Development of a targeted metagenomic approach to study a genomic region involved in light harvesting in marine Synechococcus

Florian Humily, Gregory K. Farrant, Dominique Marie, Frédéric Partensky, Sophie Mazard, Morgan Perennou, Karine Labadie, Jean-Marc Aury, Patrick Wincker, Audrey Nicolas Segui, et al.

Development of a targeted metagenomic approach to study a genomic region involved in light harvesting in marine *Synechococcus*

Florian Humily[1,2], Gregory K. Farrant[1,2], Dominique Marie[1,2], Frédéric Partensky[1,2], Sophie Mazard[3,4], Morgan Perennou[5], Karine Labadie[6], Jean-Marc Aury[6], Patrick Wincker[6], Audrey Nicolas Segui[1,2], David J. Scanlan[3] and Laurence Garczarek[1,2*]

[1]UPMC-Université Paris VI, Station Biologique, CS 90074, 29688 Roscoff cedex, France

[2]CNRS, UMR 7144 Adaptation and Diversity in the Marine Environment, Oceanic Plankton group, Marine Phototrophic Prokaryotes team, Place Georges Teissier, CS 90074, 29688 Roscoff cedex, France

[3]University of Warwick, School of Life Sciences, Gibbet Hill Road, Coventry CV4 7AL, UK.

[4]Macquarie University, Department of Chemistry and BioMolecular Science, North Ryde, NSW 2109, Australia

[5]CNRS, FR2424, Service Information et Génomique, Station Biologique, 29688 Roscoff cedex, France

[6]Commissariat à l'Energie Atomique, Institut de Génomique/Genoscope, 91000 Evry, France

[*]**Correspondence**. E-mail laurence.garczarek@sb-roscoff.fr; Tel. (+33) 2 98 29 25 38; Fax (+33) 2 9829 2324.

Keywords

marine cyanobacteria; phycobilisomes; whole genome amplification; flow cytometry cell sorting; fosmid library

Revised for FEMS Microbiology Ecology on 14/11/2013

Abstract

Synechococcus, one of the most abundant cyanobacteria in marine ecosystems, displays a broad pigment diversity. However, the *in situ* distribution of pigment types remains largely unknown. In this study, we combined flow cytometry cell sorting, whole-genome amplification and fosmid library construction to target a genomic region involved in light-harvesting complex (phycobilisome) biosynthesis and regulation. *Synechococcus* community composition and relative contamination by heterotrophic bacteria were assessed at each step of the pipeline using terminal restriction fragment length polymorphism targeting the *petB* and 16S rRNA genes, respectively. This approach allowed us to control biases inherent to each method and select reliable WGA products to construct a fosmid library from a natural sample collected off Roscoff (France). Sequencing of 25 fosmids containing the targeted region led to the assembly of whole or partial phycobilisome regions. Most contigs were assigned to clades I and IV consistent with the known dominance of these clades in temperate coastal waters. However, one of the fosmids contained genes distantly related to their orthologs in reference genomes, suggesting that it belonged to a novel phylogenetic clade. Altogether, this study provides novel insights about *Synechococcus* community structure and pigment type diversity at a representative coastal station of the English Channel.

One sentence summary

In this study, we developed a method combining flow cytometry cell sorting, whole-genome amplification and fosmid library construction to target a genomic region involved in light-harvesting complex biosynthesis and regulation.

Introduction

During the last decade, the development of culture-independent approaches has allowed scientists to make a leap forward in microbial ecology by showing that the genetic diversity of natural populations is strongly under-represented in culture collections (Rusch *et al.*, 2007, Pedrós-Alió, 2012). Indeed, the advent of high throughput sequencing has granted scientists access to extensive gene content information on the different members of natural communities and even sometimes to entire metabolic pathways or whole genomes (Tringe & Rubin, 2005, Martín *et al.*, 2006, Wooley *et al.*, 2010). This approach proved particularly successful in environments with low complexity (Tyson *et al.*, 2004, Cuadros-Orellana *et al.*, 2007). However, environmental genomics (also called metagenomics) often fails to relate diversity information of specific microbes to their ecological functions in more complex environments, in which this link is often only accessible for the dominant members of the microbial community (Pedrós-Alió, 2012).

To characterize the gene repertoire of a particular group, reducing the sample complexity appears to be an appropriate strategy. Analysis of sub-communities exhibiting lower diversity, such as specific taxa or populations (Kalyuzhnaya *et al.*, 2008) or even single cells (for a review, see Stepanauskas, 2012), is now possible by an approach termed "targeted enrichment" (Hallam *et al.*, 2006). The latter comprises focusing on a particular cell population, based on its specific phenotypic or spectral characteristics. In marine ecology, fluorescence-activated cell sorting by flow cytometry has become one of the most popular isolation methods, by allowing high-throughput cell separation based on various parameters such as light scatter and natural or induced fluorescence (Sekar *et al.*, 2004, Stepanauskas & Sieracki, 2007, Woyke *et al.*, 2009, Yoon *et al.*, 2011). After cell or population enrichment, sequencing efforts can also be directed towards specific genes or metabolic pathways (i.e., targeted metagenomics), selected using different screening methods (for a review, see Suenaga, 2012). However, these targeted methodologies require working with low cell numbers, and thus low DNA amounts, often limiting for current sequencing procedures. This constraint can be circumvented by the use of whole-genome amplification (WGA), enabling acquisition of enough DNA for sequencing from a very low amount of DNA or cells (Abulencia *et al.*, 2006, Podar *et al.*, 2007, Chen *et al.*, 2008,

Lepère *et al.*, 2011) or even single cells (Zhang *et al.*, 2006, Rodrigue *et al.*, 2009). Multiple displacement amplification (MDA) emerged as the reference technique, due to the high processivity and strand displacement activity of the phi29 enzyme (Dean *et al.*, 2001). Although this technique is not devoid of biases, due to chimera formation (Lasken & Stockwell, 2007, Chen *et al.*, 2008) or to differential amplification of some taxa with regard to others (Abulencia *et al.*, 2006), the combination of flow cell sorting and WGA led to significant advances in recent years. This notably includes the study of distribution, genetic diversity and gene content of various uncultured phytoplankton groups (Cuvelier *et al.*, 2010, Lepère *et al.*, 2011, Mazard *et al.*, 2011, Vaulot *et al.*, 2012, Malmstrom *et al.*, 2013) or the acquisition of environmental genomes, such as UCYN-A, starting from only 5,000 flow sorted cells (Tripp *et al.*, 2010).

In this context, *Synechococcus* constitutes a particularly good candidate for targeted metagenomic studies since it is the second most abundant single-celled cyanobacterium in marine ecosystems (Scanlan *et al.*, 2009) and plays a key role in the global carbon cycle (Li, 1994, Richardson & Jackson, 2007, Jardillier *et al.*, 2010, Buitenhuis *et al.*, 2012). This oxygenic photoautotroph, which is found in the upper lit layer across large environmental gradients from coastal waters to the open-ocean, exhibits a high genetic diversity (Zwirglmaier *et al.*, 2008, Huang *et al.*, 2012, Mazard *et al.*, 2012). Four clades (I to IV) are particularly abundant, with clades I and IV prevailing in nutrient-rich coastal waters, clade II in (sub)tropical waters and clade III in oligotrophic regimes (Zwirglmaier *et al.*, 2008, Mella-Flores *et al.*, 2011, Mazard *et al.*, 2012). *Synechococcus* also displays considerable pigment diversity (Six *et al.*, 2007). Like most cyanobacteria, it collects light using phycobilisomes (PBS), i.e., pigment-protein complexes composed of phycobiliproteins. Three main pigment types have been defined based on the major phycobiliprotein found in PBS rods: type 1 contains only phycocyanin (PC), type 2 PC and phycoerythrin I (PEI) and type 3 PC, PEI and PEII (Six *et al.*, 2007). The latter type has been subdivided into 5 sub-types based on the ratio of the two phycobilins, phycourobilin (PUB, A_{max} = 495 nm) or phycoerythrobilin (PEB, A_{max} = 545 nm), linked to these phycobiliproteins (Humily *et al.*, in press) . This PUB to PEB ratio, generally assessed by the fluorescence excitation ratio of these two chromophores, can be low (3a), medium (3b), high (3c) or variable (3d-e). The latter

two sub-types correspond to cells performing type IV chromatic acclimation (CA4) that are able to match their pigmentation with incident light (Palenik, 2001, Everroad & Wood, 2006, Humily *et al.*, in press). CA4 sub-types can be further distinguished, by adding a suffix (A or B), to indicate the occurrence of one of the two CA4 genomic island types, CA4-A or CA4-B, in their genomes. The seven pigment types/sub-types also differ in their content of genes involved in the synthesis and regulation of PBS rods, which are gathered into a specific genomic region ranging in size between 9 and 30 kbp (Six *et al.*, 2007). Although this pigment diversity likely allows these organisms to colonize the large variety of light environments existing in the marine ecosystem, from green turbid coastal to blue oligotrophic waters (Olson *et al.*, 1988, Olson *et al.*, 1990, Partensky *et al.*, 1999), the influence of pigmentation on the niche partitioning of *Synechococcus* spp. remains poorly understood.

In this study, we have developed a metagenomics approach to specifically target PBS genomic regions, the gene content of which can provide valuable information regarding the pigment types present in natural *Synechococcus* populations without prior knowledge of their spectral properties as well as about the evolutionary origin and regulation of phycobilisome biosynthesis genes,. This method, which consists of a unique combination of cell sorting by flow cytometry, WGA and fosmid library construction, is actually applicable to retrieve any large DNA fragments from uncultured *Synechococcus* cells.

Materials and methods

Sample collection and cell sorting

A seawater surface sample (8L) was collected on May 29th, 2012 at 1 m depth at the Service d'Observation en Milieu LITtoral (SOMLIT) Astan long-term station off Roscoff, France (latitude: 48° 46' 40" N, longitude : 3° 56' 15" W), using a Niskin bottle mounted on a CTD frame. Characteristics of the seawater at the sampling date were the following: temperature: 12.55 °C; salinity: 35.24 psu; Chl *a*: 0.60 µg.L^{-1};. Nutrients were also measured according to standard protocols (Koroleff, 1970, Aminot & Kérouel, 2007): NH_4^+: 1.40 µM; NO_3^-:1.10 µM; NO_2^-: 0.17 µM; PO_4^{3-}: 0.22 µM. Within two hours

after being collected the sample was pre-filtered through 3 µm polycarbonate filters (Millipore, Molsheim, France) to remove large planktonic cells and particles, then concentrated (229-fold) by tangential flow filtration (TFF) on a 100,000 MWCO membrane, as previously described (Marie *et al.*, 2010). Picophytoplankton cell concentrations were determined by flow cytometry using a FACS Canto II (Becton Dickinson, San Jose, CA), as detailed in (Marie *et al.*, 1999).

Synechococcus cells were sorted using a FACS Aria flow cytometer (Becton Dickinson, San Jose, CA). In order to prevent sample contamination, the flow cytometer was cleaned by a succession of 2% detergent solution (Hellmanex® II, Hellma GmbH & Co. KG, Germany) in warm water, for at least 2 h followed by overnight rinsing with MilliQ water, as described by Stepanauskas & Sieracki (2007). The Phosphate Buffered Saline (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 2 mM KH_2PO_4, pH 7.5) used as sheath fluid (pressure of 70 psi) was prepared by dissolving combusted (4 h at 450 °C) salts in MilliQ water, UV-treated for 2 h, autoclaved at 121 °C for 20 min and then filtered through a 0.2 µm pore size Stericup® filter (Millipore, Molsheim, France) before transfer into a UV-treated tank. Prior to sorting, a 0.2 µm pore size Sterivex® filter (Millipore) was connected on the sheath fluid line injection, just before the flow chamber.

A three-step sorting procedure was used to sort *Synechococcus* cells to reduce as much as possible the number of co-occurring heterotrophic cells (Fig. S1). Cells were first sorted based on their natural autofluorescence using the “purity” mode and by triggering the signal on both red chlorophyll fluorescence and side scatter. Then, two additional sorts were performed to screen out heterotrophic bacteria and picoeukaryotes after DNA staining for 15 min using a combination of two nucleic acid stains, SYTO-9 and SYTO-13, at a final concentration of 5 µM each (Invitrogen, Carlsbad, CA; del Giorgio *et al.*, 1996, Lebaron *et al.*, 1998, Gasol *et al.*, 1999). Between each sort, the sample injection and flow chamber lines were cleaned for 10 min using an alkaline detergent (BioRad, Hercules, CA) and rinsed for at least 10 min with autoclaved MilliQ water. Sorted cells were collected in UV-treated Eppendorf tubes, quickly spun down and kept on ice until WGA or DNA extraction.

Whole genome amplification

WGA reactions were carried out under a HEPA/UV3 PCR Workstation (UVP, Cambridge, UK) using the Genomiphi v2 kit (GE Healthcare, Waukesha, WI). Lysis (400 mM KOH, 100 mM DTT, 10 mM EDTA) and neutralization buffers (400 mM HCl, 600 mM Tris-HCl, pH 7.5) were filtered through a 0.2 µm pore size MiniSart syringe filters (Sartorius Stedim Biotech GmbH, Goettingen, Germany) and UV-treated for 30 min just before WGA. All samples were denatured using a chemical procedure that proved to be more suitable than thermal denaturation to obtain large size amplification products, necessary for fosmid library construction (Fig. S2). Approximately 2,500 sorted *Synechococcus* cells (corresponding to 5 µL in our sorting conditions) were lysed by adding 1 µL of lysis buffer, incubated for 10 min on ice before adding 1 µL of neutralization buffer. WGA reactions were carried out in 24 µL final volume by adding sample buffer (7 µL), reaction buffer (9 µL) and phi29 enzyme (1 µL; GE Healthcare, Waukesha, WI), then incubated for 2.5 h at 30 °C before inactivating the enzyme for 5 min at 65 °C. Blank controls with sterile water and sheath fluid collected at the exit of the chamber were also taken for each experiment. For PCR reactions, WGA products were diluted 10-fold in sterile MilliQ water. WGA products and dilutions were stored at -20 °C until processing.

PCR amplification and T-RFLP screening

Terminal-restriction fragment length polymorphism (T-RFLP) targeting either the 16S rRNA or *petB* genes was used to assess the purity of sorting and amplification biases. Genomic DNA was extracted from concentrated natural samples (~ 40,000 *Synechococcus* cells) and sorted samples (50-100,000 cells) using the DNA Blood and Tissue kit (Qiagen, Courtaboeuf, France), following the conditions described by Balzano *et al.* (2012). PCR reactions were performed in a 25 µL total volume containing 1.25 U GoTaq polymerase (Promega, Madison, WI), 1X polymerase buffer and 1.5 mM $MgCl_2$. Primers and PCR conditions are listed in Table 1. Forward primers 16S_27F and *petB*F were 5'-labeled with 6-carboxyfluorescein (6-FAM, Eurogentec, Seraing, Belgium) for fragment detection by T-RFLP. All amplifications were performed using a GeneAmp PCR system 9700 (Applied Biosystems, Carlsbad, California) and the program consisted of an initial denaturation step of 5 min at 94 °C, followed by 30 cycles (30 s at 94 °C, 30 s at 55 °C and 1 min at 72 °C) and a final extension

step of 10 min at 72 °C (Table 1). In order to assign 16S rRNA and *petB* terminal-restriction fragments (T-RFs), *Synechococcus* sequences from natural populations and cultured isolates (Mazard *et al.*, 2012) were used to generate T-RF databases by virtual *in silico* digestion using various restriction enzymes (Tables S1 and S2). While for the 16S rRNA gene, *MspI* was used as suggested by Mazard *et al.* (2011). This database allowed us to identify *HpaII/HinfII* as the most appropriate set of restriction enzymes for *petB*, for which it provided eight distinct T-RFs for the whole database (Table S2). Both amplicon types were digested for 3 h at 37 °C, using the manufacturer's recommendations (New England Biolabs, Evry, France). After inactivation of the enzyme at 80 °C for 15 min, T-RFs were diluted in Hi-Di™ Formamide (Applied Biosystems, Carlsbad, CA, USA), separated using a 3130xl Genetic Analyzer and analyzed using the Genemapper software (Applied Biosystems). Only T-RFs between 85 and 500 bp were included in the analysis.

Fosmid library construction

Twenty WGA products exhibiting no significant genetic diversity bias with regard to the initial sample, as checked by T-RFLP screening, were pooled for fosmid library construction. As previously suggested, DNA branching was reduced by cutting junctions using an enzymatic treatment with 1,000 U of Nuclease S1 (Promega, Madison, WI) for 1 h (Zhang *et al.*, 2006, Neufeld *et al.*, 2008). DNA was then purified by precipitation using SureClean reagent (Bioline France, Paris, France), washed using 70% ethanol and rehydrated in sterile MilliQ water. The metagenomic library was constructed using the CopyControl™ HTP Fosmid Library Production Kit with pCC2FOS™ Vector (Epicentre, Madison, USA). For the end-repair procedure, slight modifications to the standard protocol were made by adding 40 U of DNA polymerase I (New England Biolabs, Evry, France) to the enzymatic mix provided by the manufacturer and incubating the reaction at 25 °C for 3 h. The end-repaired DNA was separated on a 1% low melting point agarose gel (Invitrogen, Carlsbad, CA) in Tris Borate EDTA buffer 0.5x, at 200 V and a linearly ramped pulse time of 0.5 to 1.5 s for 20 h on a Pulse Field Gel Electrophoresis (PFGE) Chef DRII (BioRad, Hercules, CA). Fragments between 25 and 55 kbp were cut out of the gel after staining the DNA ladder with ethidium bromide. DNA was purified after

agarose digestion for 5 h at 45 °C in the presence of 10 $U.g^{-1}$ GELase (Epicentre), then concentrated on 100 kDa Amicon centrifugal filter units (Millipore, Molsheim, France). About 300 ng DNA was added to a ligation reaction with the pCC2FOS™ Vector and *in vitro* packaged according to the manufacturer's protocol. Once titrated, the entire phage suspension was used to infect an exponentially growing *Escherichia coli* EPI300-T1R culture. The infected metagenomic library was frozen in liquid nitrogen and stored at -80 °C in glycerol (20% final concentration) until screening. After thawing, the desired number of clones was plated on LB Agar containing 12.5 $\mu g.mL^{-1}$ chloramphenicol for subsequent screening.

Fosmid library screening

Colony picking of about 55,000 clones and library screening on a high-density colony filter were performed at the "Centre National de Ressources Génomiques Végétales" (Institut National de Recherche Agronomique, Toulouse, France), as described elsewhere (Gonthier *et al.*, 2010). In order to increase the number of fosmids containing the PBS region and the average coverage, hybridizations were performed using probes targeting three loci situated at different locations of this region. This included the *cpeC* gene encoding the PEI linker polypeptide, the *mpeBA* operon encoding the PEII α and β subunits and the *rpcBA/cpcBA* operon encoding the PC α and β subunits (see Fig. 3 in Six et al., 2007). Probes were synthesized by PCR using the pool of WGA products from the Astan station and degenerate primers that were designed with Primaclade (Gadberry *et al.*, 2005) and manually refined based on sequence alignments of these three loci (Table 1). PCR products were purified from agarose gels using the Qiaquick Gel Purification kit (Qiagen, Courtaboeuf, France) and quantified by the Quant-iT™ PicoGreen dsDNA assay (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions using a Tecan microplate reader (Tecan, Männedorf, Switzerland). Amplicons were labeled with [α-^{32}P]dCTP by random priming using Ready-To-Go DNA Labelling Beads and purified through Illustra™ ProbeQuant™ G-50 Micro Columns (GE Healthcare, Waukesha, WI). Filter hybridization were performed with a pool of the three probes as described elsewhere (Sambrook &

Russel, 2001). Filters were scanned and analyzed using a Storm 860 PhosphorImager (GE Healthcare, Waukesha, WI).

Fosmid selection and sequencing

To select fosmids to be sequenced, positives clones selected by hybridization were secondarily screened by colony PCR targeting the three above mentioned loci as well as another gene of the PBS region, *cpeU*, encoding a putative PE phycobilin lyase (Six *et al.*, 2007). After induction to high copy number, fosmid DNA was extracted using a NucleoSpin Plasmid kit (Macherey-Nagel, Hoerdt, France) and quantified with a ND-1000 spectrophotometer (Thermo Scientific, Wilmington, USA). To check the theoretical size and position of each insert within the PBS region, end-sequencing of PCR positive clones was performed according to the CopyControl™ HTP Fosmid Library Production kit instructions manual (Epicentre). The actual insert size was then determined by PFGE on 25 selected clones after fosmid DNA digestion by *NotI* as described above. These clones were pooled before generating a paired-end library, which was then sequenced using the Illumina technology on a MiSeq system (2 x 250 bp reads). This resulted in $6.9x10^6$ paired end reads with insert sizes ranging from 50 -250 bp and with about 75% of the reads overlapping over 10 bp. Reads corresponding to the vector and the *E. coli* host were removed through mapping using the Geneious Workbench (Biomatters, Auckland, New-Zealand). Remaining reads, corresponding to inserts, were then assembled using the CLC AssemblyCell© software (CLCBio, Prismet, Denmark) followed by contig elongation using CAP3 (Huang & Madan, 1999).

454 pyrosequencing

To complement the T-RFLP analyses, *petB*, *r/cpcBA* and *cpeBA* loci were sequenced by 454 pyrosequencing to assess the diversity of *Synechococcus* populations at each step of the metagenomic approach: after TFF, cell sorting and WGA. Amplifications were conducted with “fusion” primers comprising a 454 adaptor, a TCAG key used for amplicon sequencing and a multiplex identifier

(MID) tag for sample identification from a mixed pool. The *cpeBA* operon was amplified using the primer set SynB1F - SynA3R (Table 1). Amplification of the *r/cpcBA* operon was performed using the reverse primer designed by Haverkamp *et al.* (2008) and a forward primer newly designed using Primaclade (Gadberry *et al.*, 2005) on a manually refined multiple sequence alignment containing *Synechococcus* sequences from marine, brackish and freshwater environments. This primer set was then tested on a selection of 14 marine *Synechococcus* strains, representative of the different pigment types (Six *et al.*, 2007, Humily *et al.*, in press). PCR reactions were performed in triplicates, in 50 µL final volume following conditions described in Table 1, using the minimum number of cycles to obtain a thin band on agarose gel to minimize PCR biases. Amplicons were pooled and purified using the NucleoSpin® Extract II kit (Macherey-Nagel, Hoerdt, France) to obtain a final amount of at least 2 µg DNA. All samples were then pooled in equimolar amounts for 454 pyrosequencing at the Biogenouest® Genomics platform (Rennes, France) on a Roche GS-FLX system (454 Life Sciences, Roche, Brandford, CT, USA) using an emPCR amplicon kit Lib-L according to the manufacturer's protocol.

Sequence treatment and phylogenetic analysis

Low quality sequences, i.e., lacking a key, exhibiting unexpected amplicon lengths (<250 bp or >600 bp) or containing more than 7% undetermined bases, were removed from the analysis. Sequences were denoised using Acacia (v 1.52; (Bragg *et al.*, 2012) after trimming to 450 bp for *petB* and to 430 bp for the other loci. Intergenic sequences within the *r/cpcB* and *r/cpcA* operons were then removed before clustering using MOTHUR (Schloss *et al.*, 2009). To define an OTU, a cut-off value of 94% nt sequence identity was used for *petB*, as previously described (Mazard *et al.*, 2012), and 95% for the other loci. OTUs defined by singletons were removed before calculation of rarefaction curves, diversity indices and coverage values using MOTHUR. Sequences were then used to generate multiple alignments using MAFFT (G-INS-I option, v 6.953; Katoh *et al.*, 2005).

Taxonomic assignment of each OTU was made by phylogenetic analyses using i) sequences from the same database as the one used for T-RF identification for *petB*, and ii) sequences retrieved from 42

public or unpublished marine *Synechococcus/Cyanobium* genomes for the *r/cpcBA* and *cpeBA* loci Novel *r/cpcBA* and *cpeBA* sequences were deposited in Genbank under the accession numbers KF528763-KF528785 and KF528801-KF528826, respectively. Phylogenetic trees of R/CpcBA (340 aa positions) were performed by Maximum Likelihood using PhyML (v3.0; Guindon *et al.*, 2009), a LG+G model (Le & Gascuel, 2008) and a BIONJ starting tree and visualized using FigTree v1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/). Robustness of inferred topology was supported by 1000 bootstraps resampling. For *cpeBA*, bayesian inference was conducted on 999 nt positions using MrBayes (v3.1.2; Huelsenbeck & Ronquist, 2001), and started with a random tree, run for 2 million generations in four chains, sampling every 100 generations and burn-in of 5,000 trees. For both loci, pyrotag sequences and a few additional partial sequences retrieved from Genbank were then aligned with the reference alignments and added to consensus trees using the ADD_BY_PARSIMONY algorithm implemented in ARB (Ludwig *et al.*, 2004). Environmental gene sequences for *petB, r/cpcBA and cpeBA* amplicons obtained in this study were deposited in the Sequence Read Archive (SRA) database under the following accession number: SRP028891.

Results and discussion

Although major breakthroughs have been achieved in our understanding of the composition and dynamics of marine microbes (Pedros-Alio, 2006, DeLong, 2009, Gilbert & Dupont, 2011), accessing the functional diversity of natural populations remains tricky. Indeed, the ability of an organism to perform a particular biochemical pathway generally requires the co-occurrence of several genes, which in prokaryotes are often organized as clusters. One way to deal with this problem is to sequence single amplified genomes (SAG). However, this approach currently only allows one to amplify and sequence partial genomes with no control on the sequenced genomic regions (Zhang *et al.*, 2006, Rodrigue *et al.*, 2009, Malmstrom *et al.*, 2013) and requires a deep sequencing effort to access the diversity of biochemical pathways occurring within a population.

Here, we developed a metagenomic pipeline, combining flow cytometry cell sorting, WGA and fosmid library construction, which allowed us not only to specifically target *Synechococcus* cells but also particular regions of its genome (Fig. 1). Since this approach is not devoid of biases inherent to each method, we evaluated their effects at each step of the procedure by following both the genetic composition of the *Synechococcus* community and the relative abundance of contaminant heterotrophic bacteria by T-RFLP analyses and/or pyrosequencing. T-RFLP, a highly reproducible, low cost but only semi-quantitative technique, was used to rapidly screen a high number of samples, and pyrosequencing to obtain a more thorough picture of the whole genetic and pigment diversity of samples.

Efficiency of the combination of cell sorting and WGA to obtain large DNA amounts from natural *Synechococcus* populations

T-RFLP targeting the 16S rRNA gene was used to assess the proportion of co-occurring heterotrophic cells at different steps of the procedure: i) in the initial sample concentrated by TFF, ii) in flow sorted samples (including 3 technical replicates, named A to C, each comprising 3 consecutive sorts) and iii) in all samples obtained after WGA of the three final sorted samples (Figs. 2A, 2C and S3). *In silico* analysis of a 16S rRNA *Synechococcus* reference database (cf. materials and methods) predicted *Synechococcus* T-RFs of 129, 491 and 492 bp and, by extension, all other T-RFs were assumed to arise from heterotrophic contaminants (Table S1). Compared to these predictions, the sizes of *Synechococcus* T-RFs actually observed in sorted samples from the Astan station were 130 bp (*Syn*. T-RF I), 482-485 bp (*Syn*. T-RF II) and 487-489 bp (*Syn*. T-RF III; Fig. 2A), and this assignment was independently confirmed by T-RFLP analyses of *Synechococcus* isolates (data not shown) and by *petB* sequencing of sorted samples from Astan (Fig. S3). Although somewhat large (up to 7 bp), such a drift between predicted and measured T-RF sizes has already been reported and was shown to increase with T-RF size (Kaplan & Kitts, 2003, Bukovská *et al.*, 2010). The observed duplication of *Syn*. T-RF II and III may possibly be due to polymerase errors during PCR and/or during WGA. It is also worth noting that the flow cytometry set up used in this study (Exc 488 nm) was not suitable to detect PC-

rich *Synechococcus* cells (pigment type 1), sometimes reported in coastal or low salinity waters (Haverkamp *et al.*, 2008). However, a complementary flow cytometric analysis of the initial Astan sample using dual lasers emitting at 488 nm and 635 nm (exciting PC) confirmed that only PE-containing *Synechococcus* cells were present in this sample (data not shown). As expected, no *Prochlorococcus* cells were detected since this genus is systematically absent from English Channel waters (N. Simon, pers. comm).

While the concentration of *Synechococcus* in the initial seawater sample was fairly low (~ 840 cells.ml^{-1}), virtually no residual co-occurring bacterial or eukaryotic cells could be detected after cell sorting in all three replicates, either by flow cytometry (< 2%; Fig. S1) or by T-RFLP using the 16S rRNA gene (Fig. 2C). The three-step sorting procedure thus very efficiently removed all types of contaminants and was highly reproducible, since the three technical replicates gave similar results. While the first sorting step, based on natural fluorescence and side scatter, efficiently eliminated photosynthetic picoeukaryotes and large non-living particles, the two following sorting steps, performed after DNA staining using a combination of SYTO-9 and SYTO-13, allowed us to almost completely remove heterotrophic bacteria. This absence of contamination shows that this protocol also efficiently eliminated any free DNA, often present at high concentrations in environmental samples especially after TFF that can lead to cell lysis (Rodrigue *et al.*, 2009).

After the WGA step, some T-RF peaks assigned to contaminants in the initial sample re-appeared in variable proportion, i.e., between 0-46 % of the total bacterial T-RFs (Figs. 2C and S3). In one of the WGA products, contaminants even represented all of the T-RFs present in the sample. As suggested in a previous study (Swan *et al.*, 2011), a discrepancy in cell lysis efficiency between *Synechococcus* and their bacterial counterparts might explain this differential amplification. Although preliminary tests performed on cultured isolates showed that thermal lysis was slightly more efficient to break *Synechococcus* cells than chemical lysis (data not shown), we did not retain the former option, which generates smaller WGA products (Fig. S2). This differential amplification may also suggest that heterotrophic bacterial DNA could be preferentially amplified over *Synechococcus* DNA during the WGA step. It is also worth noting that the relative proportion of contaminants was higher in

WGA reactions performed on the second and third sorted replicates (sorts B and C), which chronologically were performed after sort A (Fig 2C). This may result from a lower integrity of *Synechococcus* cells and/or a higher sensitivity to DNA degradation in the TFF concentrated sample that was kept on ice for several hours before cell sorting. Such modifications of the apparent global community structure as a result of WGA have already been reported in the literature and were attributed to either amplification of free contaminant DNA (Raghunathan *et al.*, 2005, Kvist *et al.*, 2007, Woyke *et al.*, 2010) or differential amplification between taxa (Abulencia *et al.*, 2006, Chen *et al.*, 2008, Lepère *et al.*, 2011, Blainey, 2013). The latter effect was suggested to result from various parameters. These include a difference in GC%, the presence of repeated regions (Pinard *et al.*, 2006), different chromosome conformation (Schoenfeld *et al.*, 2010), chromosomal breaks, relative cell abundance, the DNA amount of the various taxa present in the sample (Chen *et al.*, 2008, Arakaki *et al.*, 2010) and/or stochastic effects during initial priming and amplification (Rodrigue *et al.*, 2009). In the present study, screening of the WGA products using T-RFLP targeting the 16S rRNA gene allowed us to circumvent this issue by selecting samples with very low contaminant load.

Effect of cell sorting and WGA on the apparent composition of *Synechococcus* communities

The aforementioned differential amplification of taxa during the WGA step may also affect the relative abundance of the different clades within the *Synechococcus* population. To estimate this potential bias, we developed a T-RFLP approach targeting the *petB* gene, encoding the cytochrome b_6 subunit of the cytochrome b_6/f complex, which provides a much finer taxonomic resolution of the genetic diversity within the marine *Synechococcus* radiation than 16S rRNA (Mazard *et al.*, 2012). It must be noted that a given T-RF may correspond to different (sub-)clades and that reciprocally some strains belonging to the same (sub-)clade may have different T-RFs (Table S2), indicating that this approach does not perfectly reflect the composition of the *Synechococcus* population at the (sub-)clade level. Still, variations in the T-RF pattern observed between the different stages of the metagenomic

pipeline can be used for rapid screening of changes in the *Synechococcus* community composition between samples.

The set of restriction enzymes *HpaII/HinfII* used for the T-RFLP analyses of the Astan samples provided two T-RFs (125 and 234 bp), which were among the eight T-RFs predicted from *in silico* analysis of a database, including sequences from both *Synechococcus* isolates and environmental samples (Fig. 2B and 2D; Table S2). These T-RF sizes were obtained from strains belonging to the four main *Synechococcus* clades dominating in oceanic regimes (I-IV; Zwirglmaier *et al.*, 2008), with T-RF 125 bp encompassing strains from clades I, III and IV and T-RF 234 bp, strains from clades II, IV but also VIII, XVI and sub-cluster 5.3-I (Table S2). Although a moderate change was observed in the relative proportion of the two T-RFs in all 3 replicates after cell sorting, the 125 bp T-RF remained dominant in the sorted samples (Fig. 2D). WGA also seemingly allowed us to preserve the diversity of the initial sample since both T-RFs were found in similar proportion after this step, despite some variability in the community composition between WGA products. In combination with 16S rRNA T-RFLP, these *petB* profiles helped us to select twenty WGA products, not only with a low contamination level by heterotrophic bacteria but also representative of the *Synechococcus* community present at the Astan site. One third of this pool, equivalent to about 3.5 µg DNA, was used to construct the fosmid library.

Assessment of the quality of the fosmid library from the Astan station

The fosmid library prepared from the Astan station was seemingly representative of the genetic diversity occurring at this station, since it contained more than 500,000 clones, among which 10 % were randomly screened by hybridization using a pool of three PBS probes targeting the *c*/*rpcBA*, *mpeBA* and *cpeC* loci, then by PCR targeting the same genes as well as *cpeU*. 94% of the clones selected by hybridization contained at least one of the four genes screened by PCR (Table S3). Twenty five clones were selected for sequencing based on their genetic diversity as assessed by BLASTN analyses of fosmid ends and by their location in the PBS region compared to reference *Synechococcus* genomes. This approach also revealed the presence of a few probable chimeric sequences, as shown

by an incongruent predicted insert size with regard to PFGE size measurements, which were not retained for sequencing. Mapping of all reads onto a reference PBS region showed that there was a good and homogeneous coverage of the region (average : 28,641 ± 6,409 fold, Fig. 3A). Assembly of these reads followed by a contig elongation step (see materials and methods) led to 64 contigs with an average size of 4,315 bp and an N50 of 7,804 bp (Table S3). Among these, 41 contigs were located within the PBS gene region, including 7 exhibiting more than 98 % nucleotide identity with others that were not included in the analysis. The 34 remaining PBS contigs had an average size of 5,033 bp and an N50 of 11,113 bp and were deposited in Genbank under accession numbers KF846537- KF846570. Mapping of these contigs onto reference picocyanobacterial genomes then allowed us to examine the PBS gene diversity and genomic context and to assign these sequences to specific *Synechococcus* clades and pigment types (see below).

Analysis of *Synechococcus* community structure and pigment type at the Astan station off Roscoff

The SOMLIT-Astan site is representative of permanently mixed waters of the Western Channel as no stratification occurs during the year (Sournia & Birrien, 1995). It is characterized by mesotrophic waters with moderate nitrate and phosphate concentrations and is weakly influenced by the Penzé estuary located near Roscoff. Picophytoplankton is dominated year-round by picoeukaryotes (Not *et al.*, 2004), while the photosynthetic prokaryote fraction is exclusively represented by *Synechococcus,* which can reach abundances exceeding 10^4 cells.mL^{-1} (N. Simon, pers. comm.). Sampling was performed in May, following the early spring bloom of *Synechococcus,* occurring in March due to water mixing.

Genetic diversity of the *Synechococcus* population was analyzed using pyrosequencing targeting the *petB* gene. After cleaning and denoising, at least 582 *petB* amplicons were obtained for each sample (Fig. 4), which proved sufficient to cover the whole *Synechococcus* diversity, as attested by a Good's coverage index close or equal to 1 (Table 2). Additionally, rarefaction curves were performed at 94% identity, a cut off previously used by Mazard et al. (2012) and shown to be suitable for

assignment of sequences at the sub-clade level (Fig. S4). Altogether, this led to the identification of 9 OTUs among the entire dataset, but only 4 were represented by more than 4 reads, suggesting that these rare OTUs might correspond to sequencing errors and that the curves were in fact saturated in all samples. It is also worth noting that based on several diversity estimators, including the S_{Chao1} and Shannon-Weaver diversity indices (Table 2; Shannon & Weaver, 1949, Chao *et al.*, 2005), no significant modifications of the community structure occurred along the metagenomic pipeline (Fig. 4A and Table 2). The *Synechococcus* population at the Astan station at the sampling date appears largely dominated by clade I, with sub-clade Ib representing more than 96 % of the total reads, clade Ia up to 3%, while the remaining community was composed of members of sub-clades Ic and IVa (Fig. 4A). These results are consistent with observations made in the Western Atlantic Ocean (Ahlgren & Rocap, 2012) or in cold North Atlantic mesotrophic waters where representatives of clade I are always dominate over clade IV (Huang *et al.*, 2012, Mazard *et al.*, 2012). However, this contrasts with abundance profiles found in California coastal waters (Tai & Palenik, 2009), or in Arctic, South Pacific and Eastern Atlantic Oceans where clade IV cells typically dominate (Zwirglmaier *et al.*, 2008).

Pyrosequencing of the *r/cpcBA* and *cpeBA* operons (encoding α and β subunits of PC and PE I, respectively) was used to assess the pigment type diversity at the Astan station, to complement the analysis of PBS-containing fosmids. It must be noted that phylogenetic analyses made with these genes are not congruent with those based on core genes (including *petB*), likely due to lateral gene transfers that would have occurred during the evolution of these antenna genes (Six *et al.*, 2007, Everroad & Wood, 2012). As for *petB*, Good's index indicated that the minimum number of sequences obtained for each operon, 3,671 for *r/cpcBA* and 260 for *cpeBA* (Table 2), was sufficient to cover properly the pigment diversity in all samples analyzed. It is noteworthy that the low number of *cpeBA* sequences is due to the fact that a high proportion (>80%) of amplicons were discarded due to the non-specific amplification of the *mpeBA* operon. Rarefaction curves led to the identification of 15 *r/cpcBA* OTUs and 11 *cpeBA* OTUs among the entire dataset but only 12 and 8, respectively, were represented by more than 4 reads, indicating that the diversity of these samples was also well covered (Fig. S4).

For both markers, the cell sorting step apparently did not alter the *Synechococcus* community composition in contrast to the WGA step, which led to a notable modification of the relative proportion of the different OTUs in most samples (Fig. 4B-C). This seemingly higher variability with regard to *petB* might be due to the larger number of OTUs discriminated by these two markers, better highlighting the differential taxa amplification caused by WGA (Fig. 5B-C). Yet, it is important to note that all dominant OTUs are still present in WGA samples.

We then performed phylogenetic analyses of these sequences together with laboratory strains of known pigmentation to try and assign OTUs obtained by pyrosequencing and contigs retrieved from fosmid sequencing to specific pigment types (Fig. 5 and 6). The *r/cpcBA* tree resolved well pigment types 1 or 2 (*sensu* Six *et al.*, 2007), since strains exhibiting these pigmentations formed two well-defined clusters. In contrast, this marker was unable to distinguish between the different sub-types of pigment type 3. None of the *r/cpcBA* sequences obtained at the Astan station, either by pyrosequencing or from fosmid libraries, clustered with pigment types 1 or 2, suggesting that there were no representatives of these two pigment types in the Astan sample. The *cpeBA* tree brings complementary information since it splits members of pigment type 3 into distinct clusters. The first one gathered pigment types 3dA strains, i.e., strains that are able to perform type IV chromatic acclimation (CA4), or related strains (e.g., 3aA) that possess a CA4-A genomic island involved in this process but display a different pigment phenotype (Humily *et al.*, in press). The second one encompassed pigment type 3c, i.e., high PUB strains, and 3dB, i.e., CA4 strains possessing a CA4-B genomic island. Finally, strains branching at the base of these two clusters all belong to pigment type 3a, i.e., low PUB strains. Thus, the *cpeBA* phylogeny can be used to putatively assign the most abundant OTU at station Astan (*cpeBA*_OTU1 in Fig. 4C) to type 3a and the second and third to type 3dA (or related strains), these abundant OTUs representing 63, 20 and 8 % of the total *cpeBA* reads, respectively. Surprisingly, all *cpeBA* sequences retrieved from fosmid clones corresponded to pigment type 3dA and none to pigment type 3a, likely because primers used to screen this library were too specific given the set of reference genomes available and have selected out fosmids containing type 3a

PBS regions. The new pyrosequencing data obtained for these markers genes in natural populations should allow us to refine these probes in the future.

The metagenomic pipeline developed in this study (Fig. 1) brought complementary information with regard to the pyrosequencing analysis, since it allowed us to sequence suites of PBS genes, while pyrosequencing can only be applied to the most highly conserved genes of the region, i.e. those that were used as probes for fosmid screening (Fig. 3). Mapping of all assembled contigs onto each reference genome showed that the majority of environmental PBS regions from the Astan station could be assigned to clades I and IV as expected from the pyrosequencing data. 24 contigs out of 36, resulting from the assembly of about 64 % of total reads, were attributable to clade IV (Fig. 3B). Indeed, these contigs only comprised genes with highest identity/similarity to either *Synechococcus* sp. BL107 or *Synechococcus* sp. CC9902, the only sequenced clade IV strains, which are closely related (100% 16S rRNA gene identity, average nucleotide identity of 91.3 %; Dufresne *et al.*, 2008). A significant microdiversity was found all over the PBS region, suggesting that 4 to 6 clade IV genotypes co-occurred in this sample. It is noteworthy that the most variable genes of the region coded for phycobilin lyases (e.g., *cpeS*, *cpeT*, *mpeU*, *cpeY*, *rpcEF*) and uncharacterized proteins (e.g. *unk2* and *unk11*). Additionally, about one third of the reads were assembled into contigs that mapped onto two clade I reference genomes: MVIR-18-1 belonging to sub-clade Ia and CC9311 to sub-clade Ib. These contigs seemingly correspond to the two main environmental genotypes. The first one, for which we retrieved a complete PBS region, comprises alternating gene cassettes with highest percentages of identity with sub-clade Ia or Ib, while the second one mainly comprises genes related to sub-clade Ib. Interestingly, the remaining reads (~4%) could be assembled into one main 12.8 kb contig, which contains many genes very distant from all reference genomes available, and likely belong to a new environmental clade.

These contigs also provide important insights about the pigment type of the cells from which they originate. As expected from *cpeBA* phylogenetic analyses (Fig. 6), most, if not all, of the contigs mapping to clades I and IV can be assigned to pigment type 3dA (i.e., CA4-A) since these environmental PBS regions displayed a remarkable synteny with regard to reference genomes of

strains belonging to this pigment type, apart from the 5'-end of these regions which presented more variability. In particular, they exhibited a succession of genes from *mpeC* to *cpeF* (previously called *mpeV*), which is typical of this pigment type (Six *et al.*, 2007, Humily *et al.*, in press). By comparison, strains exhibiting pigment type 3c possess an *unk10* gene before *mpeU* and do not possess *cpeF*, while pigment type 3a not only lack *mpeC* and *mpeU* but also have a different organization of the remaining genes of this region (i.e., *cpeY- unk11-unk12-cpeF-cpeZ*). The PBS region found in the unassigned clade (bottom of Fig. 3B) is also original in this context. Indeed, gene content and organization of the covered region is identical to pigment type 3dA, except that it lacks the *mpeU* gene, a deletion previously observed only in MVIR-18-1, and that seemingly induces a strong alteration of the cell pigment content (i.e., low PUB/PEB phenotype instead of a CA4 phenotype; Humily et al., in press). By extension, it is possible that this new environmental genotype may exhibit the same pigmentation.

Conclusions and future applications

In this paper, we described an original metagenomic pipeline that we used to explore the diversity of *Synechococcus* pigment types but that can be applied to target any other region of the genome (Fig. 1). This approach is particularly well suited for such a model organism because of its abundance in the field and its natural autofluorescence, which makes it easy to enumerate and sort by flow cytometry. The use of control molecular methods allowed us to circumvent most of the (often ignored) biases inherent to targeted metagenomics: T-RFLP to bypass the differential amplification of taxa included in the initial environmental and fosmid end-sequencing to avoid chimeras, both effects being mainly due to the WGA step. This strategy provided libraries qualitatively (though not quantitatively) representative of the genetic and functional diversity of the initial sample, as attested by pyrosequencing of *petB*, *r/cpcBA* and *cpeBA* loci. Although phylogenetic analyses of the resulting sequences showed that there was a co-dominance of pigment type 3a and 3d in the Astan sample, which was maintained along the whole metagenomic pipeline, only fosmids of the latter pigment type were apparently selected for sequencing, while those of the former pigment type have most likely been counter-selected during our screening procedure. However, this issue could be avoided through the

sequencing of a higher number of fosmids and/or the use of a more encompassing mix of probes, better representing the studied community.

The dominance of pigment type 3a at the Astan station is congruent with previous observations using dual beam flow cytometry or spectrofluorimetry, which have shown the prevalence of low-PUB cells in coastal environments (Olson *et al.*, 1990, Lantoine & Neveux, 1997, Wood *et al.*, 1998). The latter cells were shown to almost always co-occur with high-PUB cells in these environments (Olson *et al.*, 1990, Katano *et al.*, 2007). Results from the present study suggest that these coastal high-PUB strains could actually be 3d (CA4) cells and not, as previously assumed, pigment type 3c, i.e., cells with a fixed high-PUB pigmentation, since those were not observed in our sequence libraries. By extension, it is possible that pigment type 3c could rather be restricted to oceanic waters since these high-PUB cells most efficiently capture blue photons prevailing in these environments (Olson *et al.*, 1990, Kirk, 1994).

To our knowledge, this is first study that shows the presence of CA4 strains in the natural environment. Indeed, due to their variable PUB to PEB ratio, they cannot be easily discriminated from pigment types with fixed pigment ratio (mainly 3a and 3c) by fluorescence-based approaches. Although a single environmental sample was analyzed here, the application of this molecular approach to a variety of trophic environments and depths should allow one to decipher the distribution and ecological importance of CA4 cells and more globally of all *Synechococcus* pigment types.

Acknowledgements

This work was supported by the collaborative program METASYN with Genoscope, the “Agence Nationale de la Recherche” Microbial Genomics Program (PELICAN, ANR-09-GENM-030), UK Natural Environment Research Council grant (NE/I00985X/1) and the European Union's Seventh Framework Programs, MicroB3 and MaCuMBA (grant agreements 287589 and 311975, respectively). We thank Fabienne Rigaud-Jalabert for sample collection at the Astan station and Thierry Cariou ("Service d’Observation en Milieu LITtoral (SOMLIT)", INSU-CNRS, Station Biologique de

Roscoff) for providing physico chemical parameters at the sampling site. We are most grateful to the Biogenouest® Genomics core facilities for their technical support, notably Gwenn Tanguy, for genotyping operations in Roscoff and Alexandra Dheilly, Delphine Naquin and Oscar Lima for pyrosequencing in Rennes. Hélène Bergès, Arnaud Bellec and Joëlle Fourment from the CNRGV Platform are acknowledged for fosmid library screening. We are indebted to Dominique Boeuf for help with phylogenetic analyses.

References

Abulencia CB, Wyborski DL, Garcia JA, *et al.* (2006) Environmental whole-genome amplification to access microbial populations in contaminated sediments. *Appl. Environ. Microbiol.* **72**: 3291-3301.

Ahlgren NA & Rocap G (2012) Diversity and distribution of marine *Synechococcus*: multiple gene phylogenies for consensus classification and development of qPCR assays for sensitive measurement of clades in the ocean. *Front. Microbiol.* **3**: doi: 10.3389/fmicb.2012.00213.

Aminot A & Kérouel R (2007) *Dosage automatique des nutriments dans les eaux marines : méthodes en flux continu.* . Editions Quae, Versailles, France.

Arakaki A, Shibusawa M, Hosokawa M & Matsunaga T (2010) Preparation of genomic DNA from a single species of uncultured magnetotactic bacterium by multiple-displacement amplification. *Appl. Environ. Microbiol.* **76**: 1480-1485.

Balzano S, Marie D, Gourvil P & Vaulot D (2012) Composition of the summer photosynthetic pico and nanoplankton communities in the Beaufort Sea assessed by T-RFLP and sequences of the 18S rRNA gene from flow cytometry sorted samples. *ISME J.* **6**: 1480-1498.

Blainey PC (2013) The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol. Rev.* **37**: 407-427.

Bragg L, Stone G, Imelfort M, Hugenholtz P & Tyson GW (2012) Fast, accurate error-correction of amplicon pyrosequences using Acacia. *Nature Methods* **9**: 425-426.

Buitenhuis ET, Li WKW, Vaulot D, *et al.* (2012) Bacterial biomass distribution in the global ocean. *Earth System Science Data Discussions* **5**: 301-315.

Bukovská P, Jelínková M, Hršelová H, Sýkorová Z & Gryndler M (2010) Terminal restriction fragment length measurement errors are affected mainly by fragment length, G + C nucleotide content and secondary structure melting point. *J. Microbiol. Meth.* **82**: 223-228.

Chao A, Chazdon RL, Colwell RK & Shen TJ (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol. Lett.* **8**: 148-159.

Chen Y, Dumont MG, Neufeld JD, *et al.* (2008) Revealing the uncultivated majority: combining DNA stable-isotope probing, multiple displacement amplification and metagenomic analyses of uncultivated *Methylocystis* in acidic peatlands. *Environ. Microbiol.* **10**: 2609-2622.

Crosbie ND, Pöckl M & Weisse T (2003) Dispersal and phylogenetic diversity of non-marine picocyanobacteria, inferred from 16S rRNA gene and *cpcBA*-intergenic spacer sequence analyses. *Appl. Environ. Microbiol.* **69**: 5716-5721.

Cuadros-Orellana S, Martin-Cuadrado AB, Legault B, D'Auria G, Zhaxybayeva O, Papke RT & Rodriguez-Valera F (2007) Genomic plasticity in prokaryotes: the case

of the square haloarchaeon. . *ISME J* **1**: 235-245.

Cuvelier ML, Allen AE, Monier A, *et al.* (2010) Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proc. Natl. Acad. Sci. U. S. A* **107**: 14679-14684.

Dean FB, Nelson JR, Giesler TL & Lasken RS (2001) Rapid amplification of plasmid and phage DNA using phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* **11**: 1095-1099.

del Giorgio PA, Bird DF, Prairie YT & Planas D (1996) Flow cytometric determination of bacterial abundance in lake plankton with the green nucleic acid stain SYTO 13. *Limnol. Oceanogr.* **41**: 783-789.

DeLong EF (2009) The microbial ocean from genomes to biomes. *Nature* **459**: 200-206.

Dufresne A, Ostrowski M, Scanlan DJ, *et al.* (2008) Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol* **9**: R90.

Everroad CR & Wood MA (2006) Comparative molecular evolution of newly discovered picocyanobacterial strains reveals a phylogenetically informative variable region a β-phycoerythrin. *J. Phycol.* **42**: 1300-1311.

Everroad CR & Wood MA (2012) Phycoerythrin evolution and diversification of spectral phenotype in marine *Synechococcus* and related picocyanobacteria. *Mol. Phylogenet. Evol.* **64**: 381-392.

Gadberry MD, Malcomber ST, Doust AN & Kellogg EA (2005) Primaclade - a flexible tool to find conserved PCR primers across multiple species. *Bioinformatics* **21**: 1263-1264.

Gasol JM, Zweifel UL, Peters F, Fuhrman JA & Hagstrom A (1999) Significance of size and nucleic acid content heterogeneity as measured by flow cytometry in natural planktonic bacteria. *Appl. Environ. Microbiol.* **65**: 4475-4483.

Gilbert JA & Dupont CL (2011) Microbial metagenomics: beyond the genome. *Ann. Rev. Mar. Sci.* **3**: 347-371.

Gonthier L, Bellec A, Blassiau C, *et al.* (2010) Construction and characterization of two BAC libraries representing a deep-coverage of the genome of chicory (*Cichorium intybus* L., Asteraceae). *BMC Research Notes* **3**: 225.

Guindon S, Dufayard JF, Hordijk W, Lefort V & Gascuel O (2009) PhyML: Fast and accurate phylogeny reconstruction by Maximum Likelihood. *Infect. Genet. Evol.* **9**: 384-385.

Hallam SJ, Mincer TJ, Schleper C, Preston CM, Roberts K, Richardson PM & DeLong EF (2006) Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine Crenarchaeota. *PLoS Biol.* **4**: e95.

Haverkamp T, Acinas SG, Doeleman M, Stomp M, Huisman J & Stal LJ (2008) Diversity and phylogeny of Baltic Sea picocyanobacteria inferred from their ITS and phycobiliprotein operons. *Environ. Microbiol.* **10**: 174-188.

Huang S, Wilhelm SW, Rodger Harvey H, Taylor K, Jiao N & Chen F (2012) Novel lineages of *Prochlorococcus* and *Synechococcus* in the global oceans. *ISME J.* **6**: 285-297.

Huang X & Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Res* **9**: 868-877.

Huelsenbeck JP & Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754-755.

Humily F, Partensky F, Six C, *et al.* (in press) A gene island with dual evolutionary origin is involved in chromatic acclimation in marine *Synechococcus*. *PLOS one*.

Jardillier L, Zubkov MV, Pearman J & Scanlan DJ (2010) Significant CO2 fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. *ISME J* **4**: 1180-1192.

Kalyuzhnaya MG, Lapidus A, Ivanova N, *et al.* (2008) High-resolution metagenomics targets specific functional types in complex microbial communities. *Nature Biotech.* **26**: 1029-1034.

Kaplan CW & Kitts CL (2003) Variation between observed and true terminal restriction fragment length is dependent on true TRF length and purine content. *J. Microbiol. Meth.* **54**: 121-125.

Katano T, Kaneda A, Kanzaki N, *et al.* (2007) Distribution of prokaryotic picophytoplankton from Seto Inland Sea to the Kuroshio region, with special reference to 'Kyucho' events. *Aquat. Microb. Ecol.* **46**: 191-201.

Katoh K, Kuma K, Toh H & Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucl. Acids Res.* **33**: 511-518.

Kirk JTO (1994) *Light and photosynthesis in aquatic ecosystems*. Cambridge University Press, Cambridge, UK.

Koroleff F (1970) Direct determination of ammonia in natural waters as indophenol blue. . Vol. 9 ed.^eds.), p.^pp.

Kvist T, Ahring BK, Lasken RS & Westermann P (2007) Specific single-cell isolation and genomic amplification of uncultured microorganisms. *Appl. Microbiol. Biotechnol.* **74**: 926-935.

Lantoine F & Neveux J (1997) Spatial and seasonal variations in abundance and spectral characteristics of phycoerythrins in the tropical northeastern Atlantic Ocean. *Deep-Sea Res. Pt I.* **44**: 223-246.

Lasken RS & Stockwell TB (2007) Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol.* **7**: 19.

Le SQ & Gascuel O (2008) An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**: 1307-1320.

Lebaron P, Parthuisot N & Catala P (1998) Comparison of blue nucleic acid dyes for flow cytometric enumeration of Bacteria in aquatic systems. *Appl. Environ. Microbiol.* **64**: 1725-1730.

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**: 513-523.

Li W (1994) Primary productivity of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: measurements from flow cytometric sorting. *Limnol. Oceanogr.* **39**: 169-175.

Ludwig W, Strunk O, Westram R, *et al.* (2004) ARB: a software environment for sequence data. *Nucl. Acids Res.* **32**: 1363-1371.

Malmstrom RR, Rodrigue S, Huang KH, *et al.* (2013) Ecology of uncultured *Prochlorococcus* clades revealed through single-cell genomics and biogeographic analysis. *ISME J.* **7**: 184-198.

Marie D, Shi XL, Rigaut-Jalabert F & Vaulot D (2010) Use of flow cytometric sorting to better assess the diversity of small photosynthetic eukaryotes in the English Channel. *FEMS Microbiol. Ecol.* **72**: 165-178.

Marie D, Brussaard C, Partensky F, Vaulot D & Wiley J (1999) Flow cytometric analysis of phytoplankton, bacteria and viruses. *Current Protocols in Cytometry* **Supplement 10**: 11.11.11-11.11.15.

Martín HG, Ivanova N, Kunin V, *et al.* (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nature Biotech.* **24**: 1263-1269.

Mazard S, Ostrowski M, Garczarek L & Scanlan DJ (2011) A targeted metagenomic approach to determine the "Population Genome" of marine *Synechococcus*. *Handbook of Molecular Microbial Ecology II: Metagenomics in Different Habitats*,(de Bruijn FJ, ed.^eds.), p.^pp. 301-307.

Mazard S, Ostrowski M, Partensky F & Scanlan DJ (2012) Multi-locus sequence analysis, taxonomic resolution and biogeography of marine Synechococcus. *Environ. Microbiol.* **14**: 372-386.

Mella-Flores D, Mazard S, Humily F, *et al.* (2011) Is the distribution of *Prochlorococcus* and *Synechococcus* ecotypes in the Mediterranean Sea affected by global warming? *Biogeosciences* **8**: 2785-2804.

Neufeld JD, Chen Y, Dumont MG & Murrell JC (2008) Marine methylotrophs revealed by stable-isotope probing, multiple displacement amplification and metagenomics. *Environ. Microbiol.* **10**: 1526-1535.

Not F, Latasa M, Marie D, Cariou T, Vaulot D & Simon N (2004) A single species, Micromonas pusilla (Prasinophyceae), dominates the eukaryotic picoplankton in the Western English Channel. *Appl Environ Microbiol* **70**: 4064-4072.

Olson RJ, Chisholm SW, Zettler ER & Armbrust EV (1988) Analysis of *Synechococcus* pigment types in the sea using single and dual beam flow-cytometry. *Deep-Sea Res.* **35**: 425-440.

Olson RJ, Chisholm SW, Zettler ER & Armbrust EV (1990) Pigments, size, and distribution of *Synechococcus* in the North Atlantic and Pacific Oceans. *Limnol. Oceanogr.* **35**: 45-58.

Palenik B (2001) Chromatic adaptation in marine *Synechococcus* strains. *Appl. Environ. Microbiol.* **67**: 991-994.

Partensky F, Blanchot G & Vaulot D (1999) Differential distribution and ecology of *Prochlorococcus* and *Synechococcus* in oceanic waters : a review. *Bulletin de l'Institut océanographique* **19** (Marine Cyanobacteria): 457-475.

Pedros-Alio C (2006) Genomics and marine microbial ecology. *Int Microbiol* **9**: 191-197.

Pedrós-Alió C (2012) The rare bacterial biosphere. *Ann. Rev. Mar. Sci.* **4**: 449-466.

Pinard R, de Winter A, Sarkis GJ, *et al.* (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* **7**: DOI: 10.1186/1471-2164-1187-1216.

Podar M, Abulencia CB, Walcher M, *et al.* (2007) Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Appl. Environ. Microbiol.* **73**: 3205-3214.

Raghunathan A, Ferguson HR, Bornarth CJ, Song WM, Driscoll M & Lasken RS (2005) Genomic DNA amplification from a single bacterium. *Appl. Environ. Microbiol.* **71**: 3342-3347.

Richardson TL & Jackson GA (2007) Small phytoplankton and carbon export from the surface ocean. *Science* **315**: 838-840.

Rodrigue S, Malmstrom RR, Berlin AM, Birren BW, Henn MR & Chisholm SW (2009) Whole genome amplification and *de novo* assembly of single bacterial cells. *PLoS One* **4**: e686.

Rusch DB, Halpern AL, Sutton G, *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* **5**: e77.

Sambrook J & Russel DW (2001) Hybridization of bacterial DNA on filters. *Molecular Cloning : A laboratory Manual*, Vol. Third ed.^eds.), p.^pp. 138-142. Cold Spring Harbor Labratory Press, New-York.

Scanlan DJ, Ostrowski M, Mazard S, *et al.* (2009) Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev* **73**: 249-299.

Schloss PD, Westcott SL, Ryabin T, *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**: 7537-7541.

Schoenfeld T, Liles M, Wommack KE, Polson SW, Godiska R & Mead D (2010) Functional viral metagenomics and the next generation of molecular tools. *Trends Microbiol.* **18**: 20-29.

Sekar R, Fuchs BM, Amann R & Pernthaler J (2004) Flow sorting of marine bacterioplankton after fluorescence in situ hybridization. *Appl. Environ. Microbiol.* **70**: 6210-6219.

Shannon CE & Weaver W (1949) *The mathematical theory of communication*. University of Illinois Press, Urbana, IL.

Six C, Thomas JC, Garczarek L, *et al.* (2007) Diversity and evolution of phycobilisomes in marine *Synechococcus* spp.: a comparative genomics study. *Genome Biol.* **8**: R259.

Sournia A & Birrien J-L (1995) La série océanographique côtière de Roscoff (Manche occidentale) de 1985 à 1992. *Cahiers de Biologie Marine* **36**: 1-8.

Stepanauskas R (2012) Single cell genomics: an individual look at microbes. *Curr. Opin. Biotechnol.* **15**: 613-620.

Stepanauskas R & Sieracki ME (2007) Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc. Natl. Acad. Sci. U. S. A* **104**: 9052-9057.

Suenaga H (2012) Targeted metagenomics: a high-resolution metagenomics approach for specific gene clusters in complex microbial communities. *Environ. Microbiol.* **14**: 13-22.

Swan BK, Martinez-Garcia M, Preston CM, *et al.* (2011) Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333**: 1296-1300.

Tai V & Palenik B (2009) Temporal variation of *Synechococcus* clades at a coastal Pacific Ocean monitoring site. *ISME J.* **3**: 903-915.

Tringe SG & Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. *Nature Rev. Gen.* **6**: 805-814.

Tripp HJ, Bench SR, Turk KA, *et al.* (2010) Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* **464**: 90-94.

Tyson GW, Chapman J, Hugenholtz P, *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37-43.

Vaulot D, Lepère C, Toulza E, *et al.* (2012) Metagenomes of the picoalga *Bathycoccus* from the Chile coastal upwelling. *PLoS One* **7**: e39648.

Wood AM, Phinney DA & Yentsch CS (1998) Water column transparency and the distribution of spectrally distinct forms of phycoerythrin-containing organisms. *Mar. Ecol.-Prog. Ser.* **162**: 25-31.

Wooley JC, Godzik A & Friedberg I (2010) A primer on metagenomics. *PLoS Comput. Biol.* **6**: e1000667.

Woyke T, Tighe D, Mavromatis K, *et al.* (2010) One bacterial cell, one complete genome. *PLoS One* **5**: e10314.

Woyke T, Xie G, Copeland A, *et al.* (2009) Assembling the marine metagenome, one cell at a time. *PLoS One* **4**: 10.

Yoon HS, Price DC, Stepanauskas R, *et al.* (2011) Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332**: 714-717.

Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW & Church GM (2006) Sequencing genomes from single cells by polymerase cloning. *Nature Biotech.* **24**: 680-686.

Zwirglmaier K, Jardillier L, Ostrowski M, *et al.* (2008) Global phylogeography of marine *Synechococcus* and *Prochlorococcus* reveals a distinct partitioning of lineages among oceanic biomes. *Environ. Microbiol.* **10**: 147-161.

Table legends

Table 1. List of primers used in this study.

Table 2. Description and statistical analyses of the pyrosequenced samples obtained during this study.

Figure legends

Fig. 1. Flowchart of the targeted metagenomic pipeline used in this study.

Fig. 2. T-RFLP analysis of the diversity of *Synechococcus* and co-occurring heterotrophic bacteria at station Astan as assessed at different stages of the metagenomic pipeline by targeting the 16S rRNA (A, C) and *petB* (B, D) genes. (A) Composite profile of 16S rRNA T-RFs. (B) T-RFLP profile targeting *petB* for the initial sample concentrated by TFF. (C, D) Histograms of relative T-RF abundance of the *Synechococcus* and/or bacterial community composition based on T-RFLP targeting 16S rRNA (C) or *petB* (D) genes. Mean and standard deviations were calculated on the 47 WGA reactions (out of 48) for which *Synechococcus* T-RFs were detected.

Fig. 3. Partial or complete PBS regions retrieved from a natural population from the Astan station off Roscoff (France). (A) Read coverage using as a reference the PBS region from *Synechococcus* sp. CC9902. (B) Contigs or supercontigs assembled from sequencing of 25 environmental PBS regions and mapped onto the CC9902 genome. Colors represent the clades/subclades of the strain giving the best BLASTX hit among the 42 *Synechococcus/Cyanobium* genomes of the reference database, whereas the shading degree represents the percentage of identity of the environmental sequence to the closest reference genome (cf. insert legend). Blue indicates a best match to clade IV (CC9902 or BL107), red to sub-clade Ia (CC9311 or WH8020) and orange to sub-clade Ib (MVIR18-1) strains. Highly conserved genes, *mpeBA*, *cpeBA* and *rpcBA* are shaded in grey. Environmental sequences from different contigs that exhibit more than 95% nucleotide identity are surrounded by blue frames.

Abbreviations: C, contigs obtained using the CLC-Assembly Cell; SC, supercontigs obtained by contig elongation using CAP3; misc., miscellaneous; PE, phycoerythrin; PC, phycocyanin.

Fig. 4. *Synechococcus* community structure determined at each step of the metagenomic pipeline by pyrosequencing targeting *petB* (A), *r/cpcBA* (B) and *cpeBA* (C) loci. OTUs were calculated using 94% and 95% similarity cut-off for *petB* and the two operons, respectively.

Fig. 5. Phylogenetic tree of *r/cpcBA* sequences from reference strains and environmental samples. The phylogenetic tree was built using Maximum Likelihood based on 340 aligned amino acids, 42 reference strains and the LG+G model. Environmental sequences retrieved from public databases and from this study were added by ARB_Parsimony to the initial consensus tree shown in bold lines. Bold and red letters correspond to sequences obtained by pyrosequencing and fosmid library sequencing, respectively. Asterisks indicate the three most abundant OTUs in clone libraries sequenced by pyrosequencing. *Gloeobacter violaceus* PCC 7421 was used as an outgroup. Only bootstrap values >70% are shown. Clades and/or pigment types were named following the nomenclature reported in previous studies (Crosbie *et al.*, 2003, Six *et al.*, 2007, Haverkamp *et al.*, 2008, Humily *et al.*, in press).

Fig. 6. Phylogenetic tree of *cpeBA* sequences from reference strains and environmental samples. The tree was built by Bayesian Inference analysis based on 999 nt positions from 42 marine *Synechococcus* isolates. Environmental sequences retrieved in public databases and from this study were added by ARB_Parsimony to the initial consensus tree shown in bold lines. Bold and red letters correspond to sequences obtained by pyrosequencing and fosmid library sequencing, respectively. Asterisks indicate the three most abundant OTUs in clone libraries sequenced by pyrosequencing. *Gloeobacter violaceus* PCC 7421 was used as an outgroup. Only bootstrap values >70% are shown. Clades and/or pigment types are named following the nomenclature reported in previous studies (Six *et al.*, 2007, Haverkamp *et al.*, 2008, Humily *et al.*, in press).

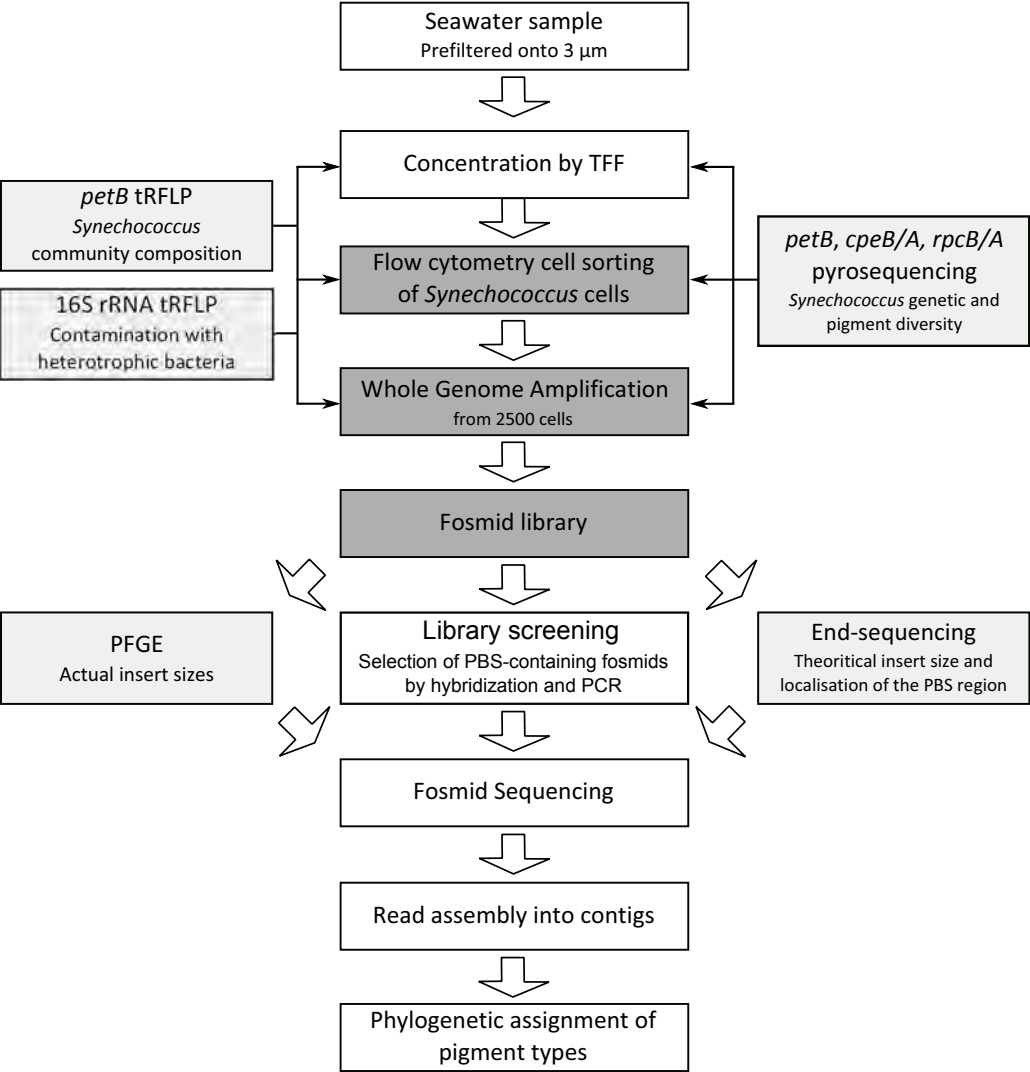
Table 1. List of primers used in this study.

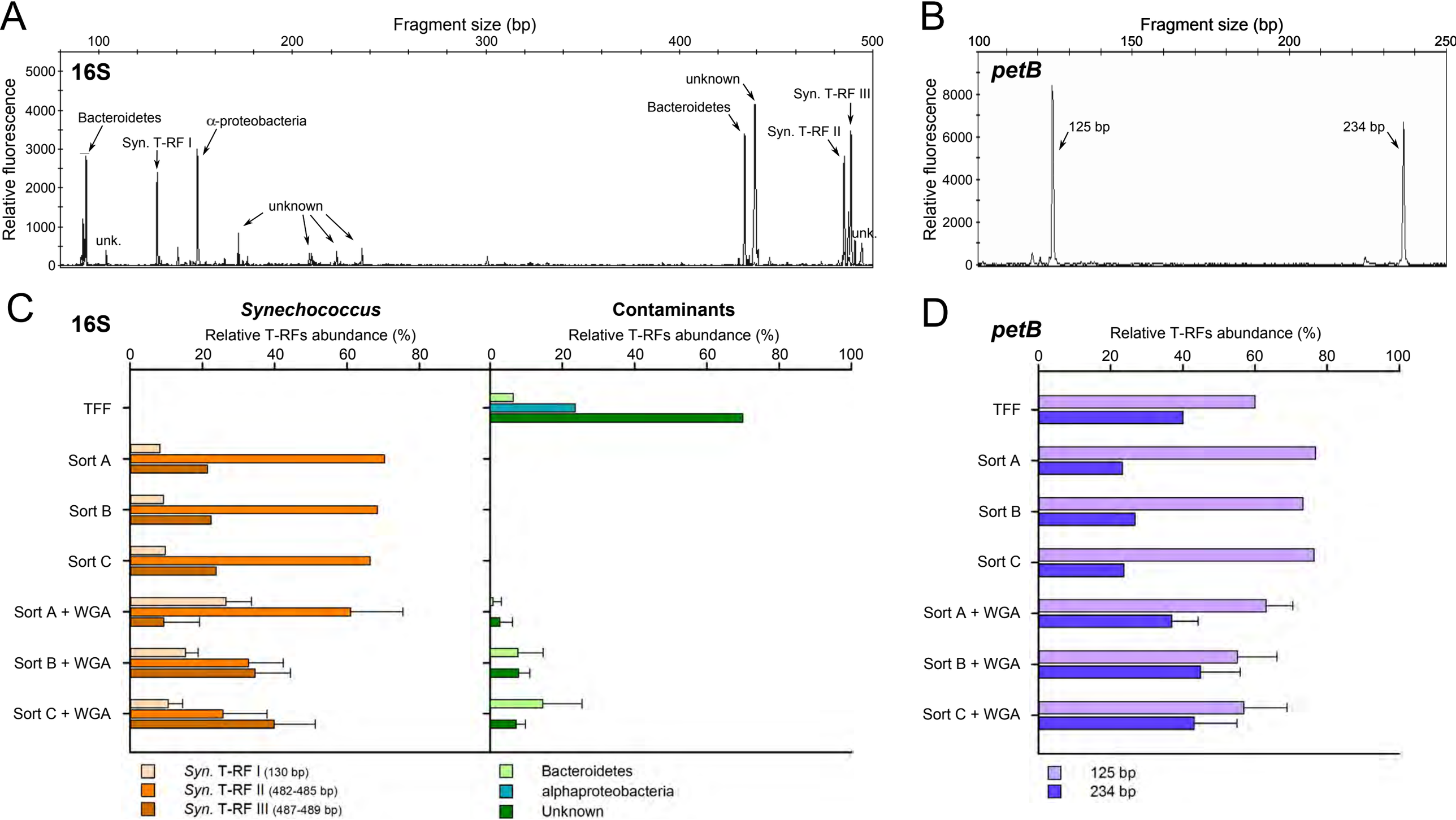
Primer name	Sequence (5' to 3')	Targeted gene/operon	Use	Product	Amplicon length[a]	Annealing Temp. (°C)	dNTPs (µM)	Primers conc. (nM)	# cycles	References
16S_27F[b]	AGAGTTTGATCMTGGCTCAG	16S rRNA	T-RFLP	16S rRNA Eubacteria	1448	55	100	400	30	(Field *et al.*, 1998)
16S_1492R	TACGGYTACCTTGTTACGACTT									(Lane, 1991)
petBF[b]	TACGACTGGTTCCAGGAACG	*petB*	T-RFLP/Pyrosequencing	b_6 subunit of cytochrome b_6/f	597	55	250	1 µM	30	(Mazard *et al.*, 2012)
petBR	GAAGTGCATGAGCATGAA								28-30[c]	(Mazard *et al.*, 2012)
mpeB_220F	TACACCAACCGCAARATGGC	*mpeBA*	Screening (hybridization/PCR)	PEII α- and β- subunits	626	55	300	500	30	This study
mpeA_263R	ACRAAGTCRCGCTTGCACTT									This study
SynB1F	ATGCTCGACGCATTCTCCMG	*cpeBA*	Pyrosequencing	PEI α- and β- subunits	635	55	250	500	30-35[c]	(Everroad & Wood, 2006)
SynA3R	CGGTGGTGAYGACGGAYTTCAT									(Everroad & Wood, 2006)
c/rpcB_244F	TGCCTGCGCGACATGGAGATC	*r/cpcBA*	Pyrosequencing	PC α- and β- subunits	523	50	250	500	30-40[c]	This study
SyncpcA-Rev	ATCTGGGTGGTGTAGGG		Screening (hybridization/PCR)						30	(Haverkamp *et al.*, 2008)
cpeC_187F	TTCAAGCTYGGTGARATCAG	*cpeC*	Screening (hybridization/PCR)	PE associated linker	645	55	300	500	35	This study
cpeC_832R	TGAAYTGCTCNGAGAGYTTG									This study
cpeU_224F	GATTTTGGTGGGARAGYAA	*cpeU*	Screening (PCR)	Putative phycobilin lyase	228	55	30	500	30	This study
cpeU_452R	ACAAACCARCAKCGYTCDAT									This study

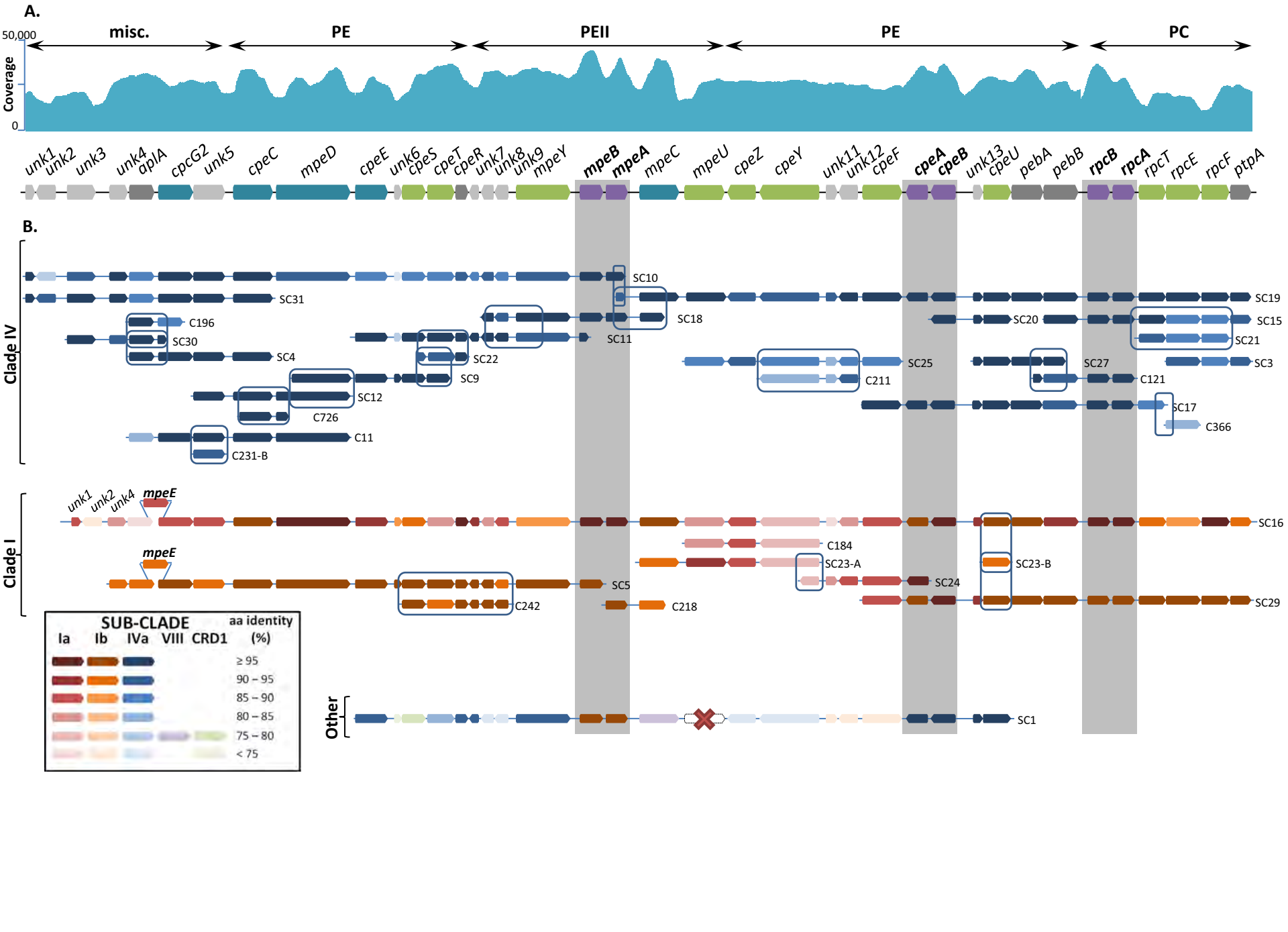
[a] Amplicon length was determined on *Synechococcus* sp. WH8102 (GenBank accession number BX548020).
[b] 16s rRNA and *petB* forward primers used for T-RFLP were 5'-labeled with 6-carboxyfluorescein (6-FAM)
[c] For pyrosequencing, the number of cycles was adjusted to obtain a faint band on an agarose gel

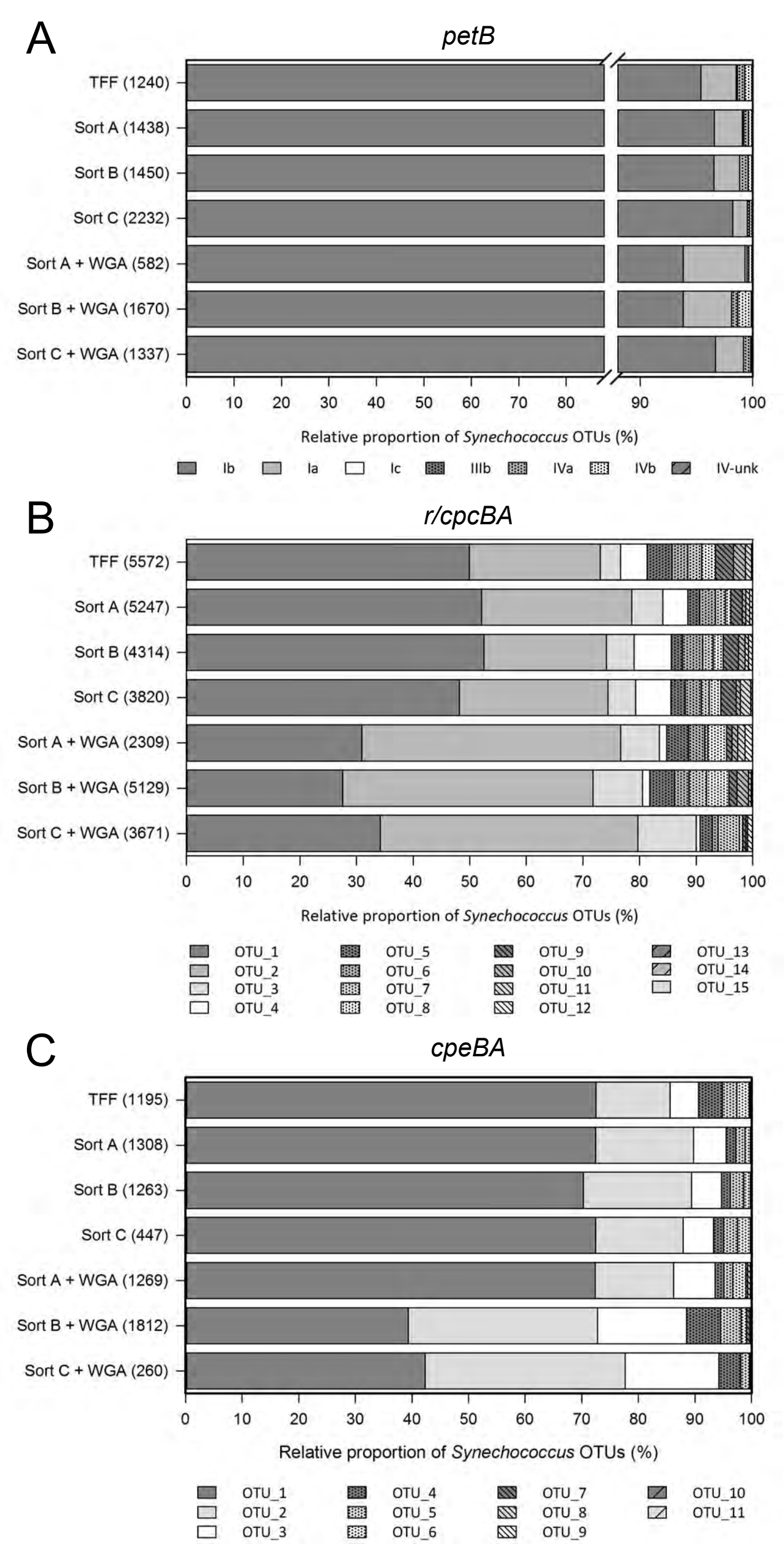
Table 2. Description and statistical analyses of the pyrosequenced samples obtained during this study.

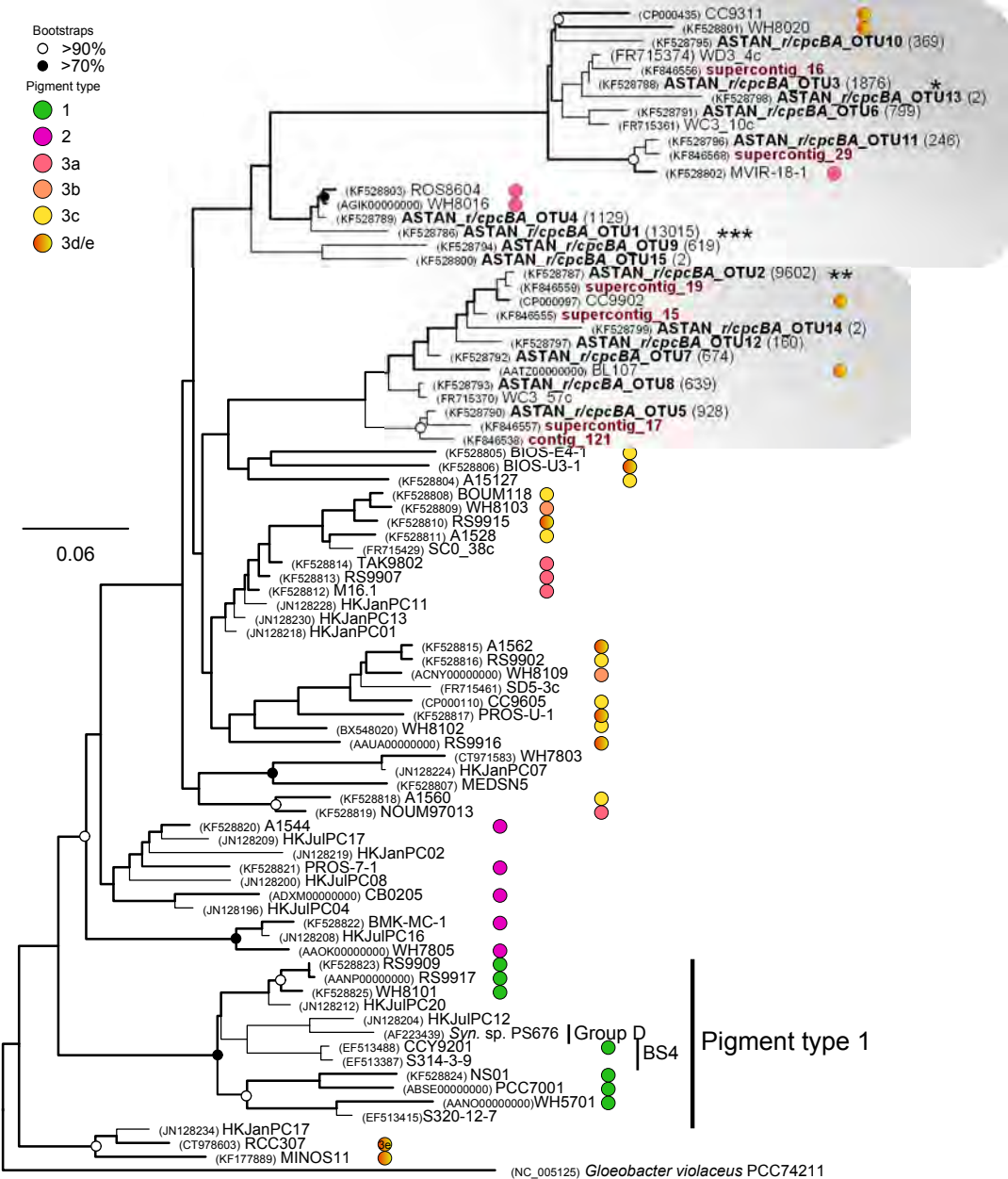
Gene/ Operon	Sample	# raw reads	# cleaned reads	% removed reads	# OTU	Good's coverage	S_{Chao-1}	Shannon index
petB	TFF	2119	1240	41	6	0.99	6.0	0.23
	Sort A	2116	1438	32	7	1.00	7.5	0.18
	Sort B	2335	1450	38	6	1.00	4.0	0.18
	Sort C	3545	2232	37	6	1.00	6.0	0.16
	Sort A + WGA	1061	582	45	5	1.00	5.0	0.20
	Sort B + WGA	3388	1670	51	7	0.99	6.0	0.28
	Sort C + WGA	2457	1337	46	5	1.00	4.0	0.16
r/cpcBA	TFF	6494	5572	14	13	0.99	14	1.55
	Sort A	6361	5247	18	12	1.00	11	1.39
	Sort B	5259	4314	18	13	1.00	13	1.50
	Sort C	4704	3820	19	13	0.99	12	1.53
	Sort A + WGA	2694	2309	14	12	1.00	11	1.51
	Sort B + WGA	6456	5129	21	13	0.99	26	1.52
	Sort C + WGA	4890	3671	25	13	0.99	12	1.24
cpeBA	TFF	6494	1195	82	8	1.00	9.0	0.98
	Sort A	6361	1308	79	7	1.00	7.0	0.89
	Sort B	5259	1263	76	8	0.99	7.0	0.85
	Sort C	4704	447	90	6	1.00	7.0	0.94
	Sort A + WGA	2694	1269	53	9	1.00	7.0	1.01
	Sort B + WGA	6456	1812	72	9	1.00	10	1.41
	Sort C + WGA	4890	260	95	6	0.99	6.0	1.24

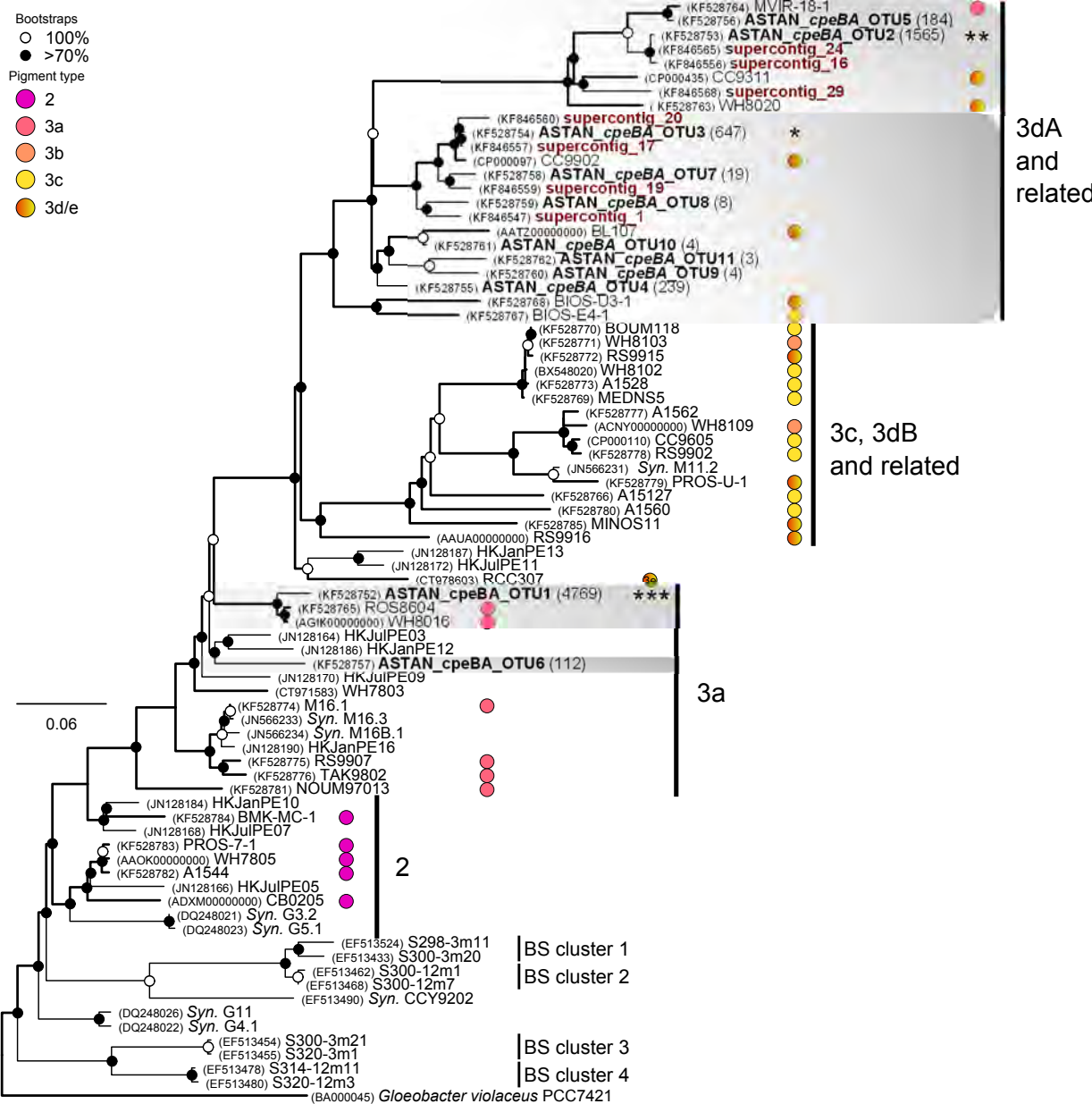

Seawater sample
Prefiltered onto 3 µm
Concentration by TFF
petB tRFLP
Synechococcus
community composition
Flow cytometry cell sorting
of Synechococcus cells
16S rRNA tRFLP
Contamination with
heterotrophic bacteria
petB, cpeB/A, rpcB/A
pyrosequencing
Synechococcus genetic and
pigment diversity
Whole Genome Amplification
from 2500 cells
Fosmid library
PFGE
Actual insert sizes
Library screening
Selection of PBS-containing fosmids
by hybridization and PCR
End-sequencing
Theoritical insert size and
localisation of the PBS region
Fosmid Sequencing
Read assembly into contigs
Phylogenetic assignment of
pigment types


A
16S
Fragment size (bp)
Relative fluorescence
Bacteroidetes
Syn. T-RF I
α-proteobacteria
unknown
unk.
unknown
Bacteroidetes
Syn. T-RF II
Syn. T-RF III
unk.
B
petB
Fragment size (bp)
Relative fluorescence
125 bp
234 bp
C
16S
Synechococcus
Contaminants
Relative T-RFs abundance (%)
TFF
Sort A
Sort B
Sort C
Sort A + WGA
Sort B + WGA
Sort C + WGA
Syn. T-RF I (130 bp)
Syn. T-RF II (482-485 bp)
Syn. T-RF III (487-489 bp)
Bacteroidetes
alphaproteobacteria
Unknown
D
petB
Relative T-RFs abundance (%)
TFF
Sort A
Sort B
Sort C
Sort A + WGA
Sort B + WGA
Sort C + WGA
125 bp
234 bp


A.
misc.
PE
PEII
PE
PC
50,000
Coverage
0
unk1 unk2 unk3 unk4 aplA cpcG2 unk5 cpeC mpeD cpeE unk6 cpeS cpeT cpeR unk7 unk8 unk9 mpeY mpeB mpeA mpeC mpeU cpeZ cpeY unk11 unk12 cpeF cpeA cpeB unk13 cpeU pebA pebB rpcB rpcA rpcT rpcE rpcF ptpA
B.
Clade IV
SC31
C196
SC30
SC4
SC12
C726
C231-B
C11
SC22
SC9
SC10
SC18
SC11
SC25
C211
SC19
SC20
SC15
SC21
SC27
C121
SC3
SC17
C366
Clade I
unk1 unk2 unk4 mpeE
mpeE
SC5
C242
C184
SC23-A
C218
SC24
SC23-B
SC16
SC29
Other
SC1
SUB-CLADE
Ia Ib IVa VIII CRD1
aa identity (%)
≥ 95
90 – 95
85 – 90
80 – 85
75 – 80
< 75


A
petB
TFF (1240)
Sort A (1438)
Sort B (1450)
Sort C (2232)
Sort A + WGA (582)
Sort B + WGA (1670)
Sort C + WGA (1337)
0 10 20 30 40 50 60 70 80 90 100
Relative proportion of Synechococcus OTUs (%)
Ib Ia Ic IIIb IVa IVb IV-unk
B
r/cpcBA
TFF (5572)
Sort A (5247)
Sort B (4314)
Sort C (3820)
Sort A + WGA (2309)
Sort B + WGA (5129)
Sort C + WGA (3671)
0 10 20 30 40 50 60 70 80 90 100
Relative proportion of Synechococcus OTUs (%)
OTU_1 OTU_2 OTU_3 OTU_4 OTU_5 OTU_6 OTU_7 OTU_8 OTU_9 OTU_10 OTU_11 OTU_12 OTU_13 OTU_14 OTU_15
C
cpeBA
TFF (1195)
Sort A (1308)
Sort B (1263)
Sort C (447)
Sort A + WGA (1269)
Sort B + WGA (1812)
Sort C + WGA (260)
0 10 20 30 40 50 60 70 80 90 100
Relative proportion of Synechococcus OTUs (%)
OTU_1 OTU_2 OTU_3 OTU_4 OTU_5 OTU_6 OTU_7 OTU_8 OTU_9 OTU_10 OTU_11


Bootstraps
>90%
>70%
Pigment type
1
2
3a
3b
3c
3d/e
0.06
(CP000435) CC9311
(KF528801) WH8020
(KF528795) ASTAN_r/cpcBA_OTU10 (369)
(FR715374) WD3_4c
(KF846556) supercontig_16
(KF528798) ASTAN_r/cpcBA_OTU3 (1876) *
(KF528798) ASTAN_r/cpcBA_OTU13 (2)
(KF528791) ASTAN_r/cpcBA_OTU6 (799)
(FR715361) WC3_10c
(KF528796) ASTAN_r/cpcBA_OTU11 (246)
(KF846568) supercontig_29
(KF528802) MVIR-18-1
(KF528803) ROS8604
(AGIK00000000) WH8016
(KF528789) ASTAN_r/cpcBA_OTU4 (1129)
(KF528786) ASTAN_r/cpcBA_OTU1 (13015) ***
(KF528794) ASTAN_r/cpcBA_OTU9 (619)
(KF528800) ASTAN_r/cpcBA_OTU15 (2)
(KF528787) ASTAN_r/cpcBA_OTU2 (9602) **
(KF846559) supercontig_19
(CP000097) CC9902
(KF846555) supercontig_15
(KF528799) ASTAN_r/cpcBA_OTU14 (2)
(KF528797) ASTAN_r/cpcBA_OTU12 (160)
(KF528792) ASTAN_r/cpcBA_OTU7 (674)
(AATZ00000000) BL107
(KF528793) ASTAN_r/cpcBA_OTU8 (639)
(FR715370) WC3_57c
(KF528790) ASTAN_r/cpcBA_OTU5 (928)
(KF846557) supercontig_17
(KF846538) contig_121
(KF528805) BIOS-E4-1
(KF528806) BIOS-U3-1
(KF528804) A15127
(KF528808) BOUM118
(KF528809) WH8103
(KF528810) RS9915
(KF528811) A1528
(FR715429) SC0_38c
(KF528814) TAK9802
(KF528813) RS9907
(KF528812) M16.1
(JN128228) HKJanPC11
(JN128230) HKJanPC13
(JN128218) HKJanPC01
(KF528815) A1562
(KF528816) RS9902
(ACNY00000000) WH8109
(FR715461) SD5-3c
(CP000110) CC9605
(KF528817) PROS-U-1
(BX548020) WH8102
(AAUA00000000) RS9916
(CT971583) WH7803
(JN128224) HKJanPC07
(KF528807) MEDSN5
(KF528818) A1560
(KF528819) NOUM97013
(KF528820) A1544
(JN128209) HKJulPC17
(JN128219) HKJanPC02
(KF528821) PROS-7-1
(JN128200) HKJulPC08
(ADXM00000000) CB0205
(JN128196) HKJulPC04
(KF528822) BMK-MC-1
(JN128208) HKJulPC16
(AAOK00000000) WH7805
(KF528823) RS9909
(AANP00000000) RS9917
(KF528825) WH8101
(JN128212) HKJulPC20
(JN128204) HKJulPC12
(AF223439) Syn. sp. PS676
(EF513488) CCY9201
(EF513387) S314-3-9
(KF528824) NS01
(ABSE00000000) PCC7001
(AANO00000000) WH5701
(EF513415) S320-12-7
Group D
BS4
Pigment type 1
(JN128234) HKJanPC17
(CT978603) RCC307
(KF177889) MINOS11
3e
(NC_005125) Gloeobacter violaceus PCC74211


Bootstraps
100%
>70%
Pigment type
2
3a
3b
3c
3d/e
(KF528764) MVIR-18-1
(KF528756) ASTAN_cpeBA_OTU5 (184)
(KF528753) ASTAN_cpeBA_OTU2 (1565)
**
(KF846565) supercontig_24
(KF846556) supercontig_16
(CP000435) CC9311
(KF846568) supercontig_29
(KF528763) WH8020
(KF846560) supercontig_20
(KF528754) ASTAN_cpeBA_OTU3 (647)
*
(KF846557) supercontig_17
(CP000097) CC9902
(KF528758) ASTAN_cpeBA_OTU7 (19)
(KF846559) supercontig_19
(KF528759) ASTAN_cpeBA_OTU8 (8)
(KF846547) supercontig_1
(AATZ00000000) BL107
(KF528761) ASTAN_cpeBA_OTU10 (4)
(KF528762) ASTAN_cpeBA_OTU11 (3)
(KF528760) ASTAN_cpeBA_OTU9 (4)
(KF528755) ASTAN_cpeBA_OTU4 (239)
(KF528768) BIOS-D3-1
(KF528767) BIOS-E4-1
3dA and related
(KF528770) BOUM118
(KF528771) WH8103
(KF528772) RS9915
(BX548020) WH8102
(KF528773) A1528
(KF528769) MEDNS5
(KF528777) A1562
(ACNY00000000) WH8109
(CP000110) CC9605
(KF528778) RS9902
(JN566231) Syn. M11.2
(KF528779) PROS-U-1
(KF528766) A15127
(KF528780) A1560
(KF528785) MINOS11
3c, 3dB and related
(AAUA00000000) RS9916
(JN128187) HKJanPE13
(JN128172) HKJulPE11
(CT978603) RCC307
(KF528752) ASTAN_cpeBA_OTU1 (4769)
***
(KF528765) ROS8604
(AGIK00000000) WH8016
(JN128164) HKJulPE03
(JN128186) HKJanPE12
(KF528757) ASTAN_cpeBA_OTU6 (112)
(JN128170) HKJulPE09
(CT971583) WH7803
3a
0.06
(KF528774) M16.1
(JN566233) Syn. M16.3
(JN566234) Syn. M16B.1
(JN128190) HKJanPE16
(KF528775) RS9907
(KF528776) TAK9802
(KF528781) NOUM97013
(JN128184) HKJanPE10
(KF528784) BMK-MC-1
(JN128168) HKJulPE07
(KF528783) PROS-7-1
(AAOK00000000) WH7805
(KF528782) A1544
(JN128166) HKJulPE05
(ADXM00000000) CB0205
2
(DQ248021) Syn. G3.2
(DQ248023) Syn. G5.1
(EF513524) S298-3m11
(EF513433) S300-3m20
BS cluster 1
(EF513462) S300-12m1
(EF513468) S300-12m7
BS cluster 2
(EF513490) Syn. CCY9202
(DQ248026) Syn. G11
(DQ248022) Syn. G4.1
(EF513454) S300-3m21
(EF513455) S320-3m1
BS cluster 3
(EF513478) S314-12m11
(EF513480) S320-12m3
BS cluster 4
(BA000045) Gloeobacter violaceus PCC7421

Supplementary information

Table S1. Characteristics of the 16S rRNA T-RFs found at different steps of the metagenomic pipeline. To identify unknown T-RFs, 16S rRNA PCR products exhibiting the highest abundance of this unknown peak in T-RFLP profiles were selected and sequenced before *in silico* digestion by *MspI*.

Gene	Restriction enzyme	*in silico* T-RFs (bp)	Experimental T-RFs (bp)	Phylogenetic assignment	Closest species
16S rRNA	*MspI*	90-91	89-93	*Bacteroidetes/Cytophaga*	*Microscilla* sp., *Marinoscillum furvescens*
			104	n.d	
			123	n.d	
		129	130	cyanobacteria	*Synechococcus* ALMO3 (Ib)
		150	151	α-*proteobacteria*	*Candidatus Pelagibacter ubique*
			172	n.d.	
			210	n.d.	
			221	n.d	
			234	n.d	
		435	434	*Bacteroidetes/*shingobacteria	*Mucilaginibacter rigui*
			439	n.d.	
		461	456	γ-proteobacteria	*Halomonas* sp.
			482	n.d	
		491-492	485-487	cyanobacteria	*Synechococcus*
		491-492	487-489	cyanobacteria	*Synechococcus*
			491	n.d.	
			494	n.d.	

Table S2. Characteristics of the *petB* T-RFs generated *in silico* using a database of cultured and environmental *Synechococcus* sequences and experimentally detected at the Astan station.

Gene	Restriction enzymes used	Reference *Synechococcus/Cyanobium* isolates	Clade	*In silico* T-RF (bp)	Experimental T-RF (bp)
petB	*HinfI, HpaII*	BIOS-S15-1/MITS9220	CRD1a	101	99-101
		Biosope_45 C4Y/BIOS-H3-2	CRD1b		
		PROSOPE_107/CC9311/WH8020	Ia	123	125
		ALMO3/SYN20/WH8016	Ib		
		Biosope_141-D	Ic		
		A1511/RS9905/WH8102/WH8103	IIIa		
		A1528	IIIb		
		BL107	IVa		
		A15-37/RS9919	IIa	137	n.d.
		PROS-U-2	IIh		
		A15-144	WPC1		
		CB0101	CB4		
		CB0205	CB5		
		RS9901/RS9916/RS9921	X	140	137-139
		WH5701	Sub 5.2		
		RS9911	IIa	231	231-236
		A15-146/A5-19	IIb		
		CC9605	IIc		
		PROS-3-1	IIe		
		A15-72	IIf		
		CC9902	IVa		
		WH8101	VIII		
		PROS-7-2/PROS-3-2/M21B.3	XVI		
		RCC307	Sub 5.3-I		
		A15-43/A15-46/A15-48A1/WH7803	V	279	279-283
		WH8109/RS9904	IIa	308	307-308
		WH7805/BL8	VIa		
		PROSOPE_130	VIc		
		RS9906/RS9917	VII		
		NS01/PCC 7001	*Cyanobium*		
		A15-147	IIh	347	347
		A15-60/A15-74	VII		
		MINOS11	Sub 5.3-I		

Table S3. Characteristics of the fosmid library obtained at the Astan station

Step	Parameter	Value
Fosmid library	Total number of clones	506,296
	Number of screened clones	55,296
Screening by hybridization	Number of positive clones	768
Screening by PCR	Number of positive clones	268
	at least 1 gene	94 %[a]
	at least 2 genes	58 %[a]
	at least 3 genes	19 %[a]
	4 genes	4 %[a]
Validations	Number of clones screened by end-sequencing	80
	Sequences assigned to *Synechococcus*	93 %
	Average insert size (bp) (*n*=25)	35,386
Sequencing/Assembly	Number of fosmids sequenced	25
	Total Nb Contigs	64
	Maximum size	30,279 bp
	Average size	4,315 bp
	N50	7,804 bp
	N90	2,137 bp
Final selection of contigs	Total Nb Contigs	34
	Maximum size	29,890 bp
	Average size	5,033 bp
	N50	11,113 bp
	N90	2,305 bp

[a] This percentage is relative to the number of positive clones as determined by hybridization

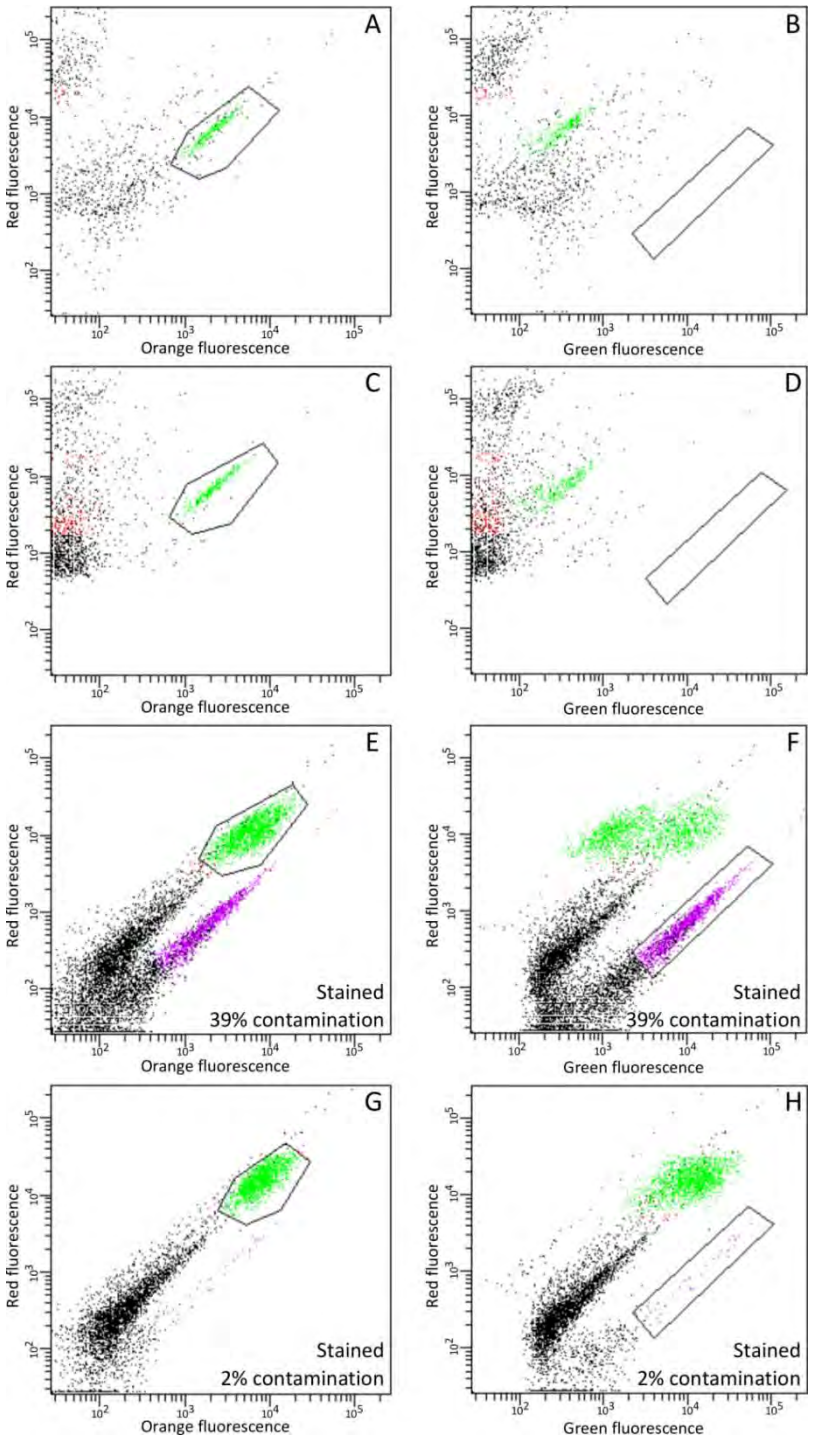


Fig. S1. Flow cytometry analysis of the picophytoplankton community from the Astan station. (A,B) Initial unconcentrated sample (seawater); (C, D) after tangential flow filtration; (E, F) after 1 step of cell sorting; (G, H) after 3 steps of cell sorting. The red, orange and green florescence signals are proxy for chlorophyll, phycoerythrin and DNA content, respectively. The population indicated in green corresponds to *Synechococcus* cells, while co-occurring heterotrophic bacteria are shown in purple.

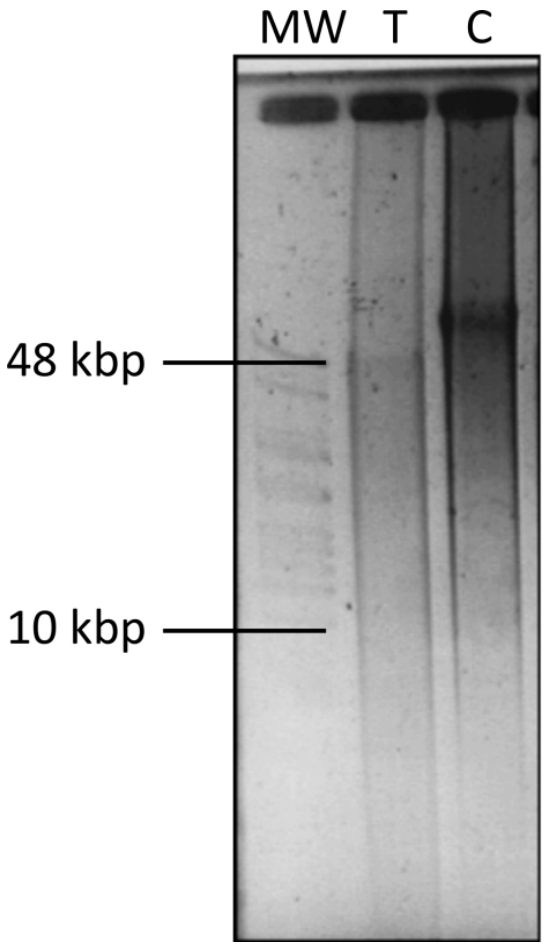


Fig. S2. Pulse Field Gel Electrophoresis analyses of WGA products from a sorted sample from the Astan station, amplified using Genomiphi *v2* kit and denaturated either thermally (T, cells were boiled for 3 min at 95 °C) or chemically (C, alkaline denaturation and neutralization). MW (Molecular weight marker): GeneRuler^{TM} High Range DNA Ladder (Fermentas, Villebon sur Yvette, France).

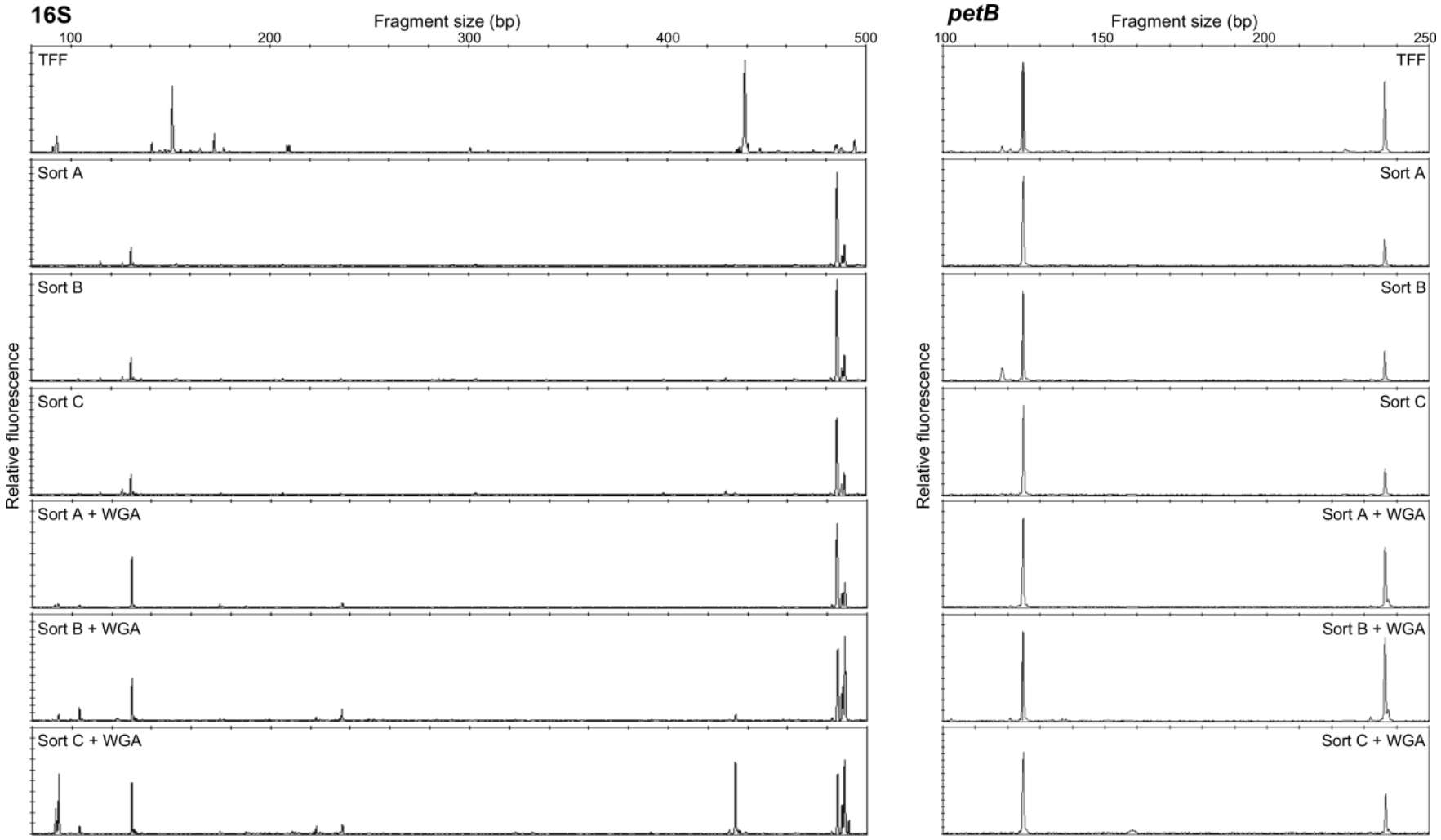


Fig. S3. Representative T-RFLP profiles at each step of the metagenomic pipeline. TFF: tangential flow filtration, WGA: whole genome amplification.

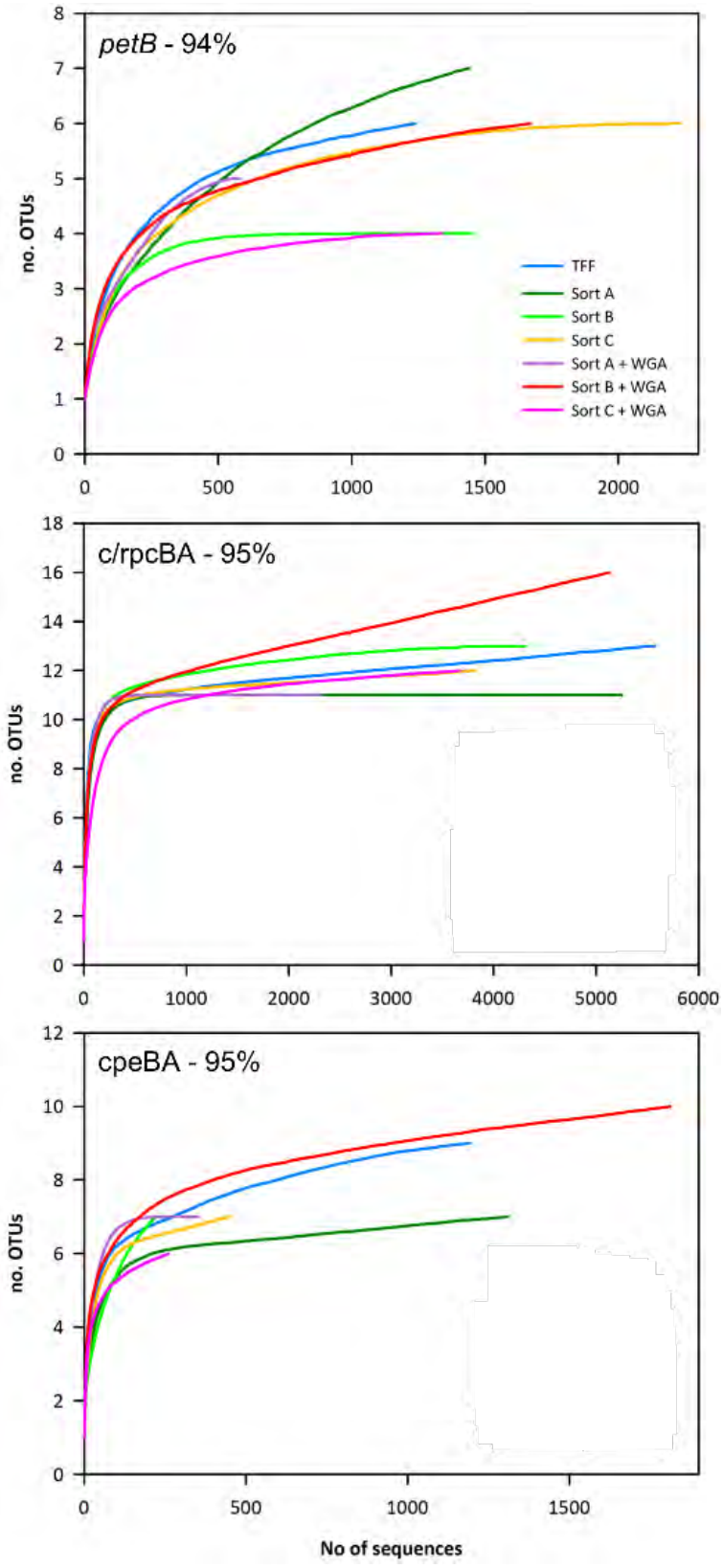


Fig. S4. Rarefaction curves of the number of OTUs for each sample at 94% and 95% similarity cut-off for the *petB, c/rpcBA* and *cpeBA* loci, respectively.