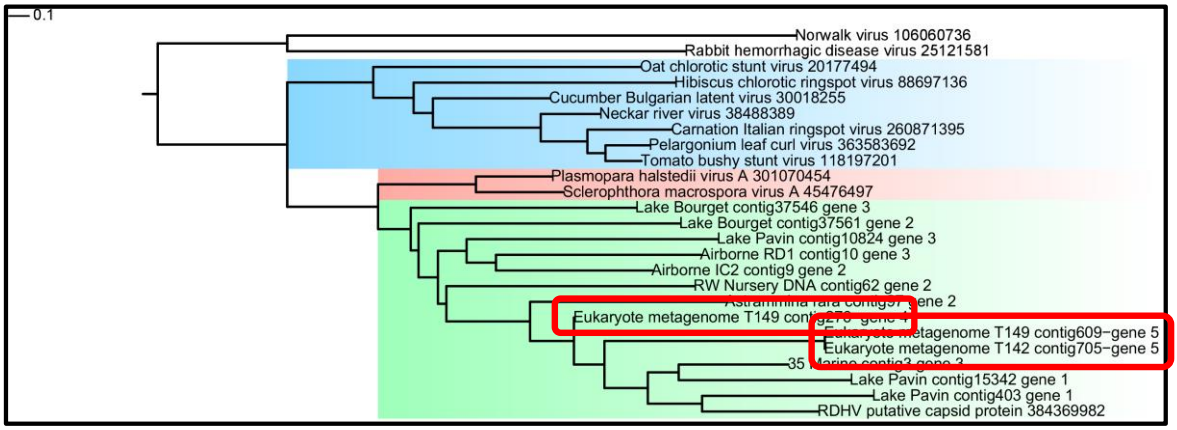


Figure 5: Coat protein phylogeny from hybrid viruses metagenomes including three recovered from samples T142 and T149 (Roux et al. 2013)

Open ocean metagenomes from sorted picoplankton cells

Picoplankton metagenome analysis

In contrast to the coastal samples discussed above, the oceanic photosynthetic picoplankton (Figure 1) contained members of taxonomic groups (such as chrysophytes or prasinophytes, Table 1) for which we do not have reference genomes and even in many cases no cultures available. We first analysed contigs to extract ORFs using the RAMMCAP pipeline. A BLASTP search against proteins from GenBank yielded a very low number of hits especially for groups which we knew dominated the samples. For example for surface samples T19 and T41 that contained mostly prasinophytes from cultured clade VII and uncultured clade IX (Table 1), only 6 and 11% of the hits corresponded to the green lineage (Figure 6 Top). Recently, a large number of transcriptomes from protist, especially photosynthetic ones have been released (MMETSP project, Keeling et al. 2014). When we used these transcriptomes as the target database for BLASTP query, a much larger number of hits was

obtained, in particular for groups related for our samples. For example for samples T19 and T41 more than half of the hits corresponded to Chlorophyta transcriptomes especially from the Chlorodendrophyceae genus *Tetraselmis* (Figure 6 Bottom). As in the case of the coastal samples discussed above this underline the necessity to have good reference database to interpret metagenomics data.

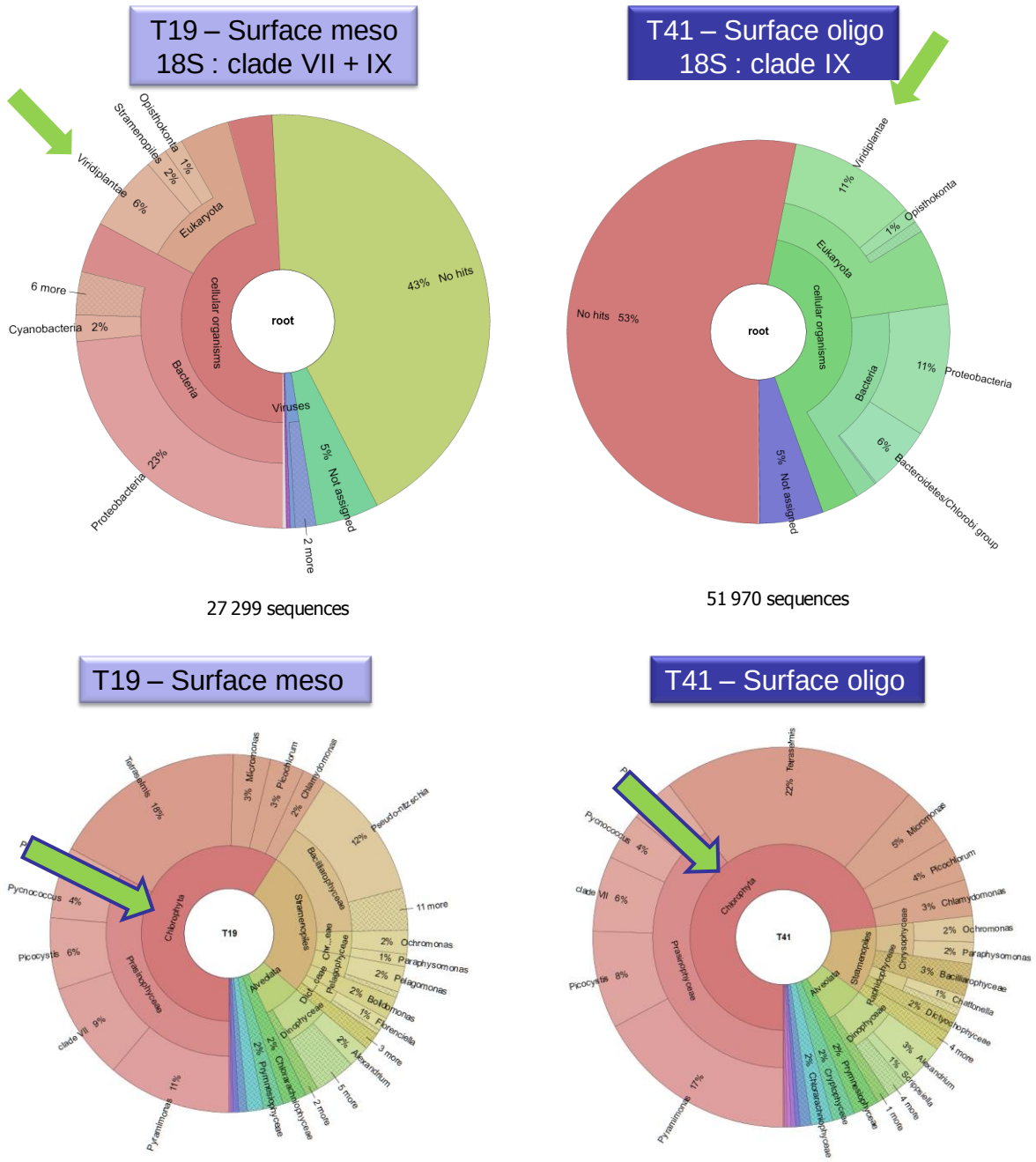


Figure 6: BLASTP analysis of ORFs recovered from contigs for samples T19 and T41 from BIOSOPE. Top: against GenBank. Bottom: against protists transcriptomes (MMETSP)

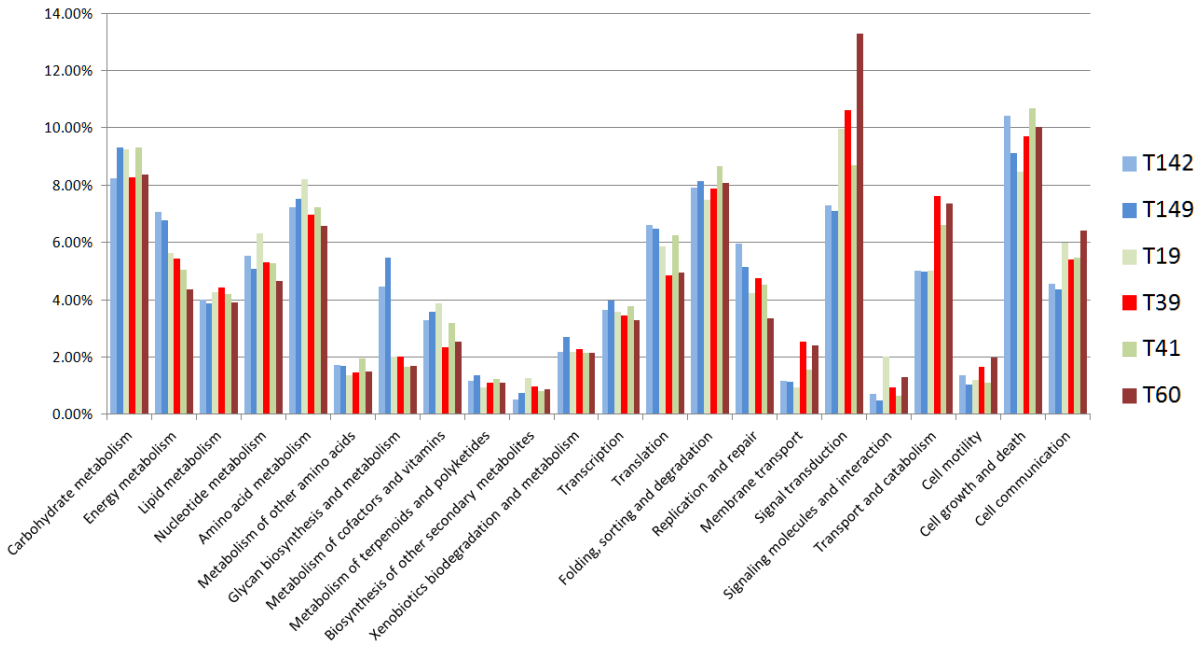


Figure 7: KEGG profiles of ORFs recovered from contigs for BIOSOPE sorted samples

In order to gain more insight into the genomics potential of these natural pico-plankton populations, we determined the KEGG affiliation of the ORFs obtained with RAMMCAP (Figure 7). Interestingly, the profiles obtained were very similar between the 6 samples analyzed with KEGG categories like Signal transduction or Cell growth and Death dominating.

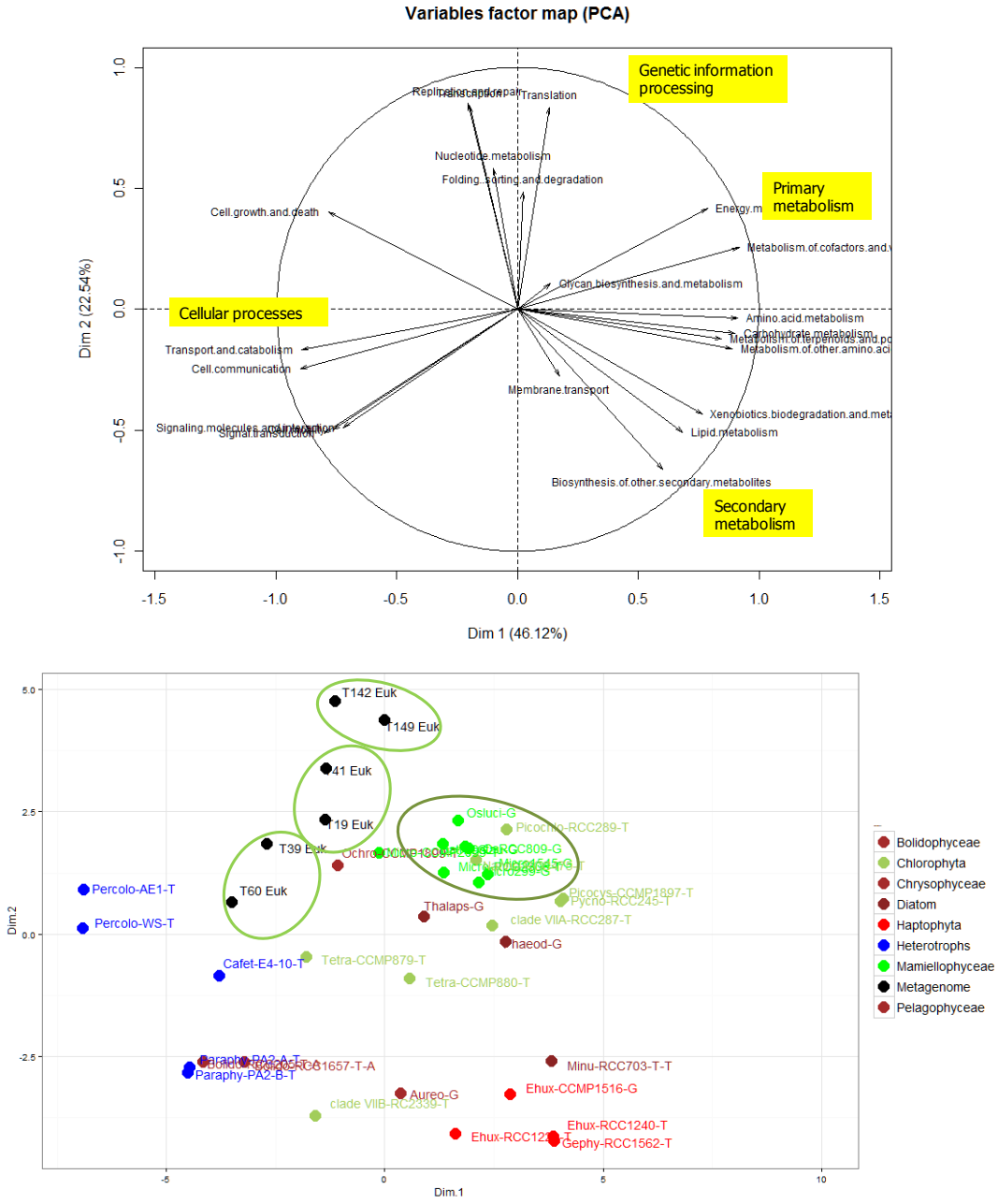


Figure 8: Principal Component Analysis (PCA) of KEGG profiles from BIOSOPE sorted samples using MMETSP protein transcriptomes for reference. Top. Contribution of KEGG categories to the first 2 components. Bottom. Map of the samples and transcriptomes

We compared these KEGG signatures with those of available picoplankton genomes (e.g. *Ostreococcus*) and some of the transcriptomes from the MMETSP project using principal component analysis (Figure 8). The first two components explained almost 70% of the variance with Primary metabolism and Cellular processes categories contributing to the first axis and Genetic Information processing and Secondary metabolism contributing to the second axis. Interestingly, transcriptomes and genomes from the same taxonomic groups

regrouped together (e.g. Mamiellophyceae in green on Figure 8). Transcriptomes from heterotrophic protists such as *Percolomonas* or *Paraphysomonas* (in blue on Figure 8) grouped together discriminated by a higher contribution of KEGG categories such Transport or Signalling. Metagenomes grouped together according to the taxonomic composition of the sorted population. For example T142 and T149, which were dominated by *Bathycoccus*, or T41 and T19 dominated by prasinophytes. Surprisingly, they did not cluster with transcriptomes from these groups.

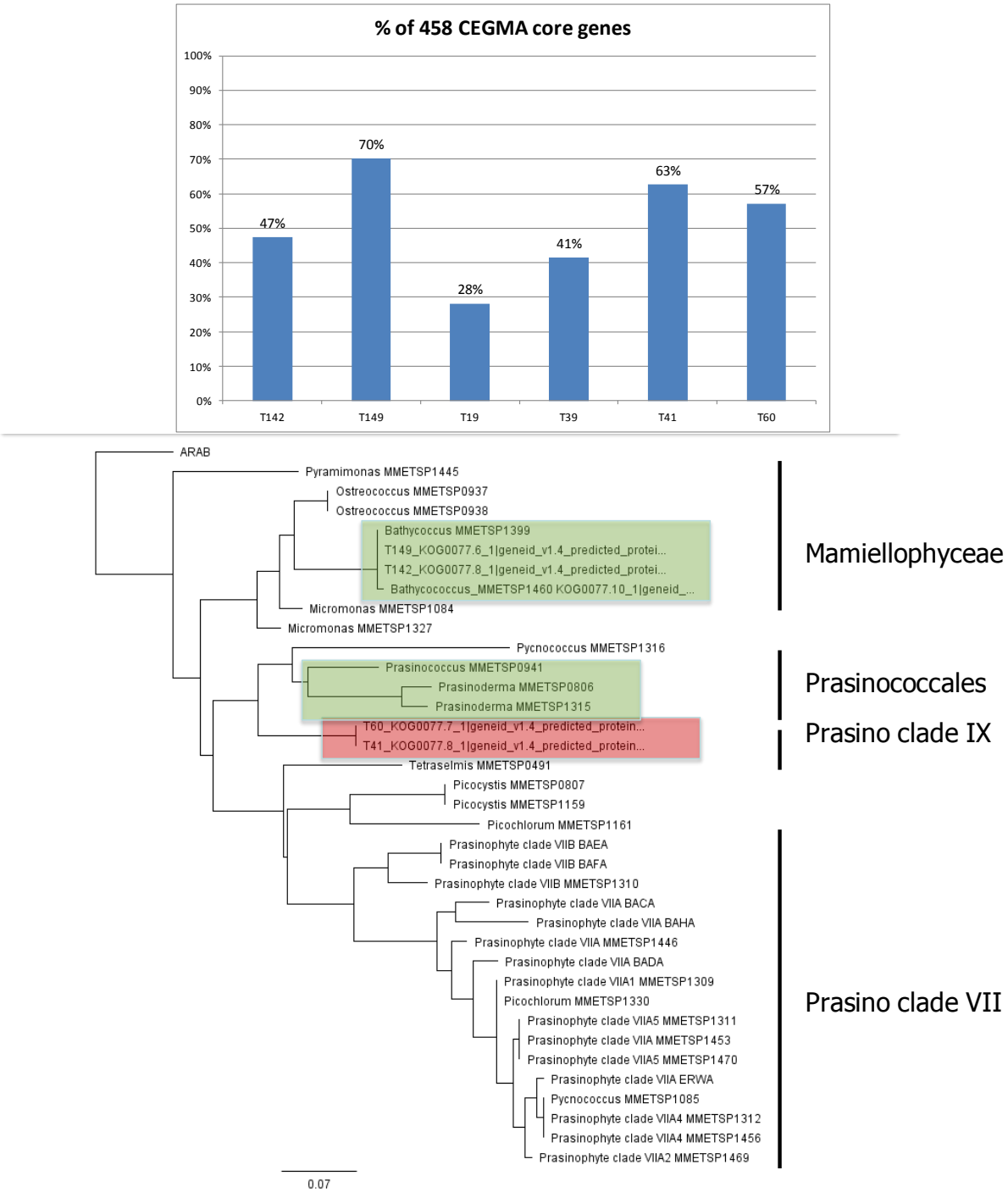


Figure 9: Top. Percentage of core genes recovered from each of the metagenomic samples. Bottom. Phylogeny of vesicle coat complex COPII protein (KOG0077) recovered from samples T60 and T41 by CEGMA analysis

Another approach which was tested on this data set was to extract a subset of conserved core genes using the CEGMA pipeline (Parra et al. 2007). Depending on the sample, from 28 to 70% of the initial set of 458 core genes were recovered (Figure 9 Top). The advantage of this approach is to allow to build phylogenies for groups which are yet uncultured and for which the only phylogenetic information originates from the 18S rRNA gene. For example we recovered from 2 samples (T41 and T60) identical sequences for one protein affiliated to KOG0077(Figure 9 Bottom). This protein probably comes from a prasinophyte from uncultured clade IX, an organism present in both samples (Table 1). Phylogenetic analysis, suggests that clade IX could be closely related to Prasinococcales, a relationship already suggested by 18S rRNA phylogeny.

Symbiosis between nitrogen fixing-cyanobacteria and haptophytes

Symbioses between nitrogen (N)2–fixing prokaryotes and photosynthetic eukaryotes are important for nitrogen acquisition in N-limited environments. A widely distributed planktonic uncultured nitrogen-fixing cyanobacterium (UCYN-A) was found to have unprecedented genome reduction, including the lack of oxygen-evolving photosystem II and the tricarboxylic acid cycle(Zehr et al. 2008, Tripp et al. 2010), which suggested partnership in a symbiosis. Thompson et al. (2012) showed that UCYN-A has a symbiotic association with a unicellular prymnesiophyte, closely related to calcifying taxa present in the fossil record. The partnership is mutualistic, because the prymnesiophyte receives fixed N in exchange for transferring fixed carbon to UCYN-A. This unusual partnership was independently confirmed by the fact that in one of the BISOPE samples (T60) we found the sequence of the host (a prymnesiophyte close to the species *Braarudosphaera bigelowii*) and elements of the genome of the nitrogen-fixing symbiont UCYN-A (Figure 10).

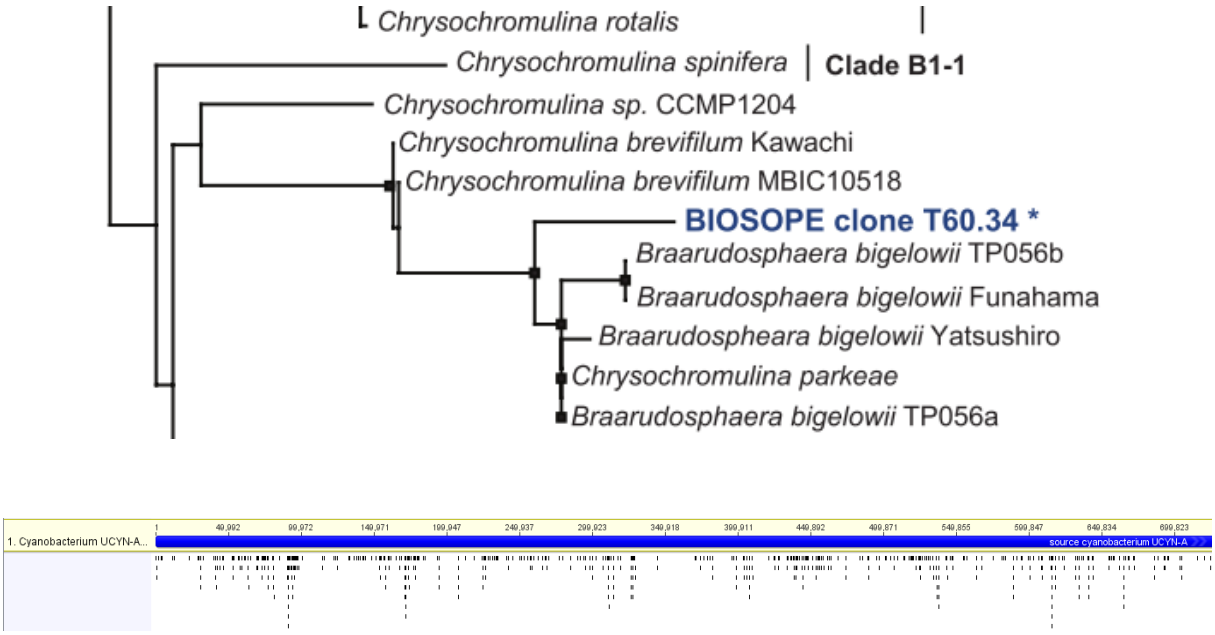


Figure 10: Top. 18S phylogenetic tree of UCYN-A prymnesiophyte host found in sample T60. Bottom. Assembly of metagenome from sample T60 against the Genome of UCYN-A nitrogen fixing-cyanobacterium (Thompson et al. 2012)

Conclusion

This work illustrates the power of coupling flow cytometry sorting to target specific populations followed by NGS to obtain sequence data on photosynthetic microbes. The availability of sequenced genomes or transcriptomes for the dominant organism in these samples is crucial for the data analysis. The application of this approach, provided that higher and more uniform coverage can be obtained, may bring in the future unique information on uncultured small photosynthetic eukaryotes that appear to dominate in the more oligotrophic oceanic region.

Recommendation for future work

- Improve genome amplification (WGA) to reduce differences in coverage depth
- Use novel sequencing approaches (Illumina) for higher coverage
- Develop better assembly and annotation tools for eukaryote metagenomes
- Obtain more reference genomes or transcriptomes from photosynthetic protists

Published papers related to Deliverable 6.6

- Roux, S., Enault, F., Bronner, G., Vaulot, D., Forterre, P. & Krupovic, M. 2013. Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. *Nat. Commun.* 4.
- Thompson, A.W., Foster, R.A., Krupke, A., Carter, B.J., Musat, N., Vaulot, D., Kuypers, M.M.M. et al. 2012. Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Science*. 337:1546–50.
- Vaulot, D., Lepère, C., Toulza, E., de la Iglesia, R., Poulain, J., Gaboyer, F., Moreau, H. et al. 2012. Metagenomes of the Picoalga *Bathycoccus* from the Chile coastal upwelling. *PLoS One*. 7:e39648.

These publications are available at http://microb3.eu/media-material/publications

Cited references

Derelle, E., Ferraz, C., Rombauts, S., Rouze, P., Worden, A.Z., Robbens, S., Partensky, F. et al. 2006. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl. Acad. Sci. U. S. A.* 103:11647–52.

Diemer, G.S. & Stedman, K.M. 2012. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses.

Fuller, N.J., Campbell, C., Allen, D.J., Pitt, F.D., Le Gall, F., Vaulot, D. & Scanlan, D.J. 2006. Analysis of photosynthetic picoeukaryote diversity at open ocean sites in the Arabian Sea using a PCR biased towards marine algal plastids. *Aquat. Microb. Ecol.* 43:79–93.

Huson, D.H., Auch, A.F., Qi, J. & Schuster, S.C. 2007. MEGAN analysis of metagenomic data. *Genome Res.* 17:377–86.

Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V. et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol*. 12:e1001889.

Lepère, C., Demura, M., Kawachi, M., Romac, S., Probert, I. & Vaulot, D. 2011. Whole Genome Amplification (WGA) of marine photosynthetic eukaryote populations. *FEMS Microbiol. Ecol.* 76:516–23.

Li, W. 2009. Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics*. 10:359.

Monier, A., Welsh, R.M., Gentemann, C., Weinstock, G., Sodergren, E., Armbrust, E.V., Eisen, J.A. et al. 2012. Phosphate transporters in marine phytoplankton and their viruses: cross-domain commonalities in viral-host gene exchanges. *Environ. Microbiol.* 14:162–76.

Moreau, H., Verhelst, B., Couloux, A., Derelle, E., Rombauts, S., Grimsley, N., Van Bel, M. et al. 2012. Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular specializations at the base of the green lineage. *Genome Biol.* 13:R74.

Parra, G., Bradnam, K. & Korf, I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*. 23:1061–7.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A.E. et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*. 499:431–7.

Roux, S., Enault, F., Bronner, G., Vaulot, D., Forterre, P. & Krupovic, M. 2013. Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. *Nat. Commun.* 4.

Shi, X.L., Lepère, C., Scanlan, D.J. & Vaulot, D. 2011. Plastid 16S rRNA gene diversity among eukaryotic picophytoplankton sorted by flow cytometry from the South Pacific Ocean. *PLoS One*. 6:e18979.

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:e7657.

Thompson, A.W., Foster, R.A., Krupke, A., Carter, B.J., Musat, N., Vaulot, D., Kuypers, M.M.M. et al. 2012. Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Science*. 337:1546–50.

Tripp, H.J., Bench, S.R., Turk, K.A., Foster, R.A., Desany, B.A., Niazi, F., Affourtit, J.P. et al. 2010. Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature*. 464:90–4.

Vaulot, D., Lepère, C., Toulza, E., de la Iglesia, R., Poulain, J., Gaboyer, F., Moreau, H. et al. 2012. Metagenomes of the Picoalga *Bathycoccus* from the Chile coastal upwelling. *PLoS One*. 7:e39648.

Worden, A.Z., Lee, J.-H., Mock, T., Rouzé, P., Simmons, M.P., Aerts, A.L., Allen, A.E. et al. 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science*. 324:268–72.

Yoon, H.S., Price, D.C., Stepanauskas, R., Rajah, V.D., Sieracki, M.E., Wilson, W.H., Yang, E.C. et al. 2011. Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science*. 332:714–7.

Zehr, J.P., Bench, S.R., Carter, B.J., Hewson, I., Niazi, F., Shi, T., Tripp, H.J. et al. 2008. Globally distributed uncultivated oceanic N2-fixing cyanobacteria lack oxygenic photosystem II. *Science*. 322:1110–2.