Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: http://www.elsevier.com/locate/syapm

Ecogenomic characterization of widespread, closely-related SAR11 clades of the freshwater genus "*Candidatus* Fonsibacter" and proposal of *Ca.* Fonsibacter lacus sp. nov

Despina Tsementzi[a,1], Luis M. Rodriguez-R[a,1], Carlos A. Ruiz-Perez[b], Alexandra Meziti[a], Janet K. Hatt[a], Konstantinos T. Konstantinidis[a,b,*]

[a] *School of Civil and Environmental Engineering, Georgia Institute of Technology, Ford Environmental Science & Technology Building, 311 Ferst Drive, Atlanta, GA 30332, United States*

[b] *School of Biological Sciences, Georgia Institute of Technology, Ford Environmental Sciences & Technology Building, 311 Ferst Drive, Atlanta, GA 30332, United States*

ARTICLE INFO

Article history:
Received 21 January 2019
Received in revised form 29 March 2019
Accepted 29 March 2019

Keywords:
Fonsibacter
Freshwater SAR11
Metagenome assembled genomes
Lake Lanier
Chattahoochee basin
LD12

ABSTRACT

The ubiquitous alpha-proteobacteria of the order "*Candidatus* Pelagibacterales" (SAR11) are highly abundant in aquatic environments, and among them, members of the monophyletic lineage LD12 (also known as SAR11 clade IIIb) are specifically found in lacustrine ecosystems. Clade IIIb bacteria are some of the most prominent members of freshwater environments, but little is known about their biology due to the lack of genome representatives. Only recently, the first non-marine isolate was cultured and described as "*Candidatus* Fonsibacter ubiquis". Here, we expand the collection of freshwater IIIb representatives and describe a new IIIb species of the genus "*Ca.* Fonsibacter". Specifically, we assembled a collection of 67 freshwater metagenomic datasets from the interconnected lakes of the Chattahoochee River basin (GA, USA) and obtained nearly complete metagenome-assembled genomes (MAGs) representing 5 distinct IIIb subclades, roughly equivalent to species based on genomic standards, including the previously described "*Ca.* F. ubiquis". Genomic comparisons between members of the IIIb species revealed high similarity in gene content. However, when comparing their abundance profiles in the Chattahoochee basin and various aquatic environments, differences in temporal and spatial distributions among the distinct species were observed implying niche differentiation might be underlying the coexistence of the highly functionally similar representatives. The name *Ca.* Fonsibacter lacus sp. nov. is proposed for the most abundant and widespread species in the Chattahoochee River basin and various freshwater ecosystems.

Introduction

Bacteria of the order "*Ca.* Pelagibacterales", commonly referred to as SAR11, comprise the most abundant microbial group in oceans with an estimated population size of 10^{28} cells [30]. Members of the marine SAR11 belong to diverse bacterial lineages (or clades) and can occupy different environmental niches exhibiting distinct abundance profiles in correlation with various environmental parameters, such as, depth in the water column, salinity, and seasonality [1,11,59,60]. SAR11 clade IIIb (also known as LD12) represents the only lineage with members strictly adapted to freshwater environments and not found in marine systems [29]. Representatives from other lineages are also found in lacustrine environments. For example, SAR11 metagenomic assembled genomes (MAGs) from other (non-IIIb) clades have been recovered from Lake Baikal, Siberia, Russia [2], Lake Qinghai, Qinghai, China [34], and estuarine or brackish ecosystems with intermediate salinity [20]. However, unlike these other clades, SAR11 bacteria of the IIIb clade are globally and exclusively distributed in freshwater environments [32], where they are among the most dominant members, making up to 20% of the lacustrine bacterial communities [16,49].

Our current knowledge on the metabolic adaptations of freshwater SAR11 bacteria mostly stems from a handful of studies employing culture-independent techniques. Tracer isotope experiments have confirmed that the clade IIIb members are aerobic

[*] Corresponding author at: School of Civil and Environmental Engineering, and School of Biology, Georgia Institute of Technology, 311 Ferst Dr., Atlanta, GA 30332-0512, United States.
E-mail address: kostas@ce.gatech.edu (K.T. Konstantinidis).
[1] Equal contribution authors.

https://doi.org/10.1016/j.syapm.2019.03.007
0723-2020/© 2019 Elsevier GmbH. All rights reserved.

heterotrophic bacteria like the majority of their seawater counterparts [49]. The limited number of freshwater SAR11 genomes publicly available up to date (n = 13) have been recovered from only a few studies employing culture-independent techniques, including metagenomic genome binning [8,12] and single-cell genomics [11,64]. The only available isolate representing clade IIIb was recently obtained from Lake Borgne, LA, USA, a coastal lagoon of the Gulf of Mexico [17]. Genomic and experimental data from these previous studies such as, the acquisition of the Embden–Meyerhof–Parnas glycolysis pathway (EPM), the presence of a complete glyoxylate shunt, differentiated C1 metabolism, and a general trend towards de novo synthesis rather than uptake of many important amino acids, osmolytes, and other compounds in most marine clades (reviewed in Ref. [11]) have provided critical insights into the genomic adaptations of clade IIIb representatives compared to their marine relatives. However, due to the limited number of available genomes, little is known about the universality of the functional traits among different freshwater ecosystems and the genomic diversity between freshwater SAR11 populations.

Here, we expand the collection of SAR11 clade IIIb MAGs, and describe the functional differences and geographic distributions of five distinct subclades of clade IIIb (genomospecies). We leveraged a large collection of freshwater datasets obtained from interconnected lakes along the Chattahoochee River in the Southeast USA and a newly developed iterative assembling and binning approach resulting in the identification of 67 "*Ca.* Pelagibacterales" partial MAGs. We describe commonalities and differences in gene content among the subclades represented by these MAGs and the spatial and temporal dynamics that differentiate them from "*Ca.* Fonsibacter ubiquis".

Materials and methods

Sample acquisition and processing

Water samples (20 L) were collected using a horizontal sampler (Wildco Instruments) from the epilimnion (typically 3–5 m depth) of Lakes Lanier (GA), West Point (GA/AL), Harding (GA/AL), Eufaula (GA/AL), and Seminole (GA/FL), and two locations in the Apalachicola estuary off the coasts of Apalachicola Bay and East Point (Florida, USA). The northernmost lake closest to the river basin source, Lake Lanier, has been sampled nearly bimonthly for six years, while the other locations were sampled for five years during late August or early September. The final sample collection consists of 67 samples (Fig. S1), including 36 samples from Lake Lanier and 31 samples from the downstream lakes and estuaries. All samples were immediately stored at 4 °C and processed typically within 1–4 h, and no more than a day post-collection. Water was sequentially filtered with a peristaltic pump through 2.0 AP20 Glass fiber filters (Millipore) μm and 1.6 μm GF/A filters (Whatman), to exclude large particles and eukaryotic cells, and cells were captured on 0.2 μm Sterivex filters (Millipore). Thus, all sequenced metagenomes represent the 1.6−0.2 μm cell size fraction. Filters were preserved at −80 °C until used for extraction.

DNA extraction and sequencing

DNA extraction was performed as previously described [7] with minor modifications. Briefly, frozen filters were placed in microcentrifuge tubes with lysis buffer (50 mM Tris−HCl pH 8.3, 40 mM EDTA, and 0.75 M sucrose) and 1 mg/ml lysozyme and incubated at 37 °C for 30 min. Reactions were subsequently incubated with 1% SDS, 10 mg/ml proteinase K, and 150 μg/ml RNase for 4 h at 55 °C in a rotating hybridization oven. DNA was isolated from lysate using phenol and chloroform extraction, precipitated with ethanol, and eluted in Tris-EDTA (TE) buffer. On average, DNA yield was 1.7 μg per liter of water filtered. One nanogram of DNA was used to prepare sequencing libraries using the Nextera XT Kit described by the manufacturer (Illumina). Equimolar concentrations of sample libraries were sequenced using Illumina GA II, HiSeq, or MiSeq following manufacturer's protocols (Supplementary Table S1).

Quality control of metagenomic datasets

All sequenced metagenomic datasets were subjected to quality control and those not passing minimum requirements (see below) were re-sequenced (n = 3). Sequencing reads were trimmed using SolexaQA++ [6] with minimum PHRED quality score of 20 and minimum fragment length of 50 bp, and clipped to remove residual sequencing adaptor contamination (if any) using Scythe (https://github.com/vsbuffalo/scythe). Read redundancy analysis using Nonpareil (v. 2.4) [43] was applied to estimate abundance weighted average coverage of each metagenome, i.e., the percentage of total extracted community DNA that was sequenced for each metagenomic dataset. A minimum 50% achieved coverage estimated with Nonpareil and 1Gb total size were required for all datasets in this study; otherwise, samples were re-sequenced.

Recovery of metagenomic assembled genomes (MAGs)

MAGs were recovered by applying an iterative binning methodology. Briefly, an initial set of MAGs (termed LLD) was obtained using only 27 metagenomic datasets derived from Lake Lanier (2009–2014) and a single co-assembly generated by IDBA-UD [38] and MetaBAT for the population genome binning step with default settings [23]. Next, metagenomic datasets from all samples were grouped based on MASH distances of whole metagenomes (MASH v1.0.2, sketch size 10,000) [35] using the Markov Clustering algorithm (MCL v14-137, MASH distances < 0.07) [13] as implemented in ogs.mcl.rb from the enveomics collection (Inflation of 1.5) [44]. The obtained groups of metagenomes were next co-assembled and, in some cases for groups over 20 Gbp in total size, also subsampled at various sizes and then co-assembled. Assemblies were generated with IDBA-UD (IDBA v1.1.1, default parameters) [38]. All assemblies were independently binned using MaxBin (v2.1.1, default parameters; MaxBin was used to complement MataBAT from the first binning step) [63], and the resulting MAGs (termed WB) were evaluated using MiGA (v0.3.1.4) [42] and CheckM [37] to estimate genome completeness and contamination. High-quality genomes were defined as those with genome quality (completeness − 5 × contamination) above 50. All high-quality bins were combined in a single collection. Each metagenomic dataset was mapped to this collection using Bowtie2 (v2.3.2, default parameters) [27] and unmapped reads were extracted using Samtools (v1.0) [28]. The entire process was repeated on unmapped reads iteratively until <1% gain in mapped reads was reached (9 iterations).

Identification and phylogenetic characterization of SAR11 MAGs

MAGs that were classified within the "*Ca.* Pelagibacterales" order using the MiGA taxonomic classifier [42] against NCBI_Prok were selected for further phylogenetic reconstructions and additional quality control using CheckM (v1.0.3) [37] (Table S2). All MAGs as well as selected SAR11 reference genomes (Table S3) were processed using MiGA, including universal gene detection. Briefly, genes were predicted using Prodigal (v2.6.1, single-genome mode) [21] and universal marker genes, found in single copy in almost all bacterial genomes (n = 106), were identified using Hidden Markov models and HMMER3 (v3.1b1; http://hmmer.janelia.org/) with default settings and the recommended cutoff [9], as implemented in the script HMM.essential.rb from the enveomics collection. Alignments for each of the 106 identified single copy genes were constructed with Clustal Omega (v1.2.1) [52] and concatenated using the script Aln.cat.rb from the enveomics collection, which removes the invariable sites and maintains protein coordinates. Maximum likelihood phylogeny was built from the concatenated alignment using RAxML, with 1,000 bootstraps and the PROTGAMMAAUTO function which identifies the best amino acid substitution model for each protein. MAGs were subsequently assigned to SAR11 subclades in accordance with the previously published subclade nomenclature [56,59].

Average Nucleotide Identity (ANI) for all SAR11 MAG pairs was estimated as described previously [25] using the ani.rb script from the enveomics collection. Species-level clusters were identified by MiGA using the Markov Clustering algorithm (MCL) as implemented in ogs.mcl.rb from the enveomics collection in order to identify clusters of genomes with ANI ≥ 95%, hereafter termed "PEL" genomospecies. It has previously been shown that a value of 95% ANI is a good proxy to group genomes into the same species, with an accuracy of 99.8% when applied to genomes of named bacterial species [22,45]. Genome representatives from each identified PEL species were selected based on the highest genome quality score (completeness − 5 × contamination) for subsequent analysis. Average Amino Acid Identities (AAI) were estimated using the aai.rb script from the enveomics collection for each PEL genome and selected genome representatives from each of the SAR11 clades. The resulting AAI matrix was used to confirm the SAR11 clade classification that was obtained based on the phylogenetic reconstructions.

16S ribosomal RNA gene identification, phylogeny, and classification

16S ribosomal RNA (rRNA) sequences were identified in the MAGs using RNAmmer (v. 1.2) [26] and their taxonomy was confirmed using the RDP classifier [61]. Recovered and previously reported SAR11 representative sequences [17,59] were aligned using the SINA aligner server (v. 1.2.11) [40]. The 16S rRNA gene phylogenetic tree was reconstructed using RAxML (v. 8.2.11) with a GTRGAMMAI model and 1000 bootstraps [54].

Assessing spatial and temporal abundance distributions

The abundance of each genomospecies in the Chattahoochee datasets (Table S1) and representative aquatic metagenomic datasets (Table S4) was estimated based on competitive read mapping. First, a non-redundant database of representative genomes was constructed which included all MAGs identified from the Chattahoochee River datasets, as well as publically available SAR11 genomes representing all known clades (Table S3). The genome database was clustered with ANI > 95% (to identify genomospecies), and only one representative genome per cluster was retained to eliminate redundancy. When no isolate genome was available for a genomospecies, the representative genome was chosen based on the best genome quality of available MAGs or SAGs. Quality-trimmed reads were mapped against the non-redundant genome database using Bowtie2 (v2.3.2, default parameters) [27]. For each metagenomic dataset, the sequencing depth of all contigs was estimated per base with bedtools genomecov (v2.25, using – bga) [41]. The 80% central truncated average of the sequencing depth of all bases on each genome (truncated average sequencing depth, TAD) was then estimated using BedGraph.tad.rb from the enveomics collection. The TAD is estimated by removing the top-10% and bottom-10% positions of per base sequencing coverage in each genome in order to correct for expected sequencing depth biases from highly conserved regions or contig edges and intra-species gene content diversity, respectively. The corrected sequencing depth was finally used to calculate the relative abundance of each representative genome (equivalent to genomospecies) over the total bacterial fraction in each metagenome by dividing the corrected sequencing depth by the sequencing depth of the *rpoB* gene. The *rpoB* sequencing depth was estimated using ROCker (v1.1.12, default parameters, model generated on June/2015) which includes a manually curated *rpoB* database and bit-score thresholds per gene position for read mapping [36]. Finally, the presence of a genome within a given metagenomic dataset was identified as any non-zero TAD values, which equates to a minimum of 10% sequencing breadth coverage for each genome, a threshold that has been previously described to accurately indicate presence of a species in a metagenome [5].

Seasonality of SAR11 genomospecies was evaluated on 2 collections of time series datasets: 32 metagenomes from Lake Lanier and 92 metagenomes from Lake Mendota, representing five and four-year collections, respectively. Seasonality was evaluated for each representative SAR11 genome using smoothing cubic splines of log-transformed abundance by sampling date (ignoring year). In order to implement the cyclic nature of the data into the spline analysis, all the abundances were used in triplicate with sampling dates in tandem; i.e., with sampling times in radians in the ranges [-2π, 0], [0, 2π], and [2π, 4π]. Only the central fragment of the splines was used in the range [0, 2π]. Seasonal genomospecies were considered those species with Pearson's correlation index between spline-derived abundance (predicted abundance) and observed abundance above 0.5 (p-value 0.0016).

Functional annotations and genome comparisons

Genes were predicted in all analyzed genomes using Prodigal (v2.6.1, default parameters)

[21] within the MIGA pipeline [42]. Functional annotations for each genome were performed using blastp searches (score >60 bits, similarity >40%) against UniProt [62] and against Cluster of Orthologous Groups of proteins (COGs) using the eggnog-mapper and the eggnog database version 4.5 [19]. In order to identify unique and common gene content between PEL1 and PEL2 ("*Ca.* Fonsibacter ubiquis") genomospecies, a core gene set within each genomospecies was first identified by analyzing all MAGs of the genomospecies (n = 17 for PEL1 and n = 16 for PEL2). Due to the partial nature of the MAGs and in order to eliminate potential contamination biases, the species core genes were defined as those genes that were found in at least two of the MAGs of each species. The orthologous genes among MAGs were identified by pairwise reciprocal best matches (RBM) and grouped with the Markov Clustering algorithm (MCL v14-137) [13] using bit-score as RBM weights as implemented in the rbm.rb and ogs.mcl.rb scripts from the enveomics collection [44]. Genes present in the isolate genome of "*Ca.* F. ubiquis" were considered part of the consensus genome of the PEL2 genomospecies. Genes of potential phage origin were identified by evaluating each MAG contig with Virsorter [48].

Fluorescence in situ hybridization

Samples for FISH were collected and preserved using previously described protocols [14,49,51]. Briefly, water samples from Lake Lanier collected in early December 2018 were selected for FISH, as the metagenomic profiles indicated increased abundance of SAR11 cells during winter. Samples were immediately fixed in the field with formaldehyde at 2% final concentration and stored at 4 °C for 24 h. Aliquots of 3 mL were vacuum filtered onto 0.22 μm 25 mm diameter hydrophilic polycarbonate membrane filters (Millipore, GTTP02500), with the addition of excess 1X PBS to remove any trace

Table 1
Genome statistics of the recovered SAR11 MAGs from the Chattahoochee River. The recovered SAR11 MAGs represent eight distinct species as defined based on ANI genome comparisons (ANI > 95%). In total, 45 SAR11 MAGs were recovered by our iterative binning pipeline representing 3 major SAR11 clades.

Pelagibacterales species	# of recovered MAGs	Maximum completeness (%)	Average completeness (%)	Clade	Primary habitat[a]
PEL1	17	84	73	IIIb	Lakes
PEL2	16	84	79	IIIb	Lakes
PEL3	2	57	64	IIIb	Lakes
PEL4	2	56	61	IIIb	Lakes
PEL5	2	70.5	71	IIIb	Lakes
PEL6	2	66	69	Ia	Estuaries/marine
PEL7	2	65	68	IIIa	Estuaries
PEL8	2	72	72	Ia	Estuaries/marine

[a] As defined by the biogeographic distribution analysis depicted in Fig. 3.

of formaldehyde. Filters were placed in glass Petri dishes for 1 h to dry, and then placed at −20 °C until further processing.

Each filter was subsequently labeled with catalyzed reporter deposition-fluorescence in situ hybridization (CARD-FISH) [39,49]. First, filters were embedded in 0.2% low gelling point agarose and dried at 35 °C for 30 min, and subsequently incubated in lysozyme solution [10 mg/mL lysozyme, 0.05 M EDTA (pH 8.0), 0.1 M Tris−HCl (pH 7.4)] for 60 min at 37 °C under low agitation. Next the filters were washed three times in excess MilliQ H_2O and incubated in 0.01 M HCl for 10 min to inactivate endogenous peroxidases. In situ hybridizations were performed with a specific 5′ HRP-labeled probe designed for subclade IIIb (LD12) (5′-CACAAGGCAGATTCCCACAT-3′; Integrated DNA Technologies, Iowa, USA) [49]. The hybridization buffer contained 20 mM Tris−HCl (pH 7.4), 0.9 M NaCl, 10% dextran sulfate (wt/vol), 35% (vol/vol) formamide, 1% blocking reagent (Roche, Germany) and 0.01% (wt/vol) SDS. A custom hybridization chamber was built using hermetically sealable container with tissue paper soaked with the hybridization buffer (35% formamide). The filters were soaked in 100 μL hybridization buffer and 1 μL probe (50 ng/μL) and placed in a parafilm covered petri dish inside the hybridization chamber. The chamber was incubated in a hybridization oven at 35 °C for 3 h as previously reported and calculated using MathFISH (http://mathfish.cee.wisc.edu/). After incubation, filters were transferred to 50 mL or pre-warmed wash buffer (20 mM Tris–HCl (pH 7.4), 80 mM NaCl, 5 mM EDTA (pH 8.0), 0.02% (wt/vol) SDS), and incubated at 37 °C for 10 min. After the washing process, the labeled filters were transferred to 50 mL 1X PBS, and incubated for 15 min at room temperature. The hybridization chamber, with tissue paper soaked with MilliQ H_2O, was prepared for the amplification step. Filters were dabbed on tissue paper to remove excess PBS and placed on a parafilm-covered petri dish inside the hybridization chamber. Once in the chamber, 100 μL of the amplification mix were added to each filter and incubated at 37 °C for 45 min in the dark. The amplification mix consisted of the amplification buffer (1X PBS (pH 7.6), 0.1% blocking reagent, 2 M NaCl, 10% (wt/vol) dextran sulfate), mixed with a 0.15% (vol/vol) H_2O_2 solution at a 100:1 ratio. The substrate mix included 4 parts of Alexa Fluor 488-labelled tyramide and 1000 parts of the amplification mix. The filters were then transferred to 50 mL 1X PBS, incubated for 7 min at room temperature in darkness and washed three times in excess MilliQ H_2O, twice in 96% ethanol at room temperature in the dark and then air-dried. DAPI staining was performed using a 0.2 ug/mL DAPI solution, followed by washing in MilliQ H2O and in 100% ethanol. Finally, the filters were air dried (in darkness) and the slide viewed with a Zeiss confocal epifluorescence microscope using DAPI (Zeiss filter set 49) and GFP (Filter set 41012, Chroma, USA) filters.

Data availability

High-quality MAGs produced in this study (Table S2), distances, and other taxonomic analyses are available at http://microbial-genomes.org/projects/WB_binsHQ. Assembled genomes were also deposited in the NCBI GenBank database under BioProject PRJNA495371. All metagenomic datasets from the Chattahoochee riverine ecosystem that were used in this study (Table S1) are available in the NCBI SRA database as part of the BioProjects PRJNA214105 (BioSamples SAMN02302271, SAMN02302272, SAMN02302397, and SAMN02302398) and PRJNA497294 (all BioSamples).

Results and discussion

Sequencing, metagenome assembly, and binning of freshwater SAR11

Sequenced metagenomes (average size of 3.35 Gb, Supplementary Table S1) resulted in an abundance-weighted average coverage between 60% and 89% of the sampled microbial communities (average: 75.9%, n: 68; meaning 75.9% of the extracted DNA was sequenced and 1 dataset discarded due to low coverage) as determined by Nonpareil, which is typically adequate coverage for assembly and genome binning [43]. Subtractive iterative binning resulted in a total of 1126 high-quality MAGs with an average length of 1.97 Mbp (Inter-Quartile Range–IQR–: 1.2–2.5 Mbp) and an average N50 of 10.8 Kbp (IQR: 3.6–11.4 Kbp). We selected a subset of 45 MAGs classified in the order “*Ca.* Pelagibacterales”, with an average length of 769 Kbp (IQR: 710–805 Kbp) and an average N50 of 4.8 Kbp (IQR: 4.1–5.6 Kbp). These MAGs were clustered in eight groups with inter-group ANI < 95% (PEL genomospecies) (Table 1 and Fig. 1), and the genome sequence with the highest quality of each group was selected as a representative (Table S2). Representative genomes had an average length of 831 Kbp (IQR: 751–898 Kbp) with estimated completeness ranging between 51% and 87.7% (average: 73.6%) and estimated contamination between 0% and 2.7% (average: 1.0%). Among the PEL species, PEL1 and PEL2 were the most frequently represented in the MAG collection (n = 17 and 16, respectively), and they also had the most complete genomes, with representative genomes showing >81% completeness and <1.2% contamination (Table S2). The iterative subtractive binning methodology implemented here was meant to capture MAGs with a wide range of abundance and intra-population diversity, in order to reduce the bias towards more abundant clonal populations. However, the use of co-assemblies could also increase the rate of chimerism in the genomes. In order to evaluate the potential increase in chimerism by the co-assembly strategy implemented here, we also targeted the reconstruction of PEL1 genomes from a single metagenomic sample. We subsampled the dataset with the highest abundance of this genomospecies at 2%, 5%, and 10% (LL_1011A; Suppl Fig S3) and assembled and binned these subsamples using the entire dataset for abundance estimation. This resulted in no bins from the 2% sample, and only two bins from each of the other samples, two of which had a high similarity to PEL1: LL_1011A_x005_001 (from the 5% sample) and LL_1011A_x010_001

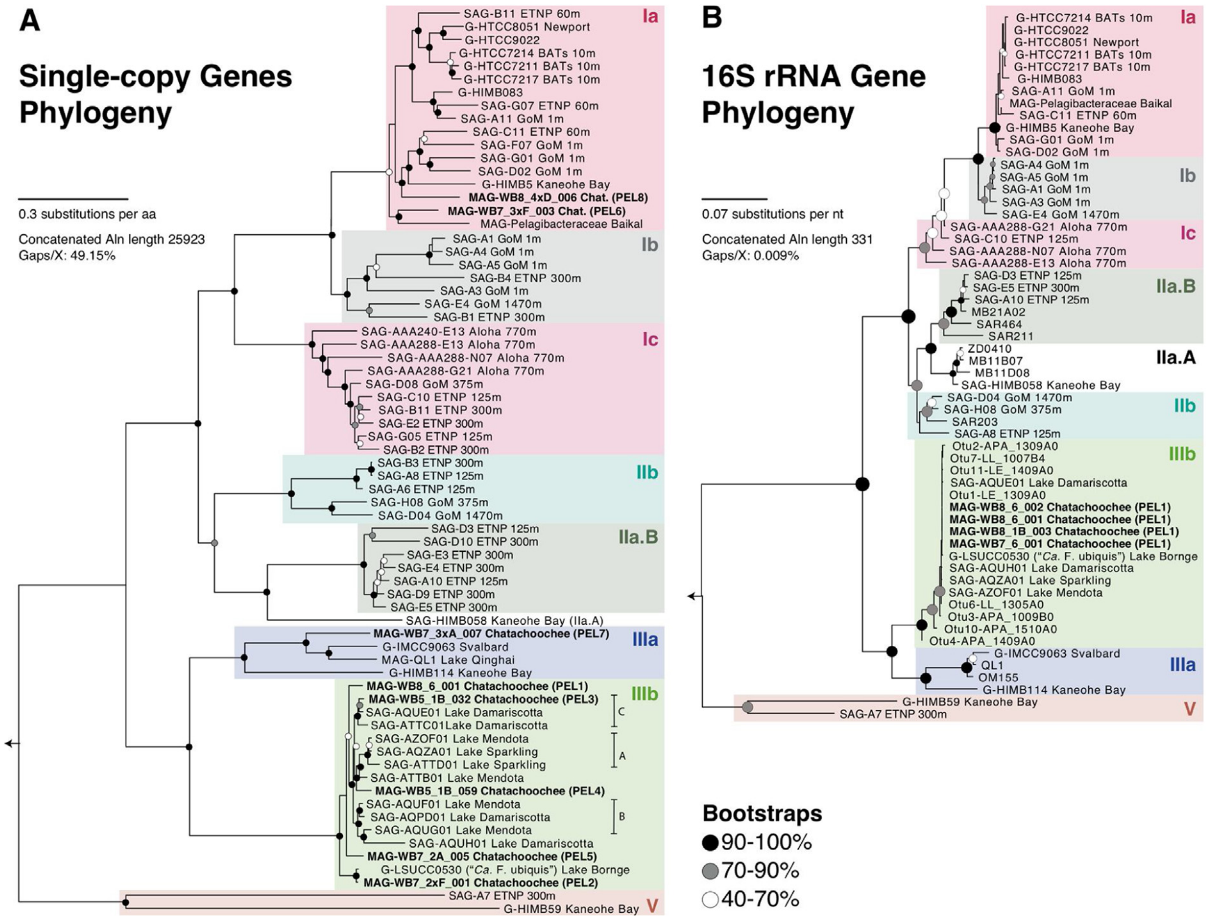


Fig. 1. (A) Phylogenetic reconstruction of recovered freshwater SAR11 MAGs in relation to publicly available SAR11 genomes and MAGs based on the concatenated alignment of 106 single-copy housekeeping genes. The recovered SAR11 MAGs from the Chattahoochee ecosystem are shown in bold (one representative genome for each genomospecies). (B) Phylogeny based on 16S rRNA gene sequences indicating the placement of the PEL1 genomospecies recovered from the Chattahoochee River within the freshwater IIIb clade (LD12). MAGs with a 16S rRNA gene were only recovered for the PEL1 group; the remaining MAGs had 16S rRNA gene sequences too similar to PEL1 16S rRNA sequences to distinguish bioinformatically (>99% nucleotide identity). The dashed line indicates the specificity of the FISH probe used to identify freshwater SAR11 cells. In both trees all included sequences (except clone sequences in the 16S rRNA gene tree) are labeled based on the origin of each organism: G: isolate genome; SAG: Single-cell Amplified Genome; MAG: Metagenome-Assembled Genome.

(from the 10% sample). The former had a very low completeness as estimated by MiGA (47.7%) and was therefore discarded for further analyses. The latter had a total length of 2.2 Mbp in 669 contigs (N50: 5012 bp), and an estimated completeness of 84.7%. However, it had an estimated contamination of 21.6% (genome quality: −23.3), indicating that for this group the chimerism introduced by the combination of alpha diversity and lack of co-abundance data is much greater than the effect of gamma diversity alone. The most similar genome to LL_1011A_x010_001 in our collection was WB7_1B_004, with 98.87% ANI over 76% of the genome.

*Genetic diversity of "*Ca.* Pelagibacterales" MAGs*

We subsequently compared the identified eight genomospecies with 60 previously described SAR11 genomes originating from various environments and representing all known lineages (Table S3). ANI clustering at 95% identity resulted in 62 total genomospecies, i.e., most described SAR11 genomes represent a distinct species based on this ANI threshold. Among the eight PEL groups, only two were classified within the same species with previously described genomes, i.e., PEL2 and PEL3 with *Ca* Fonsibacter ubiquis [17] and the single amplified genome SAG ATTC01 isolated from lake Damariscotta [64], respectively, while the rest represented new species at the 95% ANI threshold. Five genomospecies (PEL1-5) were phylogenetically classified within the freshwater clade IIIb, one within the clade IIIa (PEL7), and two were deep branching lineages within the clade Ia (PEL6 and PEL8).

Phylogenetic reconstructions based on 16S rRNA could not resolve the clade IIIb members into distinct species (Fig. 1b), i.e. most members showed 16S rRNA identity >98.5%, a typical threshold to define OTUs (operational taxonomic units) and often species. Nevertheless, phylogenetic analysis of single copy marker genes (Fig. 1a), as well as pairwise AAI genomic comparisons (Fig. 2) of the available SAR11 genomes reflected the ANI clustering and genomospecies definition, further corroborating that the freshwater clade (as most other SAR11 clades) are represented by multiple distinct species. While ANI comparisons can adequately represent genetic relatedness of closely related groups, AAI comparisons provide extended dynamic range when comparing more distantly related genomes, *i.e.* representatives of the same genera or families [45]. Indeed AAI pairwise comparisons of the available SAR11 genomes (Fig. 2) revealed that the entire IIIb clade could be considered a single genus "*Ca.* Fonsibacter" (AAI > 82% among genomospecies of clade IIIb), designated based on the only available isolate up to date [17]. Six freshwater genomospecies were captured for the first time by our collection and represented distinct species from all previously recovered clade IIIb genomes, including SAGs from lakes Damariscotta (MN, USA), Mendota, and Sparkling (WI, USA) [11,24] and the "*Ca.* F. ubiquis" LSUCC0530^T isolate recovered from the brackish water lagoon Lake Borgne [17].

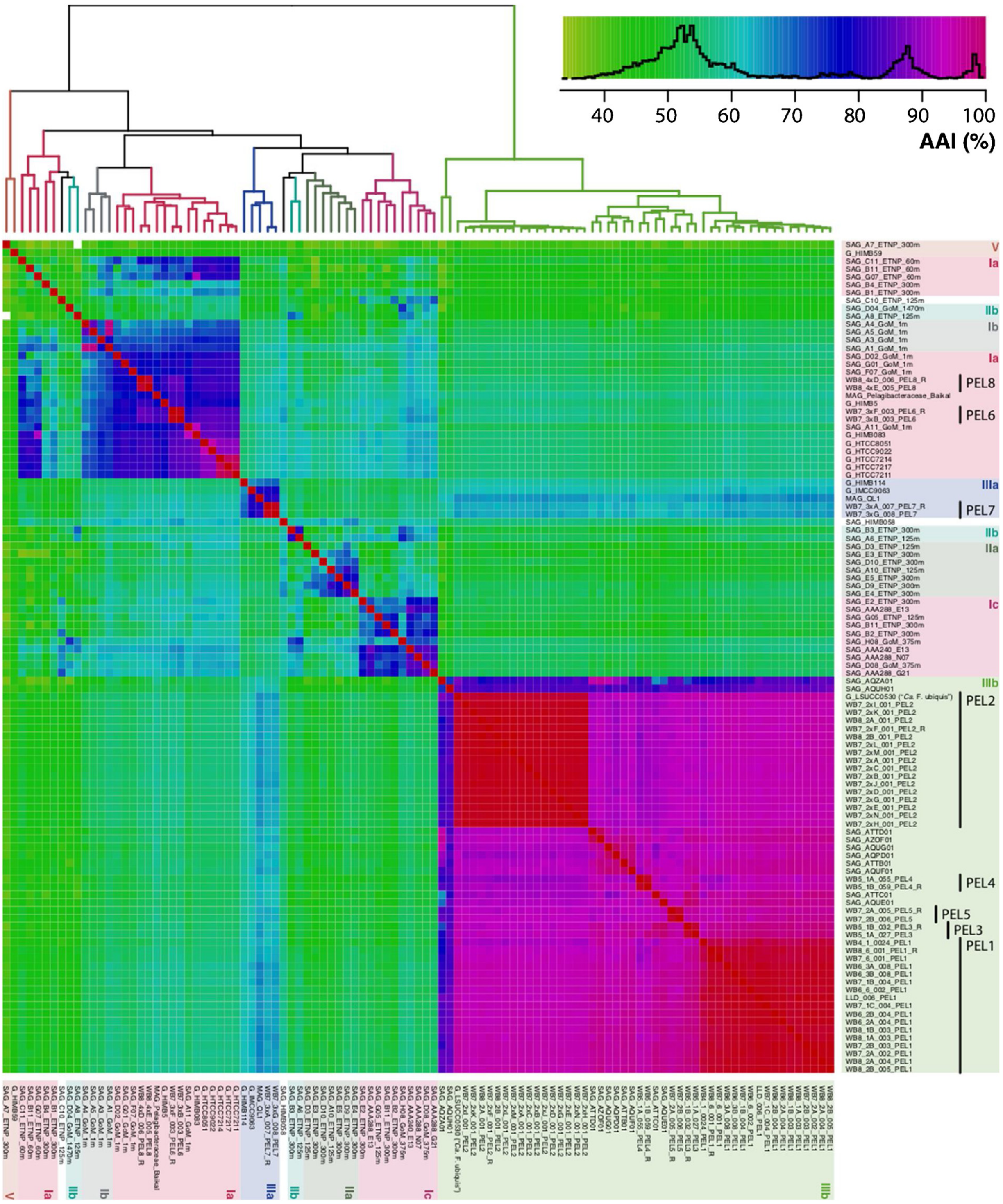
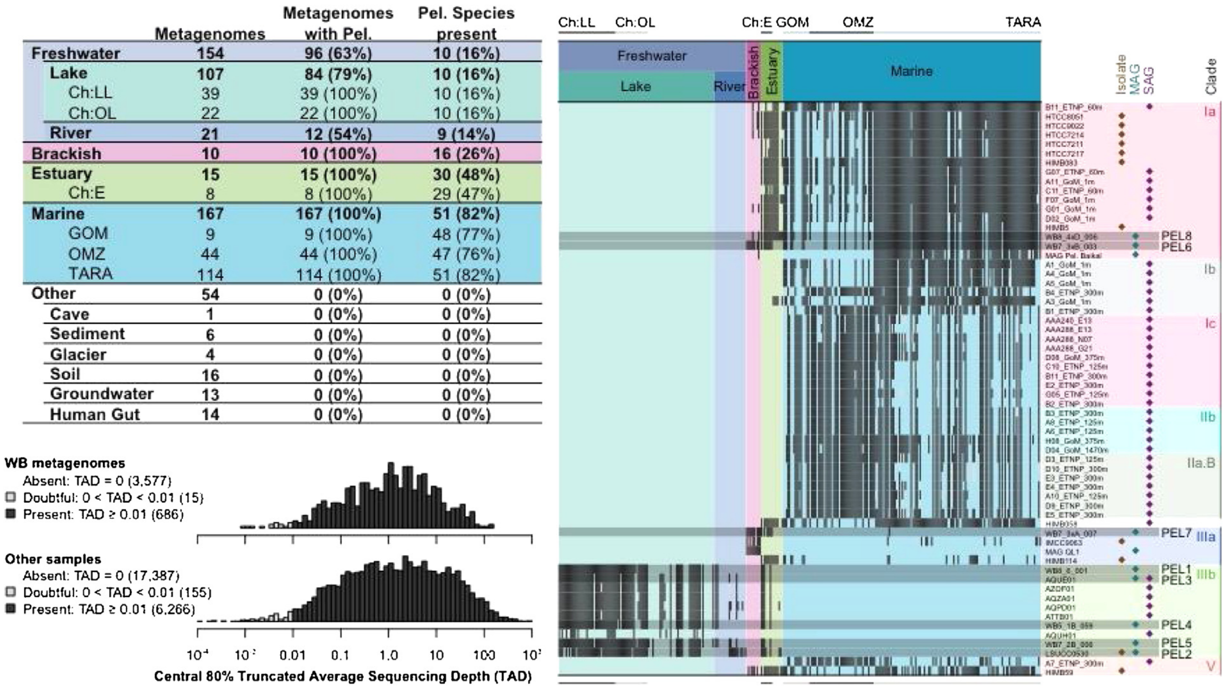


Fig. 2. Average amino-acid identity (AAI) comparisons of all recovered freshwater "*Ca.* Pelagibacterales" MAGs with previously reported representative genomes. MAGs recovered from the Chattahoochee interconnected lakes are designated based on their corresponding genomospecies (PEL1-8) and all genomes are color-labeled based on their phylogenetic clade classification from Fig. 1. Note that all genomes of the clade IIIb show AAI > 82% among them thus, could be classified within the same genus.

Among the clade IIIb species, PEL1 (n = 17, representative genome: WB8_6_001) was the most abundant and broadly distributed group in our samples (Figs. 3 and 4), and represented a deeply branching lineage within the clade IIIb. No reference genomes have been previously identified for this group (with ANI > 95%). The closest match among reference genomes was SAG ATTC01 from Lake Damariscotta (ANI: 88.5%, AAI: 89.1%), likely a member of genomospecies PEL3.

The group PEL2 (n = 16, representative genome: WB7_2xF_001) clustered in the same genomospecies with the only available isolate up to date "*Ca.* F. ubiquis" LSUCC0530^{T} (ANI: 98.1%, AAI: 97.2%). Importantly, the description of the genus "Fonsibacter" and the species "Fonsibacter ubiquis" were proposed to be included in the category *Candidatus* [17]. However, the description does not conform to the rules for the category *Candidatus* as adopted by the International Committee on Systematics of Prokaryotes (ICSP), stat-



	Metagenomes	Metagenomes with Pel.	Pel. Species present
Freshwater	**154**	**96 (63%)**	**10 (16%)**
Lake	**107**	**84 (79%)**	**10 (16%)**
Ch:LL	39	39 (100%)	10 (16%)
Ch:OL	22	22 (100%)	10 (16%)
River	**21**	**12 (54%)**	**9 (14%)**
Brackish	**10**	**10 (100%)**	**16 (26%)**
Estuary	**15**	**15 (100%)**	**30 (48%)**
Ch:E	8	8 (100%)	29 (47%)
Marine	**167**	**167 (100%)**	**51 (82%)**
GOM	9	9 (100%)	48 (77%)
OMZ	44	44 (100%)	47 (76%)
TARA	114	114 (100%)	51 (82%)
Other	**54**	0 (0%)	0 (0%)
Cave	**1**	0 (0%)	0 (0%)
Sediment	**6**	0 (0%)	0 (0%)
Glacier	**4**	0 (0%)	0 (0%)
Soil	**16**	0 (0%)	0 (0%)
Groundwater	**13**	0 (0%)	0 (0%)
Human Gut	**14**	0 (0%)	0 (0%)

Fig. 3. Biogeographic distribution of "*Ca.* Pelagibacterales" genomospecies across various environmental habitats. The presence of each MAG in a given metagenomic dataset was defined by a non-zero TAD value (truncated average sequencing depth, see Section Materials and methods). The left upper panel represents the summary of the metagenomic datasets evaluated for the presence of SAR11 species, which included the 67 freshwater samples from the Chattahoochee River ecosystem obtained in this study (Table S1) and representative metagenomes from various other habitats (Table S4). The left lower panel shows the distribution of TAD values for the SAR11 species evaluated here (n = 62 genomospecies). The right panel shows the presence/absence of each genomospecies in all metagenome assessed. Ch LL: Lake Lanier datasets, Ch OL: Other Lakes on Chattahoochee, Ch E: Estuaries of the Chattahoochee River, GOM: Nine surface water samples from the Gulf of Mexico, OMZ: 44 samples from the oxygen minimum zone of the ETNP, TARA: 125 marine samples from the TARA expedition.

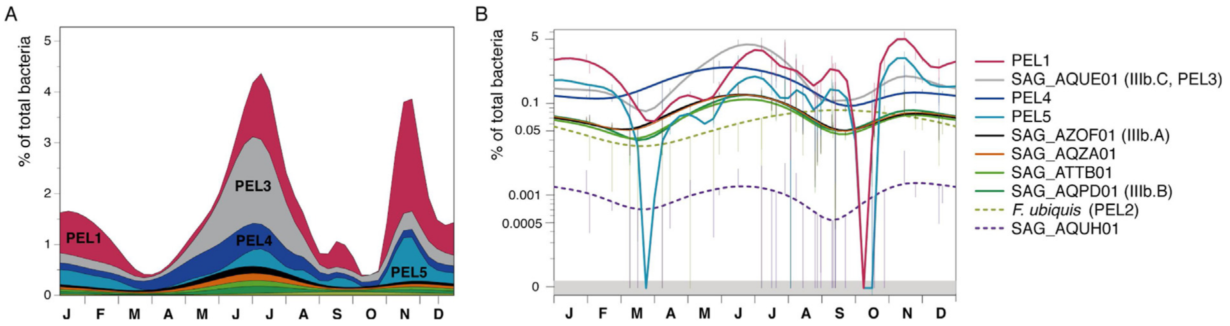


Fig. 4. Temporal dynamics of "*Ca.* Fonsibacter" PEL1-5 genomospecies in Lake Lanier. Predicted abundance of each genomospecies is shown as a percentage over the total bacteria in a stacked abundance plot (A) or separately for each species in logarithmic scale (B). Logarithmic scale was used in order to more efficiently visualize the dynamics of the low abundance members. Zeroes are indicated in a shaded band. Species that were identified to be seasonal are shown with solid lines. No seasonality could be verified for the two low-abundance members shown in dashed lines. The underlying seasonal model that was used to predict the abundances is shown in Fig. S3.

ing that the *Candidatus* category is reserved for well-characterized but as-yet uncultured organisms [53]. Therefore, the availability of the cultured strain LSUCC0530 precludes inclusion in this category. However, we maintained the designation *Candidatus* throughout this manuscript in order to conform to the original publication and to highlight the provisional status of the name.

The closest relative of the PEL3 representative genome (WB5_1B_032) among reference genomes was SAG ATTC01 (ANI: 95.9%, AAI: 92.4%), within the same genomospecies as defined by ANI > 95%. The PEL3 genomospecies is equivalent to the micro-cluster C lineage of clade IIIb (Fig. 1a), described previously for the collection of the freshwater SAR11 SAGs from lakes Damariscotta, Mendota and Sparkling [64]. No closely-matching reference genomes were identified for the remaining two freshwater genomospecies PEL4 (best match: SAG ATTB01, AAI: 88.6%) or PEL5 (best match: SAG AQUF01, AAI: 87.1%).

Additionally, the recovered SAR11 MAGs included an additional three genomospecies classified outside the clade IIIb with AAI < 61% to "*Ca.* F. ubiquis" LSUCC0530^T, possibly representing two genera yet to be described. Two of those genomospecies (PEL6 and PEL8) are classified within the clade Ia and their closest relative among previously available genomes or SAGs is "*Ca.* P. ubique" HTCC9022 (AAI: 79.0% and 77.5%, respectively). While traditionally considered a marine clade, clade I was recently recognized to encompass genomes from various environments, including a freshwater representative that was identified in metagenomic datasets from Lake Baikal in Russia [2]. As detailed below, PEL6 and PEL8 are most likely marine water representatives recovered from our estuarine metagenomes, based on their abundance distributions across the Chattahoochee riverine system.

Finally, two of the identified SAR11 MAGs were classified within clade IIIa (genomospecies PEL7), and their closest relative was "*Ca.*

Pelagibacter sp." IMCC9063 (AAI: 74.5%) isolated from the coastal waters of Svaldbard, Norway [33]. A sister clade to IIIb, clade IIIa consists of brackish/estuarine water representatives (Fig. 1a), as well as a freshwater MAG isolated from Lake Qinghai, China [34]. Based on their environmental distribution described below, PEL7 representatives most likely represent a species adapted to intermediate salinity waters (coastal, brackish) and not found to be highly abundant in marine environments.

In summary, we present here a total of 45 "*Ca.* Pelagibacterales" MAGs from 8 genomospecies, including a group belonging to the same species as the recently obtained freshwater isolate "*Ca.* F. ubiquis" (PEL2) recovered from the Lake Damariscotta (PEL3), 3 novel freshwater groups (PEL1, PEL4, and PEL5), and three novel estuarine/marine groups probably belonging to as yet-undescribed genera (PEL6, PEL7, and PEL8). All freshwater genomospecies representatives (PEL1-5) have AAI > 85% with "*Ca.* F. ubiquis" LSUCC0530^T, indicating they all belong to a single genus ("*Ca.* Fonsibacter").

Environmental distribution of identified "Ca. *Pelagibacterales" genomospecies*

The eight "*Ca.* Pelagibacterales" species recovered from the Chattahoochee riverine ecosystem exhibited similar but distinct biogeographic and temporal patterns. We first evaluated the presence of each genomospecies along the interconnected lakes of the Chattahoochee ecosystem, as well as within an additional collection of freshwater (n = 154), estuarine/brackish (n = 25), and marine (n = 164) samples from various geographic locations (Fig. 3, Table S4). Presence was defined as the coverage of at least 10% of the reference MAG sequence by highly identical metagenomic reads as described in Section "Material and methods".

Genomospecies from the clades Ia and IIIa (PEL6-8) appear to be restricted to estuarine and marine samples only (Fig. 3), and, based on the metagenome collection reported by our study, were only encountered in the estuaries of the Chattahoochee River. All the other groups (PEL1-5) from SAR11 clade IIIb were detected in most freshwater and some estuarine samples collected as part of our study and other previous collections of freshwater and estuarine metagenomes, but not in marine samples (Fig. 3). These include the genomospecies PEL2, including "*Ca.* F. ubiquis". PEL1 was found in most available freshwater samples and fewer estuarine samples compared to the PEL2 group. Similar to PEL1, the other genomospecies PEL3-5 were also found in the majority of freshwater samples evaluated here with similar distributions.

We subsequently evaluated the temporal dynamics of clade IIIb species (n = 10) in two time-series metagenomic collections from the temperate Lake Lanier (6 years, 32 samples) and Mendota (4 years, 92 samples). In Lake Lanier, freshwater SAR11 exhibit their highest relative abundance during the summer stratification (June and July) and late fall (November) (Fig. 4). Similarly, seasonal fluctuations in the abundance of the total SAR11 population (albeit less pronounced) were observed in Lake Mendota (Fig. S4). Strong seasonal variations have been previously reported for freshwater SAR11 in temperate lakes such as Lake Erken, Sweden [16] and Lake Zurich, Switzerland [49], linked with seasonal variations in physicochemical parameters. Similar to their marine counterparts [4,10] and consistent with an oligotrophic lifestyle, relative freshwater SAR11 cell numbers increase with increased temperatures [49] and water stratification, and decrease during high productivity periods [16].

All genomospecies were detected at some time in both lakes. However, only a couple of them made up the majority of the freshwater SAR11 population and exhibited a strong periodicity as described above. In Lake Lanier, four genomospecies made up the large majority of SAR11 cells, with varying contributions throughout the year. While the PEL1 genomospecies is one of the most abundant members of the SAR11 population during both SAR11 seasonal blooms, PEL3 and PEL4 dominate during the summer and PEL5 during the winter. In contrast, the most abundant genomospecies in Lake Mendota were two genomospecies represented by the previously described SAGs isolated from Lakes Mendota and Sparkling [11,64] (Fig. S4) revealing that different SAR-11 genomospecies may be predominantly abundant locally.

Gene content comparisons of "Ca. *Fonsibacter" genomospecies*

The estimated average genome size of the recovered freshwater SAR11 MAGs was 1.20 Mbp with a high average coding density of 95.3%, in accordance with the characteristics of a small genome size and high coding density that typify SAR11 species. Indeed, the average genome size of 40 reference SAR11 genomes from various clades is estimated at 1.38 Mbp with a 94.4% coding density (Table S3) (1.16Mbp and 95.6% coding density for "*Ca.* Fonsibacter ubiquis" LSUCC0530^T). Our iterative assembly and binning pipeline captured multiple genomes for each recovered genomospecies from different samples. Genomes of the PEL1 group, which is one of the most prominent SAR11 members in the Chattahoochee River, had the highest genome quality. For example, the representative PEL1 MAG WB8_6_001 was estimated to be 81% complete with minimal contamination (0.3%). Due to the more incomplete nature of the MAGs for the rest of the genomospecies relative to that of PEL1, we focused our gene content and phenotypic comparisons of PEL1 to the closest classified genome, i.e., "*Ca.* Fonsibacter ubiquis" LSUCC0530^T. We first identified the core gene set of PEL1, defined as all protein-coding genes that were found in at least 2 out of the 17 recovered PEL1 MAGs, in order to eliminate potential contamination bias, while accounting for the incompleteness of the MAGs. Similarly, we defined the species gene core for PEL2, by cataloguing genes that were either found in the complete genome "*Ca.* Fonsibacter ubiquis" LSUCC0530^T, or in at least two out of the 16S PEL2 MAGs.

This consensus core genome of the PEL1 genomospecies was estimated to be 98.0% complete, including 1256 unique genes; similar to the complete genome LSUCC0530^T that encodes 1231 genes. Overall, the gene content of PEL1 genomes included all previously described basic metabolic functions for "*Ca.* Fonsibacter ubiquis" LSUCC0530^T and freshwater SAR11 SAGs [11,17]. PEL1 is predicted to be an aerobic microorganism that relies on its chemoorganotrophic lifestyle as the main source for energy and carbon. As described for all available freshwater clade IIIb genomes up to date, PEL1 genomes encode a complete TCA cycle and a proteorhodopsin, but lack both a complete Entner–Doudoroff (ED) pathway (i.e., phosphogluconate dehydratase and aldolase) or the alternate ED pathway that has been found in several mostly coastal and marine SAR11 isolates [15,50]. Instead, PEL1 genomes encode all genes for the Embden–Meyerhof–Parnas (EMP) pathway [11]; for example, they have the 6-phosphofructokinase (*pfk*) and pyruvate kinase (*pyk*) genes. Similarly to LSUCC0530^T, PEL1 genomes differ in C1 oxidation metabolic pathways compared to their marine relatives [55], having lost genes for DMSP, methylamine, or glycine-betaine metabolism, but retaining genes for tetrahydrafolate metabolism, formate oxidation, and the glycine cleavage pathway. Additionally, with the exception of an incomplete homoserine biosynthesis and lack of the homoserine o-succinyltransferase gene, the PEL1 species was predicted to have identical (and complete) amino acid biosynthesis capabilities with the previously described freshwater genomes: complete pathways for histidine, valine, leucine, isoleucine, aspartate, asparagine, arginine, lysine, glutamate, glutamine, and proline biosynthesis. In contrast with most marine isolates, PEL1 and the freshwater genomes encode a complete pathway for serine biosynthesis (*serABC*) but lack genes for the biosynthesis of glycine [58]. Finally, and similarly to the described freshwater SAR11 isolate, PEL1 are auxotrophic for reduced sulfur since they lack a sulfite oxidase, a common feature found in marine relatives [57].

A small number of genes was found to be different between PEL1 and the LSUCC0530^T genomes. Unique gene content encoded by LSUCC0530^T but not found in any PEL1 genome included 164 genes (Table S5): 50 were located in the characteristic hypervariable region of the LSUCC0530^T and 65 genes were classified as hypothetical proteins. Amongst absent genes, several were likely of phage origin and predicted to be glycosyltransferases and genes for lipid biosynthesis and transport. The fructokinase gene responsible for the conversion of D-fructose to β-D-fructose-6-P, present in LSUCC0530^T was absent from all PEL1 MAGs (Table S5). Similarly, 225 genes were identified as unique in PEL1 genomes and not present in LSUCC0530^T, among which 132 were classified as hypothetical and 89 were predicted to be of phage origin (Table S6). PEL1 encodes proteins for both glutamine and L-glutamate biosynthesis and metabolism similar to the LSUCC0530^T genome; but unlike the LSUCC0530^T genome, PEL1 encodes an additional glutamate dehydrogenase which can utilize ammonia, α-ketoglutarate and NAD(P)H to yield glutamate. Additional genes unique to PEL1 include an amino acid ABC transporter permease and a glutamate-aspartate periplasmic binding protein and several citrate transporters (Table S6).

Comparisons among the other PEL genomospecies or between them and PEL1 and PEL2 revealed similar results, i.e., the genomospecies are all very identical to each other in terms of gene content and only a few, non-conclusive functions with respect to functional differentiation were found. These functions and their corresponding genes should be the subject of future work and the PEL genome sequences made available here can facilitate future research.

Cell morphology of “Ca. *Fonsibacter*” *species*

CARD-FISH was used instead of FISH as a more sensitive technique for the small size and low ribosomal content of SAR11 cells [39]. The CARD-FISH probes used in this study were designed to capture cells of the SAR11 LD12 clade [49]. The probes allowed visualization of SAR11 from Lake Lanier water samples. The presence of a large number of photosynthetic microorganisms in these samples resulted in high levels of auto-fluorescence and hence, high fluorescence background during microscopy. To overcome the auto-fluorescence, we pre-filtered water samples through 5 μm porosity filters, to exclude large photosynthetic cells, and treated the filters containing fixed cells with 3% H_2O_2 at room temperature for 10 min. All observed labeled cells were small, curved rods of approximately 1 μm × 0.2 μm (Fig. 5, Fig. S4), which is consistent with the morphology reported for “*Ca.* Fonsibacter ubiquis” LSUCC0530^T and the marine representative “*Ca.* Pelagibacter ubique” HTCC1062^T. While we were not able to differentiate the various PEL species based on CARD-FISH labeling due to high identity of their 16S rRNA gene sequence, we expect that most of the cells visualized belonged to the species PEL1 since it was the most abundant SAR11 organism in Lake Lanier (see above) especially during the winter month samples that were analyzed here (Fig. 4).

Concluding remarks: Freshwater SAR11 as an ideal system to study bacterial microdiversity and speciation

In summary, we recovered representative genomes of five genomospecies of the freshwater SAR11 IIIb clade, PEL1 through PEL5, and three additional genomospecies, PEL6 through PEL8, representing more divergent, estuarine clades. The described PEL1-5 genomospecies appear to be highly related to each other, e.g., showing AAI values >80%; yet, they are distinct in terms of genomic relatedness, measured by ANI or other metrics. Thus, these freshwater genomospecies were traceable over time and space based on metagenomics. Tracking the genomospecies in our time-series revealed different in-situ relative abundances but highly correlated seasonal abundance patterns, with most genomospecies being persistent across the seasons and peaking in abundance in the summertime (e.g., Fig. 4). The co-occurrence of the genomospecies in the same samples/conditions suggested that they are somewhat **functionally distinct**, otherwise environmental conditions would have selected for the most fit genomospecies and purge diversity (which was not the case; e.g., Fig. 1). Indeed, similar patterns of persistent microdiversity among closely related co-existing populations have been previously observed for abundant freshwater populations such as closely related species Actinobacteria the acI clade [31] or Polynucleobacter species [18,33], which exhibit significant gene content differences which could potentially explain their ecological diversification.

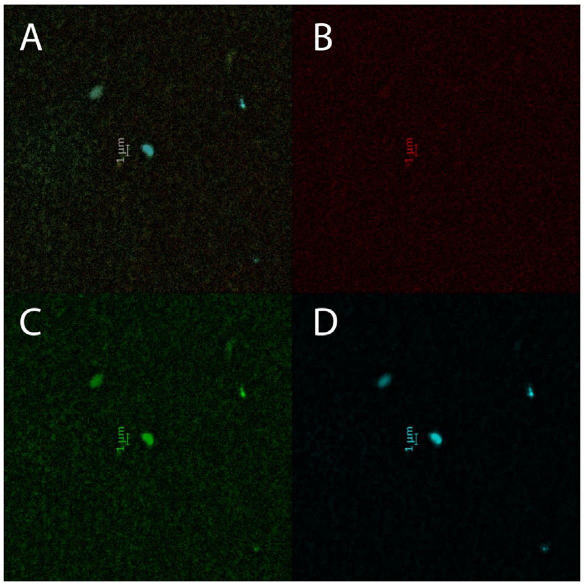


Fig. 5. Direct microscopic observation of CARD-FISH and DAPI-stained cells from Lake Lanier (December 2018 sample). CARD-FISH and DAPI was used to visualize SAR11 clade IIIb and all cells in a Lake Lanier water sample, respectively. Composite figure (A) showing autofluorescence control (red, B), CARD-FISH (green, C) and DAPI staining (blue, D) indicating the vibrioid morphology of SAR11 cells. Note that the probe used was not specific to distinguish between the different freshwater PEL species. See also supplementary figure S4 where SAR11 cells are shown adjacent to non-SAR11 cells. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

However, the gene content differences among these five genomospecies and between them and the named *Ca* Fonsibacter ubiquis (PEL3) appeared to be limited to relatively few genes of unknown or poorly characterized functions, viral predation genes, and a couple of central metabolism and organic compound utilization functions. The functional importance of these genes remains unclear at present but should be the subject of future research. Subtle genomic differences and overall high gene content conservation and metabolic adaptations is a common observation for closely related, yet distinct coexisting marine SAR11 species [3,15]. Based on the limited and mostly uncharacterized (hypothetical) gene content differences observed among the freshwater genomes identified here, it is indeed possible that the PEL genomospecies are not substantially metabolically different from each other, but rather the distinctive functional/phenotypic differences may be related to differential **phage predation** and/or **adaptation (fine tuning)** of core genes to specific environmental conditions such as temperature [46,47]. Understanding which metabolic and/or ecological factors may differentiate the PEL genomospecies will provide new insights into how closely-related and phenotypically highly-similar organisms may co-exist in the same habitat without competitive exclusion and hence, provide new insights into the process of speciation. Therefore, the freshwater SAR11 clade III represents an ideal system to study speciation since gene content (and thus, phenotypic) differences and/or genetic divergence are typically much smaller within this group relative to those observed within other microbial groups such as the *Escherichia coli*. The genome sequences determined as part of our study have already provided testable hypotheses (e.g., genes are differentially present among genomospecies), and should greatly facilitate future studies of freshwater SAR11. Obtaining cultured representatives of these genomospecies to help test the hypotheses created by the genome sequences with direct competition and laboratory experiments would greatly facilitate efforts to understand the speciation process of this important group of environmental organisms.

Description of "Candidatus *Fonsibacter lacus*"

Fonsibacter lacus *(L. gen. n.* lacus, *of a lake)*

We propose to name PEL1, the most complete and persistent genomospecies of clade IIIb in the freshwater ecosystems sample, as *Candidatus* Fonsibacter lacus. In addition to the previously described properties of the genus *Fonsibacter* [17], the proposed species is described as follows. Small, curved rod cells of approximately 1 × 0.2 μm. Partial genome sequence WB8_6_001 with estimated completeness of 77.5% and contamination of 0.9% is 1.08 Mbp in size, with 1375 predicted genes, a coding density of 92.0%, and a G + C content of 35.9%. WB8_6_001 has an ANI of 84.9% and AAI of 87.0% with *Ca.* Fonsibacter ubiquis str. LSUCC0530^T genome [17], the only currently described member of the *Ca* Fonsibacter genus. The type material for this species is the genome WB8_6_001, identified in the freshwater lakes along the Chattahoochee River, Southeast USA. The representative genome is available in GenBank under accession number SAMN10223538.

Funding

This work was supported by NSF (awards 1241046 and 1759831 to KTK).

Acknowledgements

We would like to thank Yuanqi Wang, Brittany Suttner, Minjae Kim, and Blake Lindner that have helped with sample collection and processing of the time series Lake Lanier samples. We thank Dr. Rosello-Mora for his advice on conducting the CARD-FISH.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.syapm.2019.03.007.

References

[1] Brown, M.V., Lauro, F.M., DeMaere, M.Z., Muir, L., Wilkins, D., Thomas, T., Riddle, M.J., Fuhrman, J.A., Andrews-Pfannkoch, C., Hoffman, J.M., McQuaid, J.B., Allen, A., Rintoul, S.R., Cavicchioli, R. (2012) Global biogeography of SAR11 marine bacteria. Mol. Syst. Biol. 8, 595, http://dx.doi.org/10.1038/msb.2012.28.

[2] Cabello-Yeves, P.J., Zemskaya, T.I., Rosselli, R., Coutinho, F.H., Zakharenko, A.S., Blinov, V.V., Rodriguez-Valera, F. (2018) Genomes of novel microbial lineages assembled from the sub-ice waters of Lake Baikal. Appl. Environ. Microbiol. 84 (1), http://dx.doi.org/10.1128/AEM.02132-17, e02132-17.

[3] Cameron Thrash, J., Temperton, B., Swan, B.K., Landry, Z.C., Woyke, T., DeLong, E.F., Stepanauskas, R., Giovannoni, S.J. (2014) Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. ISME J. 8 (7), 1440–1451, http://dx.doi.org/10.1038/ismej.2013.243.

[4] Carlson, C.A., Morris, R., Parsons, R., Treusch, A.H., Giovannoni, S.J., Vergin, K. (2009) Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. ISME J. 3 (3), 283–295, http://dx.doi.org/10.1038/ismej.2008.117.

[5] Castro, J., Rodriguez-R, L.M., Weigand, M.R., Hatt, J.K., Carter, M.Q., Konstantinidis, K.T. (2018) imGLAD: Accurate detection and quantification of target organisms in metagenomes. PeerJ 6, e5882 https://dx.doi.org/10.7717%2Fpeerj.5882.

[6] Cox, M.P., Peterson, D.A., Biggs, P.J. (2010) SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinf. 11 (1), 485, http://dx.doi.org/10.1186/1471-2105-11-485.

[7] DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.-U., Martinez, A., Sullivan, M.B., Edwards, R., Brito, B.R., Chisholm, S.W., Karl, D.M. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311 (5760), 496–503, http://dx.doi.org/10.1126/science.1120250.

[8] Dupont, C.L., Larsson, J., Yooseph, S., Ininbergs, K., Goll, J., Asplund-Samuelsson, J., McCrow, J.P., Celepli, N., Allen, L.Z., Ekman, M., Lucas, A.J., Hagström, Å., Thiagarajan, M., Brindefalk, B., Richter, A.R., Andersson, A.F., Tenney, A., Lundin, D., Tovchigrechko, A., Nylander, J.A.A., Brami, D., Badger, J.H., Allen, A.E., Rusch, D.B., Hoffman, J., Norrby, E., Friedman, R., Pinhassi, J., Venter, J.C., Bergman, B. (2014) Functional tradeoffs underpin salinity-driven divergence in microbial community composition. PLoS One 9 (2), e89549, http://dx.doi.org/10.1371/journal.pone.0089549.

[9] Dupont, C.L., Rusch, D.B., Yooseph, S., Lombardo, M.-J., Alexander Richter, R., Valas, R., Novotny, M., Yee-Greenbaum, J., Selengut, J.D., Haft, D.H., Halpern, A.L., Lasken, R.S., Nealson, K., Friedman, R., Craig Venter, J. (2012) Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. ISME J. 6 (6), 1186–1199, http://dx.doi.org/10.1038/ismej.2011.189.

[10] Eiler, A., Hayakawa, D.H., Church, M.J., Karl, D.M., Rappé, M.S. (2009) Dynamics of the SAR11 bacterioplankton lineage in relation to environmental conditions in the oligotrophic North Pacific subtropical gyre. Environ. Microbiol. 11 (9), 2291–2300, http://dx.doi.org/10.1111/j.1462-2920.2009.01954.x.

[11] Eiler, A., Mondav, R., Sinclair, L., Fernandez-Vidal, L., Scofield, D.G., Schwientek, P., Martinez-Garcia, M., Torrents, D., McMahon, K.D., Andersson, S.G., Stepanauskas, R., Woyke, T., Bertilsson, S. (2016) Tuning fresh: radiation through rewiring of central metabolism in streamlined bacteria. ISME J. 10 (8), 1902–1914, http://dx.doi.org/10.1038/ismej.2015.260.

[12] Eiler, A., Zaremba-Niedzwiedzka, K., Martínez-García, M., McMahon, K.D., Stepanauskas, R., Andersson, S.G.E., Bertilsson, S. (2014) Productivity and salinity structuring of the microplankton revealed by comparative freshwater metagenomics. Environ. Microbiol. 16 (9), 2682–2698, http://dx.doi.org/10.1111/1462-2920.12301.

[13] Enright, A.J., Van Dongen, S., Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30 (7), 1575–1584.

[14] Glöckner, F.O., Fuchs, B.M., Amann, R. (1999) Bacterioplankton compositions of lakes and oceans: a first comparison based on fluorescence in situ hybridization. Appl. Environ. Microbiol. 65 (8), 3721–3726.

[15] Grote, J., Thrash, J.C., Huggett, M.J., Landry, Z.C., Carini, P., Giovannoni, S.J., Rappé, M.S. (2012) Streamlining and core genome conservation among highly divergent members of the SAR11 clade. MBio 3 (5), http://dx.doi.org/10.1128/mBio.00252-12, e00252-12.

[16] Heinrich, F., Eiler, A., Bertilsson, S. (2013) Seasonality and environmental control of freshwater SAR11 (LD12) in a temperate lake (Lake Erken, Sweden). Aquat. Microb. Ecol. 70 (1), 33–44, http://dx.doi.org/10.3354/ame01637.

[17] Henson, M.W., Lanclos, V.C., Faircloth, B.C., Thrash, J.C. (2018) Cultivation and genomics of the first freshwater SAR11 (LD12) isolate. ISME J. 1, http://dx.doi.org/10.1038/s41396-018-0092-2.

[18] Hoetzinger, M., Schmidt, J., Jezberová, J., Koll, U., Hahn, M.W. (2017) Microdiversification of a pelagic polynucleobacter species is mainly driven by acquisition of genomic islands from a partially interspecific gene pool. Appl. Env. Microbiol. 83 (3), http://dx.doi.org/10.1128/AEM.02266-16, e02266-16.

[19] Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., Bork, P. (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. Mol. Biol. Evol. 34 (8), 2115–2122, http://dx.doi.org/10.1093/molbev/msx148.

[20] Hugerth, L.W., Larsson, J., Alneberg, J., Lindh, M.V., Legrand, C., Pinhassi, J., Andersson, A.F. (2015) Metagenome-assembled genomes uncover a global brackish microbiome. Genome Biol. 16, 279, http://dx.doi.org/10.1186/s13059-015-0834-7.

[21] Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinf. 11, 119, http://dx.doi.org/10.1186/1471-2105-11-119.

[22] Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., Aluru, S. (2017) High-throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. BioRxiv, 225342, http://dx.doi.org/10.1101/225342.

[23] Kang, D.D., Froula, J., Egan, R., Wang, Z. (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3, e1165, http://dx.doi.org/10.7717/peerj.1165.

[24] Kashtan, N., Roggensack, S.E., Rodrigue, S., Thompson, J.W., Biller, S.J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R.R., Stocker, R., Follows, M.J., Stepanauskas, R., Chisholm, S.W. (2014) Single-cell genomics reveals hundreds of coexisting subpopulations in wild prochlorococcus. Science 344 (6182), 416–420, http://dx.doi.org/10.1126/science.1248575.

[25] Konstantinidis, K.T., Tiedje, J.M. (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. Proc. Natl. Acad. Sci. U. S. A. 101 (9), 3160–3165, http://dx.doi.org/10.1073/pnas.0308653100.

[26] Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.-H., Rognes, T., Ussery, D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35 (9), 3100–3108, http://dx.doi.org/10.1093/nar/gkm160.

[27] Langmead, B., Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods 9 (4), 357–359, http://dx.doi.org/10.1038/nmeth.1923.

[28] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25 (16), 2078–2079, http://dx.doi.org/10.1093/bioinformatics/btp352.

[29] Logares, R., Bråte, J., Bertilsson, S., Clasen, J.L., Shalchian-Tabrizi, K., Rengefors, K. (2009) Infrequent marine–freshwater transitions in the microbial world. Trends Microbiol. 17 (9), 414–422, http://dx.doi.org/10.1016/j.tim.2009.05.010.

[30] Morris, R.M., Rappé, M.S., Connon, S.A., Vergin, K.L., Siebold, W.A., Carlson, C.A., Giovannoni, S.J. (2002) SAR11 clade dominates ocean surface bacterioplankton communities. Nature 420 (6917), 806–810, http://dx.doi.org/10.1038/nature01240.

[31] Neuenschwander, S.M., Ghai, R., Pernthaler, J., Salcher, M.M. (2018) Microdiversification in genome-streamlined ubiquitous freshwater Actinobacteria. ISME J. 12 (1), 185–198, http://dx.doi.org/10.1038/ismej.2017.156.

[32] Newton, R.J., Jones, S.E., Eiler, A., McMahon, K.D., Bertilsson, S. (2011) A guide to the natural history of freshwater lake bacteria. Microbiol. Mol. Biol. Rev. MMBR 75 (1), 14–49, http://dx.doi.org/10.1128/MMBR.00028-10.

[33] Oh, H.-M., Kang, I., Lee, K., Jang, Y., Lim, S.-I., Cho, J.-C. (2011) Complete genome sequence of strain IMCC9063, belonging to SAR11 subgroup 3, isolated from the Arctic Ocean. J. Bacteriol. 193 (13), 3379–3380, http://dx.doi.org/10.1128/JB.05033-11.

[34] Oh, S., Zhang, R., Wu, Q.L., Liu, W.-T. (2014) Draft genome sequence of a novel SAR11 clade species abundant in a Tibetan Lake. Genome Announc. 2 (6), http://dx.doi.org/10.1128/genomeA.01137-14.

[35] Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., Phillippy, A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17, 132, http://dx.doi.org/10.1186/s13059-016-0997-x.

[36] Orellana, L.H., Rodriguez-R, L.M., Konstantinidis, K.T. (2017) ROCker: accurate detection and quantification of target genes in short-read metagenomic data sets by modeling sliding-window bitscores. Nucleic Acids Res. 45 (3), http://dx.doi.org/10.1093/nar/gkw900, e14–e14.

[37] Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res., http://dx.doi.org/10.1101/gr.186072.114.

[38] Peng, Y., Leung, H.C.M., Yiu, S.M., Chin, F.Y.L. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinf. Oxf. Engl. 28 (11), 1420–1428, http://dx.doi.org/10.1093/bioinformatics/bts174.

[39] Pernthaler, A., Pernthaler, J., Amann, R. (2002) Fluorescence in situ hybridization and catalyzed reporter deposition for the identification of marine bacteria. Appl. Environ. Microbiol. 68 (6), 3094–3101.

[40] Pruesse, E., Peplies, J., Glöckner, F.O. (2012) SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 28 (14), 1823–1829, http://dx.doi.org/10.1093/bioinformatics/bts252.

[41] Quinlan, A.R., Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26 (6), 841–842, http://dx.doi.org/10.1093/bioinformatics/btq033.

[42] Rodriguez-R, L.M., Guntruru, S., Harvey, W.T., Rosselló-Móra, R., Tiedje, J.M., Cole, J., Konstantinidis, K.T. (2018) The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and bacteria at the whole genome level. Nucleic Acid Res.

[43] Rodriguez-R, L.M., Konstantinidis, K.T. (2014) Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. Bioinf. Oxf. Engl. 30 (5), 629–635, http://dx.doi.org/10.1093/bioinformatics/btt584.

[44] Rodriguez-R, L.M., Konstantinidis, K.T. (2016) The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ, http://dx.doi.org/10.7287/peerj.preprints.1900v1.

[45] Rodriguez-R, L.M., Konstantinidis, K.T. (2014) Bypassing cultivation to identify bacterial species. Microbe 9 (3), 111–118.

[46] Rodriguez-Valera, F., Martin-Cuadrado, A.-B., López-Pérez, M. (2016) Flexible genomic islands as drivers of genome evolution. Curr. Opin. Microbiol. 31, 154–160, http://dx.doi.org/10.1016/j.mib.2016.03.014.

[47] Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., Pašić, L., Thingstad, T.F., Rohwer, F., Mira, A. (2009) Explaining microbial population genomics through phage predation. Nat. Rev. Microbiol. 7 (11), 828–836, http://dx.doi.org/10.1038/nrmicro2235.

[48] Roux, S., Enault, F., Hurwitz, B.L., Sullivan, M.B. (2015) VirSorter: mining viral signal from microbial genomic data. PeerJ 3, e985, http://dx.doi.org/10.7717/peerj.985.

[49] Salcher, M.M., Pernthaler, J., Posch, T. (2011) Seasonal bloom dynamics and ecophysiology of the freshwater sister clade of SAR11 bacteria 'that rule the waves' (LD12). ISME J. 5 (8), 1242–1252, http://dx.doi.org/10.1038/ismej.2011.8.

[50] Schwalbach, M.S., Tripp, H.J., Steindler, L., Smith, D.P., Giovannoni, S.J. (2010) The presence of the glycolysis operon in SAR11 genomes is positively correlated with ocean productivity. Environ. Microbiol. 12 (2), 490–500, http://dx.doi.org/10.1111/j.1462-2920.2009.02092.x.

[51] Sekar, R., Pernthaler, A., Pernthaler, J., Warnecke, F., Posch, T., Amann, R. (2003) An improved protocol for quantification of freshwater Actinobacteria by Fluorescence In Situ Hybridization. Appl. Environ. Microbiol. 69 (5), 2928–2935, http://dx.doi.org/10.1128/AEM.69.5.2928-2935.2003.

[52] Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D., Higgins, D.G. (2014) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7 (1), http://dx.doi.org/10.1038/msb.2011.75, 539–539.

[53] Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A.D., Kämpfer, P., Maiden, M.C.J., Nesme, X., Rosselló-Mora, R., Swings, J., Trüper, H.G., Vauterin, L., Ward, A.C., Whitman, W.B. (2002) Report of the ad hoc committee for the reevaluation of the species definition in bacteriology. Int. J. Syst. Evol. Microbiol. 52 (3), 1043–1047, http://dx.doi.org/10.1099/00207713-52-3-1043.

[54] Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinf. Oxf. Engl. 22 (21), 2688–2690, http://dx.doi.org/10.1093/bioinformatics/btl446.

[55] Sun, J., Steindler, L., Thrash, J.C., Halsey, K.H., Smith, D.P., Carter, A.E., Landry, Z.C., Giovannoni, S.J. (2011) One carbon metabolism in SAR11 pelagic marine bacteria. PLoS One 6 (8), e23973, http://dx.doi.org/10.1371/journal.pone.0023973.

[56] Thrash, J.C., Boyd, A., Huggett, M.J., Grote, J., Carini, P., Yoder, R.J., Robbertse, B., Spatafora, J.W., Rappé, M.S., Giovannoni, S.J. (2011) Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. Sci. Rep. 1, http://dx.doi.org/10.1038/srep00013.

[57] Tripp, H.J., Kitner, J.B., Schwalbach, M.S., Dacey, J.W.H., Wilhelm, L.J., Giovannoni, S.J. (2008) SAR11 marine bacteria require exogenous reduced sulphur for growth. Nature 452 (7188), 741–744, http://dx.doi.org/10.1038/nature06776.

[58] Tripp, H.J., Schwalbach, M.S., Meyer, M.M., Kitner, J.B., Breaker, R.R., Giovannoni, S.J. (2009) Unique glycine-activated riboswitch linked to glycine-serine auxotrophy in SAR11. Environ. Microbiol. 11 (1), 230–238, http://dx.doi.org/10.1111/j.1462-2920.2008.01758.x.

[59] Tsementzi, D., Wu, J., Deutsch, S., Nath, S., Rodriguez-R, L.M., Burns, A.S., Ranjan, P., Sarode, N., Malmstrom, R.R., Padilla, C.C., Stone, B.K., Bristow, L.A., Larsen, M., Glass, J.B., Thamdrup, B., Woyke, T., Konstantinidis, K.T., Stewart, F.J. (2016) SAR11 bacteria linked to ocean anoxia and nitrogen loss. Nature 536 (7615), 179–183, http://dx.doi.org/10.1038/nature19068.

[60] Vergin, K.L., Beszteri, B., Monier, A., Cameron Thrash, J., Temperton, B., Treusch, A.H., Kilpert, F., Worden, A.Z., Giovannoni, S.J. (2013) High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. ISME J. 7 (7), 1322–1332, http://dx.doi.org/10.1038/ismej.2013.32.

[61] Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73 (16), 5261–5267, http://dx.doi.org/10.1128/AEM.00062-07.

[62] Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Mazumder, R., O'Donovan, C., Redaschi, N., Suzek, B. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res. 34 (Database issue), D187–191, http://dx.doi.org/10.1093/nar/gkj161.

[63] Wu, Y.-W., Simmons, B.A., Singer, S.W. (2016) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32 (4), 605–607, http://dx.doi.org/10.1093/bioinformatics/btv638.

[64] Zaremba-Niedzwiedzka, K., Viklund, J., Zhao, W., Ast, J., Sczyrba, A., Woyke, T., McMahon, K., Bertilsson, S., Stepanauskas, R., Andersson, S.G.E. (2013) Single-cell genomics reveal low recombination frequencies in freshwater bacteria of the SAR11 clade. Genome Biol. 14 (11), R130, http://dx.doi.org/10.1186/gb-2013-14-11-r130.