Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: www.elsevier.de/syapm

Candidatus Prosiliicoccus vernus, a spring phytoplankton bloom associated member of the *Flavobacteriaceae*

T. Ben Francis, Karen Krüger, Bernhard M. Fuchs, Hanno Teeling, Rudolf I. Amann*

Max Planck Institute for Marine Microbiology, Bremen, Germany

ARTICLE INFO

Keywords:
Metagenome assembled genome
Helgoland
North Sea
Laminarin
Flow cytometric sorting

ABSTRACT

Microbial degradation of algal biomass following spring phytoplankton blooms has been characterised as a concerted effort among multiple clades of heterotrophic bacteria. Despite their significance to overall carbon turnover, many of these clades have resisted cultivation. One clade known from 16S rRNA gene sequencing surveys at Helgoland in the North Sea, was formerly identified as belonging to the genus *Ulvibacter*. This clade rapidly responds to algal blooms, transiently making up as much as 20% of the free-living bacterioplankton. Sequence similarity below 95% between the 16S rRNA genes of described *Ulvibacter* species and those from Helgoland suggest this is a novel genus. Analysis of 40 metagenome assembled genomes (MAGs) derived from samples collected during spring blooms at Helgoland support this conclusion. These MAGs represent three species, only one of which appears to bloom in response to phytoplankton. MAGs with estimated completeness greater than 90% could only be recovered for this abundant species. Additional, less complete, MAGs belonging to all three species were recovered from a mini-metagenome of cells sorted via flow cytometry using the genus specific ULV995 fluorescent rRNA probe. Metabolic reconstruction indicates this highly abundant species most likely degrades proteins and the polysaccharide laminarin. Fluorescence in situ hybridisation showed coccoid cells, with a mean diameter of 0.78 mm, with standard deviation of 0.12 μm. Based on the phylogenetic and genomic characteristics of this clade, we propose the novel candidate genus *Candidatus* Prosiliicoccus, and for the most abundant and well characterised of the three species the name *Candidatus* Prosiliicoccus vernus.

Introduction

Species of the class *Flavobacteriia* are among the most numerically abundant bacteria in coastal oceans during and immediately following phytoplankton bloom events [10,11,36,49,50]. Surveys of bacterial diversity during spring blooms and throughout the year in the North Sea, as well as in other regions of the North Atlantic, have demonstrated repeated patterns of occurrence of specific clades of *Flavobacteriia*, which typically form blooms of their own in response to springtime increases in algal abundance [11,30,31,49,50]. In common with other *Bacteroidetes* such as those found in mammalian digestive tracts, it has been proposed that marine *Flavobacteriia* typically make use of higher molecular weight polymeric organic matter such as algal derived protein and polysaccharide [18,56]. Indeed the *Bacteroidetes* are known to frequently possess characteristic genomic structures known as polysaccharide utilisation loci (PULs), which are assemblages of carbohydrate active enzymes (CAZymes) [51] typically physically co-located in the genome with genes for a TonB-dependent receptor derived transport system [19,33,52,53]. Many cultivated marine flavobacterial strains such as *Gramella forsetii* KT0803^{T} [23], *Polaribacter* spp. Hel1_33_49 and Hel1_85 [57], and *Zobellia galactanivorans* Dsij^{T} [6], have been shown to possess multiple PULs in their genomes, which vary in CAZyme content dependent on the specific polysaccharide that is targeted. Detailed description of the degradative capabilities of the free-living spring bloom associated bacterioplankton has, however, largely focused on community-level analysis rather than on individual taxa [9,49,50], and currently there are only a handful of cultivated species genuinely representative of the major bloom-forming genera [21]. It is therefore of considerable value to begin describing these bloom associated communities at the level of individual genomes, given that they play a major role in the recycling of organic material and remineralisation of fixed carbon in surface oceans.

One clade identified as an important part of the post-algal bloom flavobacterial community in the North Sea has been previously referred to as belonging to the genus *Ulvibacter* [11,49,50]. This clade was found to make up as much as 20% of the total free-

* Corresponding author at: Celsiusstraße 1, 28359 Bremen, Germany.
E-mail address: ramann@mpi-bremen.de (R.I. Amann).

https://doi.org/10.1016/j.syapm.2018.08.007

living bacterial community during and after spring bloom events. Additionally it appeared that it responded concurrent with and in the immediate aftermath of the initial algal bloom, suggesting a possible niche associated with living and senescent algae or their exudates rather than dead algal or bacterial cell material. Recalcitrance of this clade to cultivation has, however, precluded inference of plausible growth substrates, and all that had been known thus far was based on 16S rRNA gene sequences and estimates of environmental abundances measured both by direct cell counting via fluorescence in situ hybridisation (FISH), and by 16S rRNA gene amplicon sequencing.

Here we describe this *Ulvibacter*-related clade as the novel genus *Candidatus* Prosiliicoccus, from the Latin *prosilio* — meaning to jump up, rush, or break forth. The genus currently comprises three North Sea species, one of which is sufficiently well represented in our data that we describe it as a novel species *Candidatus* Prosiliicoccus vernus (henceforth referred to for brevity as *Ca.* Pv), reflecting its initial identification from spring samples. Our circumscription is based on multiple metagenome assembled genomes (MAGs) derived from assembled metagenomic datasets from samples collected across spring blooms in the years 2010, 2011, and 2012. Additional data relating to the environmental abundance of this genus and its phylogenetic position also support our description.

Materials and methods

Sampling

Surface seawater samples were collected from the long-term ecological research site Kabeltonne at the island of Helgoland in the North Sea (54°11.3′N, 7°54.0′E), as described previously [49,50]. Biomass of free-living bacteria for DNA extraction was collected on 0.2 μm pore-size polycarbonate filters following prefiltration steps using both 10 and 3 μm filters to remove larger primarily eukaryotic material and debris, including particle attached bacteria.

Seawater for FISH was fixed by direct addition of formaldehyde (final concentration of 1%) to the sample, followed by filtration on 0.2 μm pore-size polycarbonate filters without pre-filtration. For cell sorting, 10 l of unfixed water sample from the 21st of April 2009 was filtered onto a polycarbonate filter (142 mm diameter, 0.2 μm pore-size) within 3 h of sampling. All filters were kept frozen at −80 °C until processing.

Fluorescence in situ hybridisation

For visualisation and size estimation, cells of *Ca.* Prosiliicoccus on a filter from 08/04/2010, when it was determined to be highly abundant [50], were labelled using catalysed reporter deposition-FISH (CARD-FISH), using probe ULV995, and following the protocol of Thiele et al. [54].

For flow cytometric cell sorting, a modified hybridisation chain reaction-FISH (HCR-FISH) [58] was done, again using ULV995. The initiator probe was the *Ca.* Prosiliicoccus specific ULV-I-995 initiatorH (5′-CCGAATACAAAGCATCAACGACTAGA-AAAA-TCCACGCCTGTCAGACTACA-3′). This was used in conjunction with the two competitors ULV-I-995 c1 (5′-TCCACTCCTGTCAGACTACA-3′) and ULV-I-995 c2 (5′-TCCACCCCTGTCAGACTACA-3′). NON EUB initiatorH (5′-CCGAATACAAAGCATCAACGACTAGA-AAAAA-ACTCCTACGGGAGGCAGC-3′) was used as a negative control. Hybridisation was done in direct-gene-FISH buffer as described by Barrero-Canosa et al. [7], and contained 35% formamide with a final probe concentration of 1 μM. The sample was then hybridised for 120 min at 46 °C, before washing for 30 min at 48 °C in washing buffer. Amplification was carried out with four-times Alexa Fluor™ 488 – labelled oligonucleotides H1 (5′-TCTAGTCGTT(G)*ATGCTTT(G)*TATTCGGCGACA(G)*ATAACCG-AATACAAA(G)*CATC-3′) and H2 (5′- CCGAATACAAA(G)-*CATCAAC(G)*ACTAGAGATGCTTT(G)*TATTCG(G)*TTATCTGTCG-3′) in amplification buffer after they had been denatured for 90 s at 95 °C followed by 30 min at 25 °C, and kept at 20 °C until further use. Amplification was done for 120 min at 37 °C. Final washing was done twice in 1× PBS at 4 °C for 5 min and in deionised water for 30 s. Filters were air dried and embedded in CitiFluor™ AF1 supplemented to a final concentration of 2 ng μl^{-1} with 4′,6-diamidino-2-phenylindole (DAPI). Microscopy was done on a Zeiss LSM 780 confocal laser scanning microscope equipped with an Airyscan detector, using a 63× Plan-Apochromat objective lens (Carl Zeiss, Jena, Germany).

Cell sorting using FISH, and sorted cell mini-metagenome generation

A piece of polycarbonate filter containing approximately 8.5 × 10^{8} cells (sample date 21/04/2009) was hybridised overnight at 35 °C with the *Ca.* Prosiliicoccus specific ULV-I-995 initiatorH (see *Fluorescence in situ hybridisation* above), washed in washing buffer for 30 min at 35 °C [58] and subsequently incubated in amplification buffer-H1/H2 probe mix in a humidity chamber at 35 °C for 2 h. Hybridised filters were then cut into small pieces, transferred into a 1.5 ml tube containing 1.3 ml of ice-cold cell resuspension buffer (150 mM NaCl, 0.05% Tween80) and vortexed for 15 min at 4 °C. The supernatant containing hybridised cells was kept on ice until analysis and cell sorting. Cell sorting was conducted using a MoFlo flow cytometer (Beckman Coulter, Krefeld, Germany). Supernatants containing resuspended cells were DAPI stained, and prefiltered through a 5 μm pore-size polycarbonate filter (Millipore, 13 mm diameter) to avoid nozzle clogging. Cells were sorted according to their combined FISH and DAPI signal and stored at −20 °C until further processing. The purity of sorted cells was checked microscopically. In order to avoid contamination, DNA amplification from sorted cells was done in a UV-treated PCR workstation using the illustra GenomiPhi V2 DNA Amplification (MDA) Kit (GE Healthcare) according to the manufacturer's instructions. Ten replicates of ∼500 sorted, ULV995-positive cells were lysed by three freeze/thaw cycles (−20 °C/room temperature) and subsequent alkaline lysis before subjection to the MDA reaction. Sorted calibration beads served as a negative control. After taxonomic verification of the MDA products by 16S rRNA gene sequencing, the genome sequencing of the MDA product with the highest yield was performed by JGI using the Illumina MiSeq platform (San Diego, CA, USA) and a 2 × 150 bp protocol. Following sequencing, raw reads were trimmed and quality filtered to remove the TruSeq adapters and low quality sequence using BBDuk v35.14 (http://bbtools.jgi.doe.gov). Options used for trimming were as follows: ktrim = r k = 28 mink = 12 hdist = 1 tbo = t tpe = t qtrim = rl trimq = 20 minlength = 100. Read quality for each sample was then confirmed using FastQC v0.11.2 [3].

Metagenome sequencing

Sequencing of ten metagenome samples at the DOE Joint Genome Institute (JGI) has been described previously [50] (sample dates: 03/03/2010; 08/04/2010; 04/05/2010; 18/05/2010; 24/03/2011; 238/04/2011; 26/05/2011; 08/03/2012; 16/04/2012; 10/05/2012). An additional 28 metagenomic samples from intervening dates across the same period were also sequenced at JGI using the same procedures (sample dates: 30/03/2010; 13/04/2010; 20/04/2010; 23/04/2010; 30/04/2010; 11/05/2010; 21/03/2011; 28/03/2011; 31/03/2011; 04/04/2011; 07/04/2011; 14/04/2011; 21/04/2011; 26/04/2011; 06/05/2011; 09/05/2011; 12/05/2011; 16/05/2011; 19/05/2011; 23/05/2011; 30/05/2011;

05/04/2012; 12/04/2012; 26/04/2012; 03/05/2012; 24/05/2012; 31/05/2012; 07/06/2012). For full details of sample preparation and sequencing for each sample see Supplementary Table S1. Briefly, extracted DNA was sheared to average length of 270 bp by sonication, and then sequenced on the Illumina HiSeq platform following a 2 × 150 bp protocol. The ten samples from Teeling et al. [50] were sequenced more deeply than the 28 additional samples, resulting in approximately four times as many reads. Following sequencing, raw reads were trimmed and quality filtered as detailed above for *Cell sorting using FISH and sorted cell mini-metagenome generation.*

Metagenome assembly and binning

Quality filtered reads from each metagenomic sample and also the sorted cell mini-metagenome were assembled individually using SPAdes v3.10.0 [37] in -meta mode with kmer lengths of 21, 33, 55, 77, and 99, and with read error correction enabled. Contigs longer than 2.5 kbp were retained for binning. Each metagenome assembly was binned using CONCOCT [1] as part of the standard anvi'o v3 workflow [16]. To generate differential coverage information for CONCOCT, SPAdes error corrected reads from the assembled sample and the reads from four other randomly selected datasets from the same year were mapped back to the assembly. Reads were mapped with BBMap v35.14 (http://bbtools.jgi.doe.gov), using 'fast' mode, minimum mapping identity (minid) of 0.99, and identity filter for reporting mappings (idfilter) of 0.97. The sorted cell mini-metagenome was binned directly using the anvi'o interactive interface (anvi-interactive function), using reads mapped in the same manner as above from all 38 metagenomic samples. SPAdes error corrected reads from the sorted mini-metagenome itself were not included as these were the product of a single MDA run and would therefore not be expected to give meaningful differential coverage information between species.

Bin selection and refinement

Bins from metagenomes deriving from *Candidatus* Prosiliicoccus populations were initially identified using the output of CheckM tree v1.0.8 [39], which produces an approximate phylogenomic placement of metagenomic bins. Bins were selected for further refinement that had a close phylogenetic relationship to known *Ulvibacter* species and Unidentified eubacterium SCB49, which is also sometimes referred to as *Ulvibacter* sp. SCB49. Additionally, bin similarity was assessed using Mash v1.1.1 [38] with the default sketch size of 1000. A mash distance cutoff of less than 0.05 – approximating average nucleotide identity (ANI) of greater than 95% – was used to determine if two bins belonged to the same species, and this then produced three clusters of bins representing the three *Ca.* Prosiliicoccus species, two of which clusters contained a sorted cell mini-metagenome MAG (metagenome assembled genome). The selected metagenome bins were then manually refined using the anvi'o interactive interface (anvi-refine function) to produce high quality MAGs. In order to produce MAGs with lower L50 and higher N50 values, the refined bins (excluding sorted cell mini-metagenome MAGs) of the *Ca.* Prosiliicoccus species were then 'co-reassembled'. Reads from each of the 38 metagenomic samples were remapped to each MAG using BBMap as in *Metagenome assembly and binning* above, and all reads mapping to MAGs of the two lower abundant *Ca.* Prosiliicoccus species were then co-assembled with SPAdes in careful mode without error correction enabled using kmers of length 21, 33, 55, 77, 99, and 127. In the case of the more abundant *Ca.* Pv, however, reassembly using reads from all 38 metagenomes did not produce an improved assembly, and so only reads from the 13/04/2010 metagenome dataset that mapped to the most complete MAG (Prosiliicoccus_vernus_Helgoland_20100413) were reassembled, using the same SPAdes parameters as for the other reassemblies. The resulting reassemblies were then refined with anvi'o by mapping reads from each of the 38 metagenomic samples back to the reassembly using BBMap as before, followed by profiling with minimum contig length of 1000 base pairs, and manual refinement in the anvi'o interactive interface. Assessments of MAG completeness and redundancy were made using both CheckM's lineage workflow, with anvi'o, and with the HMM.essential.rb script from the enveomics collection [45] in metagenome (-M) mode. The reassembled MAGs were then taken for further analyses.

Phylogenomic and 16S rRNA gene phylogenetic reconstruction

Phylogenomic reconstruction was based on concatenated sequences of 40 proteins (Supplementary Table S2) present in all three reassembled *Ca.* Prosiliicoccus MAGs, taken from the 82 phylogenetically conserved bacterial proteins listed by Soo et al. [47]. Reference amino acid sequences of other *Flavobacteriaceae* and *Bacteroidetes* were downloaded from NCBI GenBank, with the North Sea Gammaproteobacterium *Reinekea* sp. Hel_1_31_D35 used as outgroup. Amino acid sequences were predicted for the *Ca.* Prosiliicoccus MAGs using Prodigal v2.6.3 [22] in metagenomic mode (-p meta). Sequences of the 40 phylogenetic markers (Supplementary Table S2) were identified in the MAGs and reference genomes using the hmmsearch function of HMMER v3 [15]. Sequences were aligned using FAMSA v1.2 [13] with default parameters, and phylogenomic trees calculated using RAxML v8.2.9 [48] with automatic selection of substitution model, and rapid-bootstrapping with 1000 resamplings (-m PROTGAMMAAUTO -p 12345 -x 12345 -# 1000 -o Reinekea_sp_Hel1_31_D35). Trees were visualised using iTOL [29].

16S rRNA gene based phylogeny was calculated using the full length 16S rRNA gene sequences detected by anvi'o in the MAGs Prosiliicoccus_vernus_Helgoland_20110523, Prosiliicoccus_vernus_Helgoland_20100518, Prosiliicoccus_vernus_Helgoland_20110421, and Prosiliicoccus_vernus_Helgoland_20110426. These sequences were aligned with SINA v1.3.0 [41] to the SILVA NR Ref database v128 [42], along with all *Flavobacteriaceae* in SILVA v128, and the *Bacteroidetes Bacteroides vulgatus* ATCC 8482 and *Prevotella brevis* ATCC 19188, and using *Reinekea blandensis* MED297^T as outgroup. 16S rRNA genes deriving from the two lower abundant *Ca.* Prosiliicoccus species that produced less complete MAGs could not be detected in the assembled metagenomes, and thus could not be included in this part of the analysis. Phylogeny was reconstructed using RAxML with the same bootstrapping as above, but using the GTRGAMMA substitution model (-f a -m GTRGAMMA -p12345 -x 12345 -#1000 -o Reinekea_blandensis_MED297).

Phylogenetic uniqueness was assessed using both percent identity across the full length 16S rRNA gene sequences, calculated in ARB [32], and by calculation of ANI and average amino acid identity (AAI) between MAGs and the genomes of related species using the ani.rb and aai.rb scripts from the enveomics collection [45].

Estimation of environmental abundance

Direct cell counting to estimate absolute cell numbers using CARD-FISH has been described previously [50]. Cell counts from that study made using hybridisation with the ULV995 probe, which we consider to be specific to the genus *Ca.* Prosiliicoccus, were used here to plot absolute cell numbers. The relevant data from that work is reproduced here in Supplementary Table S3.

Similarly, estimates of relative abundance based on proportion of reads deriving from amplicon data from samples also collected at Helgoland have been described previously [11]. Data from Chafee et al. [11] referring to *Ulvibacter* are used here for additional monitoring of this clade. Sequence identity between the most abundant oligotype sequence classified by Chafee et al. [11] as *Ulvibacter* and the 16S rRNA gene from the *Ca.* Prosiliicoccus MAGs was 100%.

Data for global abundance and distribution was collected using IMNGS [28], using the 16S rRNA gene sequence from the MAG Prosiliicoccus_vernus_Helgoland_20110426 as a query. Minimum target size was 200, and an identity threshold of 99% was used. Percent of reads in each sequencing run was calculated from the IMNGS output, and the corresponding geographic positions for each sequencing run (where these data were available) were collected from NCBI. An arbitrary cutoff of at at least 50 reads matching the query was used for plotting.

Species relative abundance was also assessed based on the proportion of metagenomic reads recruited to individual bins. Reads from all 38 metagenomic samples were thus mapped to the reassembled MAGs of each *Ca.* Prosiliicoccus species, and the number of reads recruited were counted and normalised to the total number of reads in that sample. These numbers then estimate the proportion of reads deriving from the different *Ca.* Prosiliicoccus populations over time. Reads were mapped as detailed above in *Metagenome assembly and binning*.

Assessment of single nucleotide variation and strain diversity in Ca. Prosiliicoccus vernus

Single nucleotide variants (SNVs) in all MAGs, including the reassembled MAGs, were called by anvi'o using the anvi-gen-variability-profile tool, with -min-coverage-in-each-sample set to 20×. Metagenomic samples were selected for inclusion in this analysis based on the average detection of the MAG by each sample, as calculated by anvi'o. Samples where detection of the MAG was greater than or equal to 0.9 were included, meaning that at least 90% of the nucleotide positions in the MAG had at least one read mapping to them. Number of SNVs per thousand base pairs and average coverage of SNVs in each MAG were then calculated from the output.

Inference of number and abundance of strains represented by the reassembled *Ca.* Pv MAG was also attempted using DESMAN [43] as described on the DESMAN github pages (https://github.com/chrisquince/DESMAN). The mapping files created for measuring abundance were used as inputs. Core COGs used were the 40 identified by Mende et al. [34]. The variant filter was run with -p and -c options, and the coverage cutoff was reduced to 2 in order to include more samples from early 2011 and 2012. The same coverage cutoff was used for running the DESMAN algorithm. The DESMAN algorithm was run with 10 replicates, with -r 1000 and -i 500 as recommended.

MAG annotation and metabolic reconstruction

Initial annotation was done with Prokka v1.12 [46], modified to include prediction and annotation of partial genes by removing the -c and -m options when running Prodigal within Prokka. This annotation was then manually refined for *Ca.* Pv using searches against Pfam v31.0 [17] (pfam_scan.pl script with default parameters), and BLAST v2.6.0+ [2] searches against the most up-to-date NCBI nr database (downloaded 13/03/2018). For all three species, specific annotation of CAZymes using the dbCAN v6 database [61] and the hmmscan function of HMMER was also used. Custom e-value cutoffs for specific CAZyme families were used as described previously [50]. Comparison of CAZyme sequence identity against experimentally verified sequences was done using BLAST. Peptidases were predicted using BLAST against the MEROPS merops_scan database v12.0 [44], using the default BLAST settings recommended by MEROPS: e-value cutoff of 1×10^{-4}. Cellular localisation of glycoside hydrolase and peptidase enzymes was predicted using CELLO v2.5 [62] and PSORTb v3.0 [63]. SusC/D-like transporters were predicted using the TIGRFAM [20] profile TIGR04056 for SusC-like sequences and Pfam profiles PF07980.9, PF12741.5, PF12771.5, and PF14322.4 for SusD-like sequences. Additional TonB dependent receptors were predicted using TIGRFAM profiles TIGR01352, TIGR01776, TIGR01778, TIGR01779, TIGR01782, TIGR01783, TIGR01785, TIGR01786, TIGR02796, TIGR02797, TIGR02803, TIGR02804, TIGR02805, and TIGR04057. For TonB- dependent receptors and SusC-like genes, an e-value cutoff of 1×10^{-10} was used, and for SusD-like genes an e-value cutoff of 1×10^{-5} was used. Function of sulfatases was predicted using SulfAtlas v1.2 [5]. Metabolic pathway information was reconstructed from the Prokka output using Pathway Tools v20.5 [24].

Data availability

Metagenome reads for the 38 environmental metagenomes used in this study are available under the NCBI BioProject accession numbers listed in Supplementary Table S1. The sorted cell mini-metagenome reads are available under NCBI BioProject accession PRJNA367155. Accession numbers for the metagenome assemblies and *Ca.* Prosiliicoccus MAG sequences were deposited in ENA using the data brokerage service of the German Federation for Biological Data (GFBio) [14] in compliance with the Minimal Information about any (X) Sequence (MIxS) standard [60], and are available under the INSDC project number PRJEB28156. Anvi'o databases for individual MAGs, and the sorted cell mini-metagenome assembly, sorted MAGs, and sorted cell metagenome anvi'o database are available at doi:10.6084/m9.figshare.6139730.

Results

Metagenomic sequencing

Summary information for the metagenomic datasets from the dates 03/03/2010; 08/04/2010; 04/05/2010; 18/05/2010; 24/03/2011; 28/04/2011; 26/05/2011; 08/03/2012; 16/04/2012; and 10/05/2012 have been reported previously [50], and are reproduced in the data in Supplementary Table S4 along with summary statistics covering the rest of the metagenomic samples.

Metagenome assembly and binning

General assembly statistics for each metagenomic dataset are presented in Supplementary Table S4. The binning and MAG reassembly process produced the 40 environmental metagenome derived MAGs and seven sorted cell mini-metagenome derived MAGs described in Supplementary Table S5. The MAGs divided into three clusters we consider to represent three distinct species (Figs. 1 and 2a), as determined by both phylogenomic placement and average nucleotide identity. Redundancy of single copy marker genes of 3% or below for all MAGs indicates low levels of contamination. Estimates of completeness and redundancy as calculated by CheckM, anvi'o, and the HMM.essential.rb script of the enveomics collection are also included in Supplementary Table S5. Approximately 75% of mini-metagenome reads mapped back to the mini-metagenome derived *Ca.* Prosiliicoccus MAG sequences at 97% identity, with a ratio of approximately 770:120:1 between *Ca.* Pv and the second and third species respectively.

The reassembled MAGs showed improvements in total length and completeness without any increase in redundancy (see Supplementary Table S5). The reassembled MAG of *Ca.* Pv has completeness estimated at between 92% and 96%, with redundancy between 0.4% and 0.7% and total length of 1.9 Mbp. The reassembled MAG of *Ca.* Prosiliicoccus species 2 (*Ca.* P2) is between 69% and 91% complete with 0–0.6% contamination and length of 1.9 Mbp,

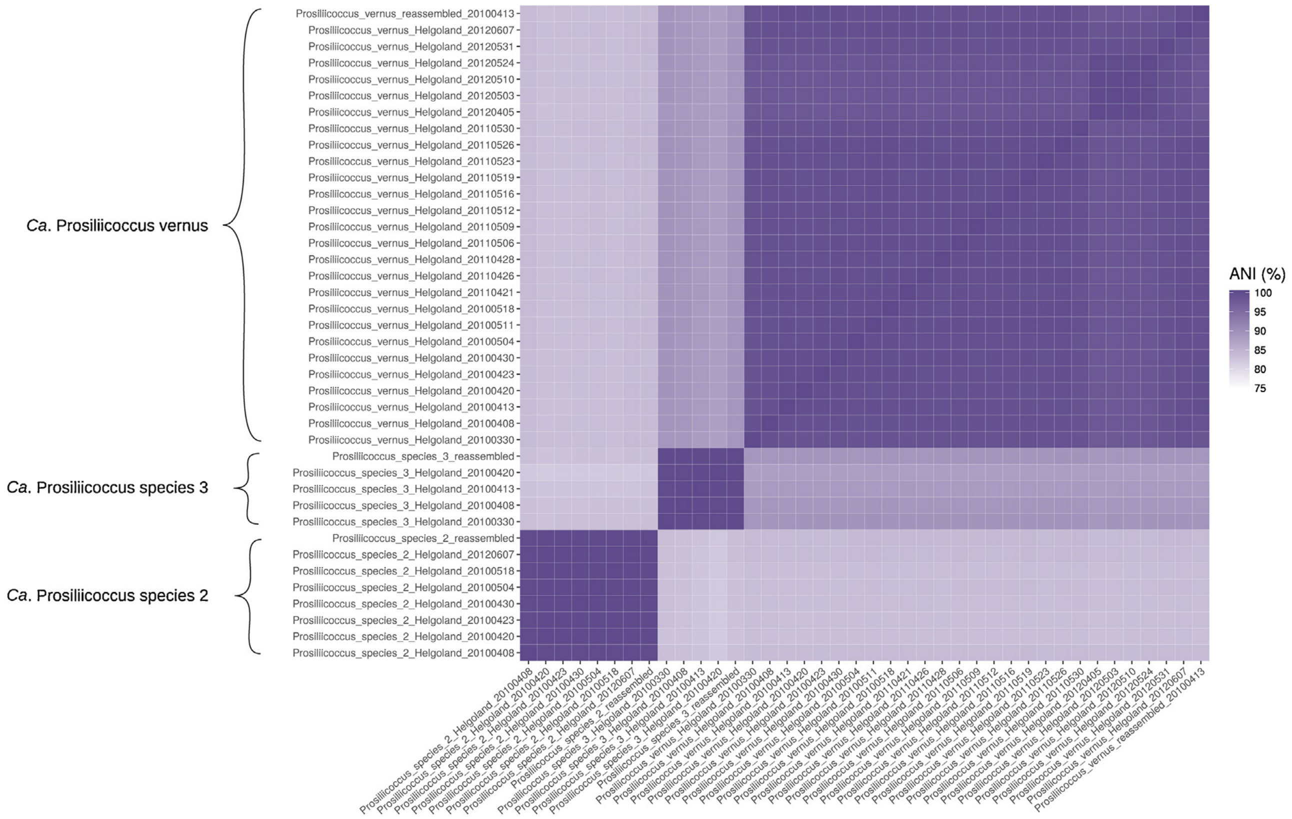

Fig. 1. Average nucleotide identity between *Candidatus* Prosiliicoccus MAGs showing separation into three species.

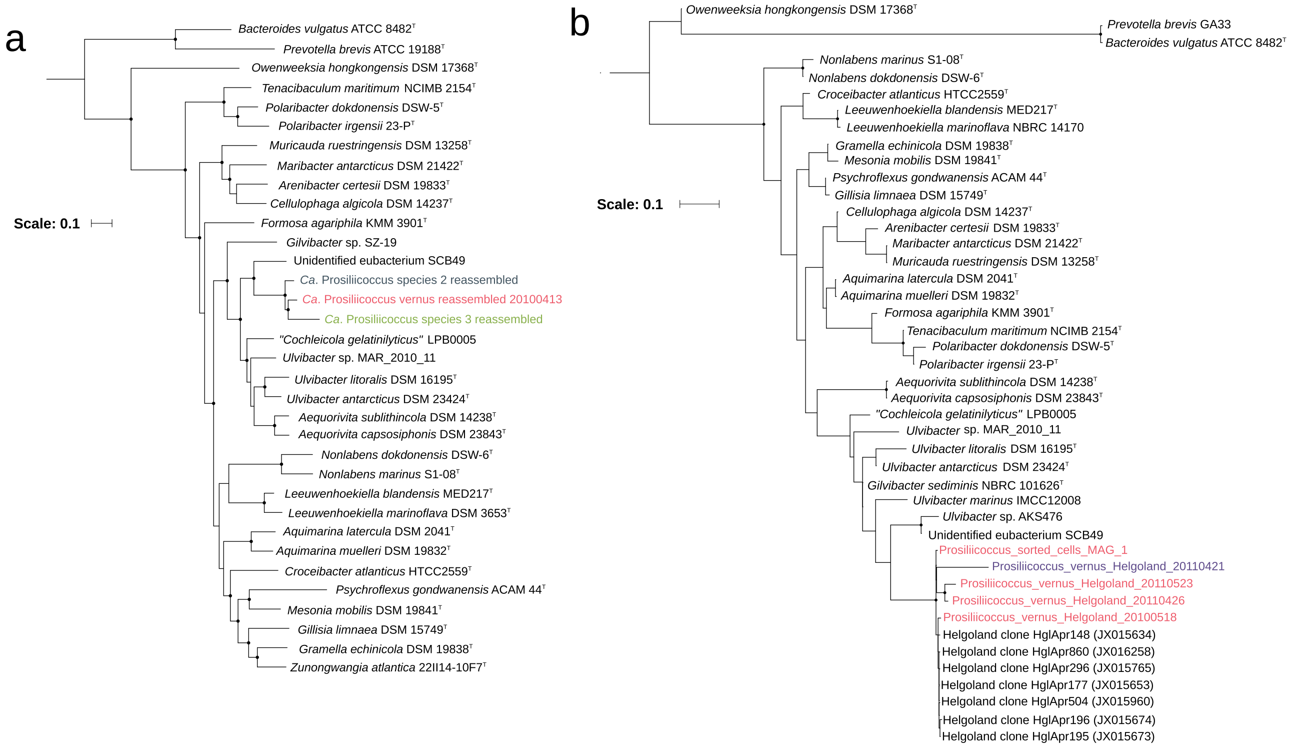
a
Bacteroides vulgatus ATCC 8482^T
Prevotella brevis ATCC 19188^T
Owenweeksia hongkongensis DSM 17368^T
Tenacibaculum maritimum NCIMB 2154^T
Polaribacter dokdonensis DSW-5^T
Polaribacter irgensii 23-P^T
Muricauda ruestringensis DSM 13258^T
Maribacter antarcticus DSM 21422^T
Arenibacter certesii DSM 19833^T
Cellulophaga algicola DSM 14237^T
Scale: 0.1
Formosa agariphila KMM 3901^T
Gilvibacter sp. SZ-19
Unidentified eubacterium SCB49
Ca. Prosiliicoccus species 2 reassembled
Ca. Prosiliicoccus vernus reassembled 20100413
Ca. Prosiliicoccus species 3 reassembled
"Cochleicola gelatinilyticus" LPB0005
Ulvibacter sp. MAR_2010_11
Ulvibacter litoralis DSM 16195^T
Ulvibacter antarcticus DSM 23424^T
Aequorivita sublithincola DSM 14238^T
Aequorivita capsosiphonis DSM 23843^T
Nonlabens dokdonensis DSW-6^T
Nonlabens marinus S1-08^T
Leeuwenhoekiella blandensis MED217^T
Leeuwenhoekiella marinoflava DSM 3653^T
Aquimarina latercula DSM 2041^T
Aquimarina muelleri DSM 19832^T
Croceibacter atlanticus HTCC2559^T
Psychroflexus gondwanensis ACAM 44^T
Mesonia mobilis DSM 19841^T
Gillisia limnaea DSM 15749^T
Gramella echinicola DSM 19838^T
Zunongwangia atlantica 22II14-10F7^T
b
Owenweeksia hongkongensis DSM 17368^T
Prevotella brevis GA33
Bacteroides vulgatus ATCC 8482^T
Nonlabens marinus S1-08^T
Nonlabens dokdonensis DSW-6^T
Croceibacter atlanticus HTCC2559^T
Leeuwenhoekiella blandensis MED217^T
Leeuwenhoekiella marinoflava NBRC 14170
Gramella echinicola DSM 19838^T
Mesonia mobilis DSM 19841^T
Psychroflexus gondwanensis ACAM 44^T
Gillisia limnaea DSM 15749^T
Scale: 0.1
Cellulophaga algicola DSM 14237^T
Arenibacter certesii DSM 19833^T
Maribacter antarcticus DSM 21422^T
Muricauda ruestringensis DSM 13258^T
Aquimarina latercula DSM 2041^T
Aquimarina muelleri DSM 19832^T
Formosa agariphila KMM 3901^T
Tenacibaculum maritimum NCIMB 2154^T
Polaribacter dokdonensis DSW-5^T
Polaribacter irgensii 23-P^T
Aequorivita sublithincola DSM 14238^T
Aequorivita capsosiphonis DSM 23843^T
"Cochleicola gelatinilyticus" LPB0005
Ulvibacter sp. MAR_2010_11
Ulvibacter litoralis DSM 16195^T
Ulvibacter antarcticus DSM 23424^T
Gilvibacter sediminis NBRC 101626^T
Ulvibacter marinus IMCC12008
Ulvibacter sp. AKS476
Unidentified eubacterium SCB49
Prosiliicoccus_sorted_cells_MAG_1
Prosiliicoccus_vernus_Helgoland_20110421
Prosiliicoccus_vernus_Helgoland_20110523
Prosiliicoccus_vernus_Helgoland_20110426
Prosiliicoccus_vernus_Helgoland_20100518
Helgoland clone HglApr148 (JX015634)
Helgoland clone HglApr860 (JX016258)
Helgoland clone HglApr296 (JX015765)
Helgoland clone HglApr177 (JX015653)
Helgoland clone HglApr504 (JX015960)
Helgoland clone HglApr196 (JX015674)
Helgoland clone HglApr195 (JX015673)

Fig. 2. Phylogeny of *Ca.* Prosiliicoccus. (**a**) Phylogenomic reconstruction based on 40 concatenated proteins (Supplementary Table S2). (**b**) 16S rRNA gene based phylogenetic reconstruction. Both phylogenies were calculated using RAxML with 1000 rapid bootstrap replicates. Black dots on nodes indicate greater than 90% bootstrap support.

while the reassembled MAG of *Ca.* Prosiliicoccus species 3 (*Ca.* P3) is between 42% and 64% complete with 0% redundancy and length of 1.35 Mbp. Note that CheckM gave the reassembled MAG of *Ca.* P2 a completeness score of above 90%, which may be considered a threshold for completeness upon which description of a candidate species may be based, while the other two approaches to measure completeness did not. Contamination/redundancy estimates were effectively nil in both CheckM (0.63%) and anvi'o (0%). Since the consensus between metrics was not in favour of the MAG being near complete however, the MAG is not formally described here as a candidate species.

Average nucleotide identities between MAGs and reassembled MAGs for the three *Ca.* Prosiliicoccus species are shown in Fig. 1 and Supplementary Table S6, indicating clear genomic divergence between the three groups. ANI between *Ca.* P2 and the other two species was approximately 84%, while ANI between *Ca.* P3 and *Ca.* Pv was higher at approximately 90%.

While MAGs 1 and 2 from the sorted cell mini-metagenome clearly belong to *Ca.* Pv and *Ca.* P2 respectively, as was predicted by Mash, sorted MAG 7 has ANI of greater than 98% to *Ca.* P3, while 3, 6, 9, and 10 all appear to belong within the genus (Supplementary Fig. S1). The clear caveat with these MAGs is that they each had lengths below 500 kbp and were not complete enough to be clustered by Mash. The low completeness of these MAGs is expected to be the result of the MDA not amplifying all parts of genomes equally. Despite this, the sorted cell mini-metagenome derived MAGs confirm the connection between 16S rRNA gene sequences, the ULV995 probe, and our metagenome derived MAGs.

Genomic divergence between the *Ca.* Pv MAGs was higher than that detected between the two lower abundant species, implying greater strain diversity in the *Ca.* Pv population. This can be seen in the higher proportion of genomic fragments with lower ANI (Supplementary Fig. S2) and the lower ANI values between whole MAGs (minimum value 97.7% between the *Ca.* Pv MAGs from 30/03/2010 and 05/04/2012, as compared to minima above 99% within species 2 and species 3 MAG sets), as can be seen in Supplementary Fig. S1 and Supplementary Table S6. It is also apparent that there is a change in the *Ca.* Pv population between non-bloom and bloom time periods, as demonstrated by the lower ANI between MAGs from earlier in 2012 and those assembled from the final two samples from 2012, which coincided with the onset of the *Ca.* Pv bloom (ANI between first and final *Ca.* Pv MAG from 2012 = 98.0%, compared to ANI of above 99% between the first four MAGs from 2012, and similarly between the final pair). The pre-bloom population from 2012 is also seemingly more genetically homogeneous than the bloom populations from 2010, 2011 and 2012, implying a change in population composition during progression of the 2012 bloom (i.e. the first four MAGs from 2012 have ANI above 99.7% with one another, compared with lower values between MAGs from consecutive sampling dates from the previous years).

ANI and AAI between the reassembled *Ca.* Prosiliicoccus MAGs and reference genomes of related species are shown in Supplementary Fig. S3 and Supplementary Table S7, indicating that *Ca.* Prosiliicoccus species belong to a separate genus with higher ANI and AAI within the genus than between *Ca.* Prosiliicoccus and other genera. ANI between MAGs within the genus *Ca.* Prosiliicoccus is approximately 84% and above, compared to values below 80% when compared to genomes of other species. AAI values within the genus are approximately 88% and above, compared to below 70% when compared with reference taxa.

Four of the *Ca.* Pv MAGs contained near full length 16S rRNA genes (Prosiliicoccus_vernus_Helgoland_20110426, Prosiliicoccus_vernus_Helgoland_20110523, Prosiliicoccus_vernus_Helgoland_20100518, and Prosiliicoccus_vernus_Helgoland_20110421), as did Prosiliicoccus_sorted_MAG_1. A further three (Prosiliicoccus_vernus_Helgoland_20100413, Prosiliicoccus_vernus_Helgoland_20110509, and Prosiliicoccus_vernus_Helgoland_20110530) contained other parts of the rRNA operon. From visual inspection of the coverage of contigs containing these operons, there appears to be approximately twofold higher coverage of the rRNA operon than the rest of the contig on which they sit, implying the presence of most likely two rRNA operons in this organism.

Phylogenomic and 16S rRNA phylogenetic reconstruction

Analysis of 40 concatenated, phylogenetically conserved protein sequences indicates the sister group relationship between *Ca.* Prosiliicoccus and *Ulvibacter*, as well as the presence of three distinct species within the genus *Ca.* Prosiliicoccus (Fig. 2a). This is consistent with the 16S rRNA gene based phylogeny, which recovers the same relationship between *Ulvibacter* and *Ca.* Prosiliicoccus (Fig. 2b). Equally, both methods are clear on the placement of *Ca.* Prosiliicoccus in the family *Flavobacteriaceae*, order *Flavobacteriales*, and class *Flavobacteriia* in the phylum *Bacteroidetes*. As is evident from the ANI data, the phylogenomic reconstruction confirms that *Ca.* Pv and *Ca.* P3 are more closely related to one another than either is to *Ca.* P2.

16S rRNA gene identity, typically used to determine the taxonomic level of divergence between clades, also confirms that *Ca.* Prosiliicoccus belongs to a novel genus, as identity between the full length *Ca.* Pv 16S rRNA sequence (using that derived from MAG Prosiliicoccus_vernus_Helgoland_20100518 as representative) and those of *Gilvibacter sediminis* NBRC 101626 [25] (94.1%), *Ulvibacter antarcticus* DSM 23424^T [12] (93.7%), and *Ulvibacter litoralis* DSM 16195^T [35] (94.7%) lie close to and below the threshold for delineating genera as recommended by Yarza et al. [59]. The similarity between *Ca.* Pv and Unidentified eubacterium SCB49 (sometimes referred to as *Ulvibacter* sp. SCB49) is 94.9%, thus also at the lower bound of belonging to genus *Ca.* Prosiliicoccus, with both the phylogenomic and 16S rRNA based reconstructions placing it as the closest relative of the three *Ca.* Prosiliicoccus species presented here.

The 16S rRNA gene assembled in the MAG Prosiliicoccus_vernus_Helgoland_20110421 only shares identity of 97% with those from the other *Ca.* Pv MAGs, and it is most likely that this sequence derives from misassembly of the gene because the coverage profile of the rest of the contig on which this gene is found is consistent with *Ca.* Pv. It appears from the 16S rRNA phylogeny that the genus *Ulvibacter* may also be paraphyletic, with *Ulvibacter marinus* IMCC 12008^T [4] belonging to a sister group to the other two described *Ulvibacter* species. Similarly the isolate *Ulvibacter* sp. MAR_2010_11 is likely a member of another genus, as is demonstrated in both phylogenies.

Cell morphology

In epifluorescence microscopic images of *Ca.* Prosiliicoccus, identified using CARD-FISH with the ULV995 16S rRNA probe, cells appear coccoid (Supplementary Fig. S4). In contrast, all species of *Ulvibacter* so far described are rods [4,12,35]. *Ca.* Prosiliicoccus cells have a size range of approximately 0.5-1 μm in diameter (Supplementary Table S8) with the mean of all cells measured being 0.78 μm, and a standard deviation of 0.12 μm.

Estimates of environmental abundance

In the years 2010, 2011, and 2012, *Ca.* Prosiliicoccus species reached high abundances during and after phytoplankton blooms (Fig. 3). Rapid doubling times are possible for the *Ca.* Prosiliicoccus population; a minimum doubling time of less than one day, implied by greater than 100% daily increases in cell number, can be seen at certain time points in Fig. 3a and Supplementary Table S3. This data refers to the population of the entire genus however, given that the ULV995 probe targets all three species.

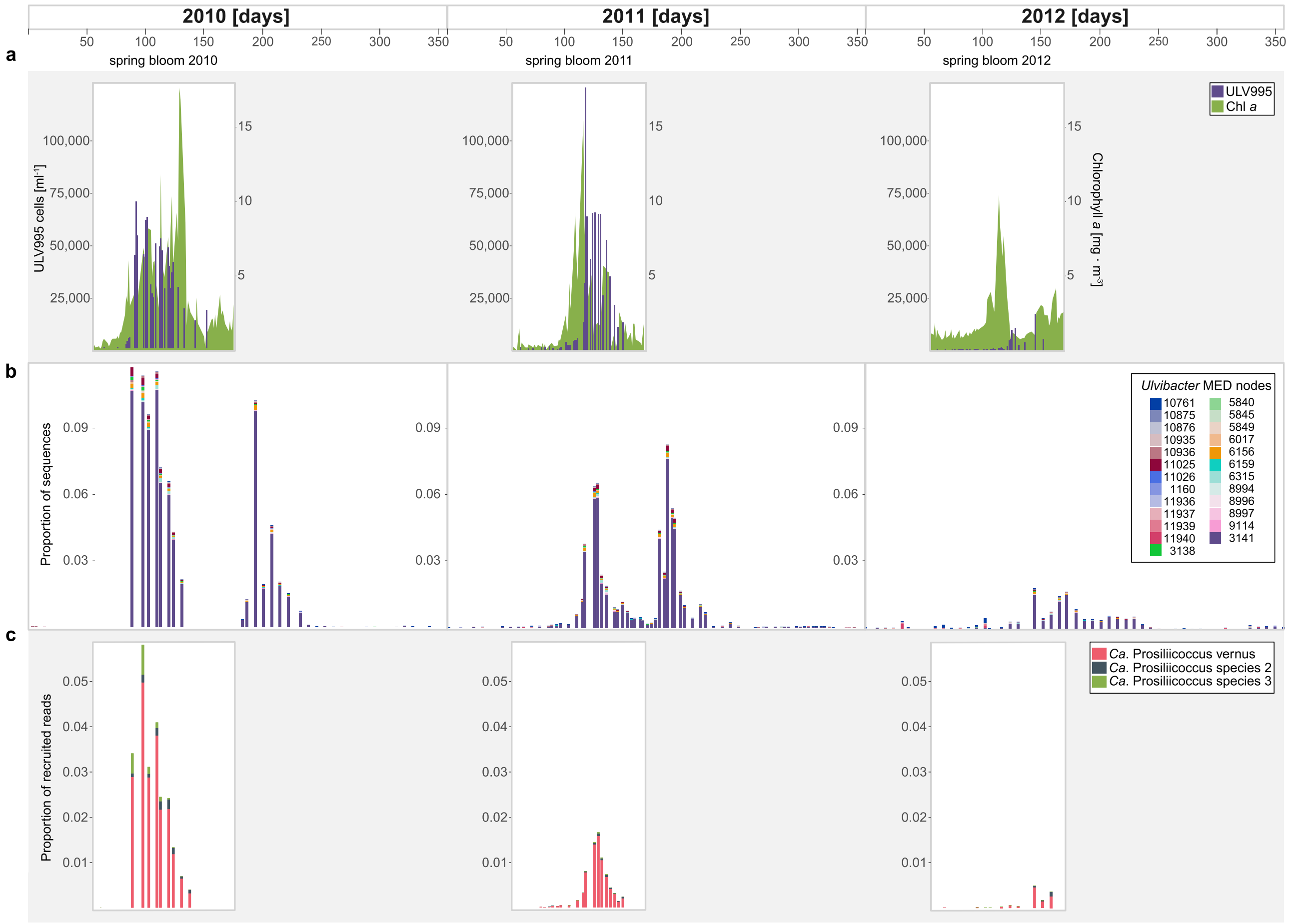

Fig. 3. Estimates of environmental abundance of *Ca.* Prosiliicoccus in the years 2010–12 at Helgoland during spring blooms, based on (**a**) CARD-FISH cell counts and Chlorophyll a data from Teeling et al. [50]; (**b**) proportion of amplicon sequences classified as *Ulvibacter* in the data from Chafee et al. [11]; (**c**) proportion of reads from each of the 38 metagenomic datasets recruited to the reassembled *Ca.* Prosiliicoccus MAGs.

From the amplicon data from Chafee et al. [11], it is apparent that at least in 2010 and 2011, *Ca.* Prosiliicoccus populations were prevalent during both spring and summer phytoplankton blooms at Helgoland, suggesting that conditions that favour this clade are not restricted to the springtime (Fig. 3b). This is also backed up by the global pattern of detection (Fig. 4), which demonstrates that sequences with high identity to the *Ca.* Pv 16S rRNA gene have been detected in regions such as the Benguela upwelling system off the coast of Namibia, and at a site in the Southern Ocean where an artificial iron seeded phytoplankton bloom was generated [55], as well as the seasonal temperate northern hemisphere locations where it can be seen in high abundance (Supplementary Table S9). Additionally, sequences have been detected in lower abundance across a number of sites in both the northern and southern hemispheres, demonstrating the ubiquity of this clade in temperate and polar regions. It is likely these data refer to the genus *Ca.* Prosiliicoccus as a whole, as based on our inability to distinguish the three species in the amplicon data of Chafee et al. [11], we might expect standard 16S rRNA gene amplicon datasets to capture a region of this gene conserved across the three species.

The metagenomic read recruitment data from Helgoland shows a similar pattern of abundance of the three *Ca.* Prosiliicoccus species to the other two datasets (Fig. 3c). It is apparent that *Ca.* Pv is the dominant *Ca.* Prosiliicoccus species during spring blooms, and based on the differential in average sequencing depth between *Ca.* Pv and the two lower abundant *Ca.* Prosiliicoccus species, it most likely makes up the majority of the *Ca.* Prosiliicoccus community detected by FISH and amplicon sequencing.

Within species variation in Ca. *Prosiliicoccus populations*

While sequencing depth was not sufficient for meaningful information to be gleaned regarding the two lower abundant *Ca.* Prosiliicoccus species in 2011 and 2012, variation within all three species in 2010 could be examined using SNV profiling. The detection of different strains within populations represented by individual MAGs was also attempted. As was seen when comparing ANI between MAGs, variability was higher within the *Ca.* Pv population than within the populations of the other two species (Supplementary Fig. S5), with the number of variable nucleotide positions per kilobase pair typically between 5 and 10, as compared to less than 3 in species 2 and species 3. The exception is the pre-bloom phase of 2012, where, as was seen in the ANI data, detected variability was lower in the *Ca.* Pv population. Strain deconvolution produced only inconclusive results. DESMAN's built in heuristic for strain number prediction suggested that *Ca.* Pv comprises 5 confidently predicted strains. However, visual inspection of the mean posterior deviance (Supplementary Fig. S6), which should stop decreasing once the true number of strains has been modelled, suggests that even at 14 strains the curve was only beginning to look as though it might plateau. This suggests the variability in this population is either genuinely very high, or it is in some way inconsistent with the underlying principles of the DESMAN algorithm. When using the run predicted as optimal by DESMAN (9 haplotypes used, 5 of which were confidently predicted, with average error of 5%) to examine strain relative abundance in the *Ca.* Pv population however, it appears that the strains that are predicted have

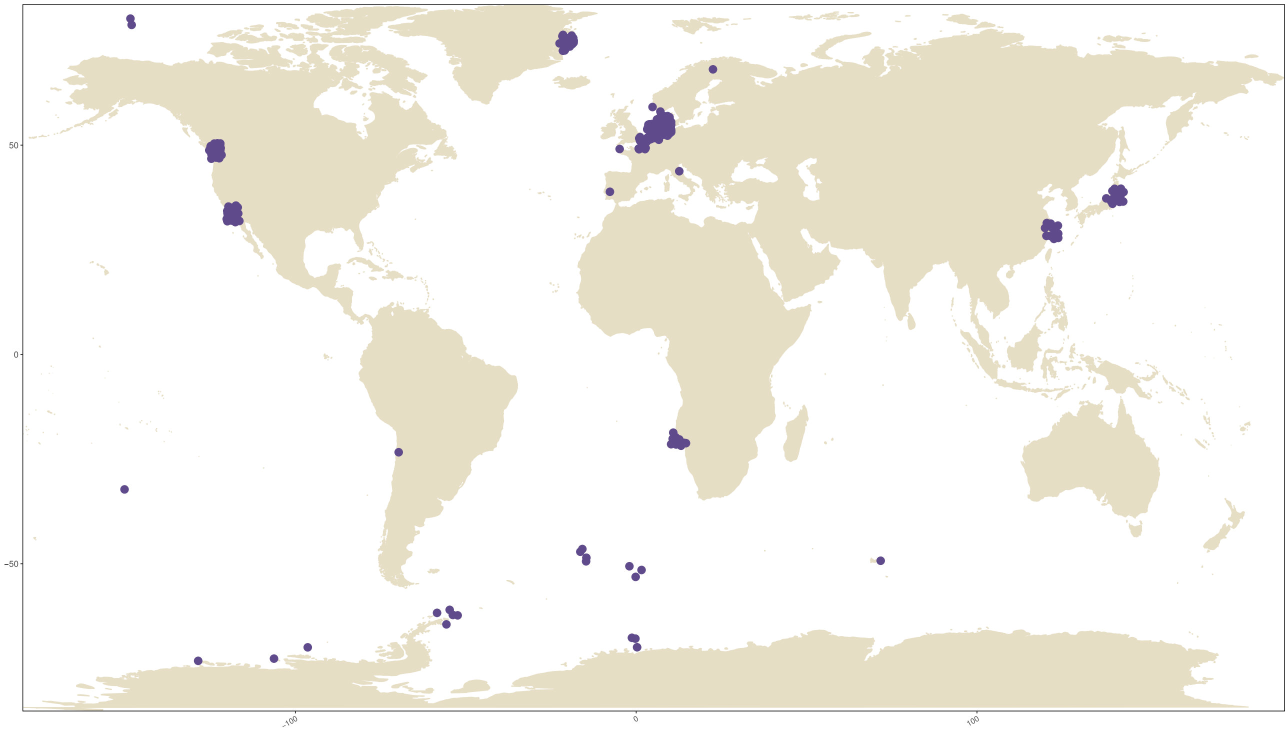

Fig. 4. Global distribution of amplicons showing greater than 99% identity to the 16S rRNA gene of *Ca.* Prosiliicoccus vernus. Points indicate detection, with an arbitrary cutoff of at least 50 reads from the sample matching the query sequence. The sample in the Pacific Ocean derives from a sponge metagenome (*Scopalina* sp.; BioProject accession PRJNA292036). Points are jittered to demonstrate where multiple sampling of the same location has taken place.

consistent abundance patterns across the bloom periods, implying a deterministic separation of function between strain-like populations of *Ca.* Pv despite the high noise (Supplementary Fig. S7). This consistency between years can also be seen when using data from runs with higher predicted strain numbers (data not shown). What is evident from Supplementary Fig. S7 is that haplotype H2 is dominant in 2010 and 2011 (~50% of the total *Ca.* Pv population), and as the *Ca.* Pv population increases in size through 2012, this haplotype increases to approach a similar proportion of the overall population. This suggests that haplotype H2 could be responding more strongly to the phytoplankton blooms than other haplotypes. The dominance of haplotype H3 before bloom onset in 2012 is also consistent with the homogeneity seen among MAGs from these dates based on ANI. There is also a noticeable shift in both 2010 and 2011 from H5 to H4 over time, again pointing potentially to deterministic rather than stochastic changes in population structure. These patterns are currently only observed in these 3 years however, and this conclusion is thus only tentative.

Annotation of the reassembled Ca. *Prosiliicoccus vernus MAG and inference of metabolic potential*

The reassembled *Ca.* Pv MAG contains 1810 predicted genes, 592 of which (33%) are annotated as hypothetical. Of these, 31 are tRNA genes, among which tRNA genes for aspartic acid are absent. tRNA genes for aspartic acid are however found in other *Ca.* Pv MAGs.

Basic energy conservation

The *Ca.* Pv MAG contains complete pathways for aerobic respiration comprising glycolysis, the non-oxidative phase of the pentose phosphate pathway, TCA cycle, and an electron transport chain (Fig. 5). There are no unambiguous indications of use of other monosaccharides than glucose, but the presence of various unspecified ABC transporters implies that different sugar monomers might also be taken up. The MAG also possesses 87 predicted proteases, some of which are expected to be secreted extracellularly (Supplementary Table S10), and predicted degradation pathways are present for the amino acids alanine, arginine, asparagine, cysteine, glutamine, histidine, isoleucine, lysine, methionine, phenylalanine, threonine, tryptophan, tyrosine, and valine. Additionally the MAG contains 10 co-located genes involved in phenylacetate degradation, which encode the multisubunit 1,2-phenylacetyl-CoA epoxidase (PaaABCDE) as well as PaaGHINY enzymes. There is also a gene for a short chain fatty acid transporter, long chain fatty acid ligase, and an alkane 1-monooxygenase gene, indicating basic processing of fatty acids either as an additional source of reduced carbon or for general metabolic purposes such as building of cell membranes.

The MAG contains genes for all 6 subunits of the Na^+-translocating NADH-quinone oxidoreductase, succinate dehydrogenase, and cytochrome c and cytochrome c oxidases of complex IV. In our data, there are no indications of adaptation for fermentation, or for use of alternative electron acceptors for anaerobic respiration.

The presence of a glycogen synthase gene suggests energy storage. Also present is a proteorhodopsin gene, which could be involved either in energy conservation, or otherwise may serve as part of a light based sensory system.

Sources of nitrogen, sulfur, phosphorous

The primary nitrogen source for *Ca.* Pv is expected to be protein and amino acids, whilst it also possesses an ammonium transporter. There is no indication of assimilatory sulfate reduction in the genome, so it is presumed that proteins are the primary sulfur source for this species. There are transporters present for both inorganic phosphate and organic phosphonates, and both polyphosphate kinase and exopolyphosphatase genes that imply storage of phosphate via polyphosphate.

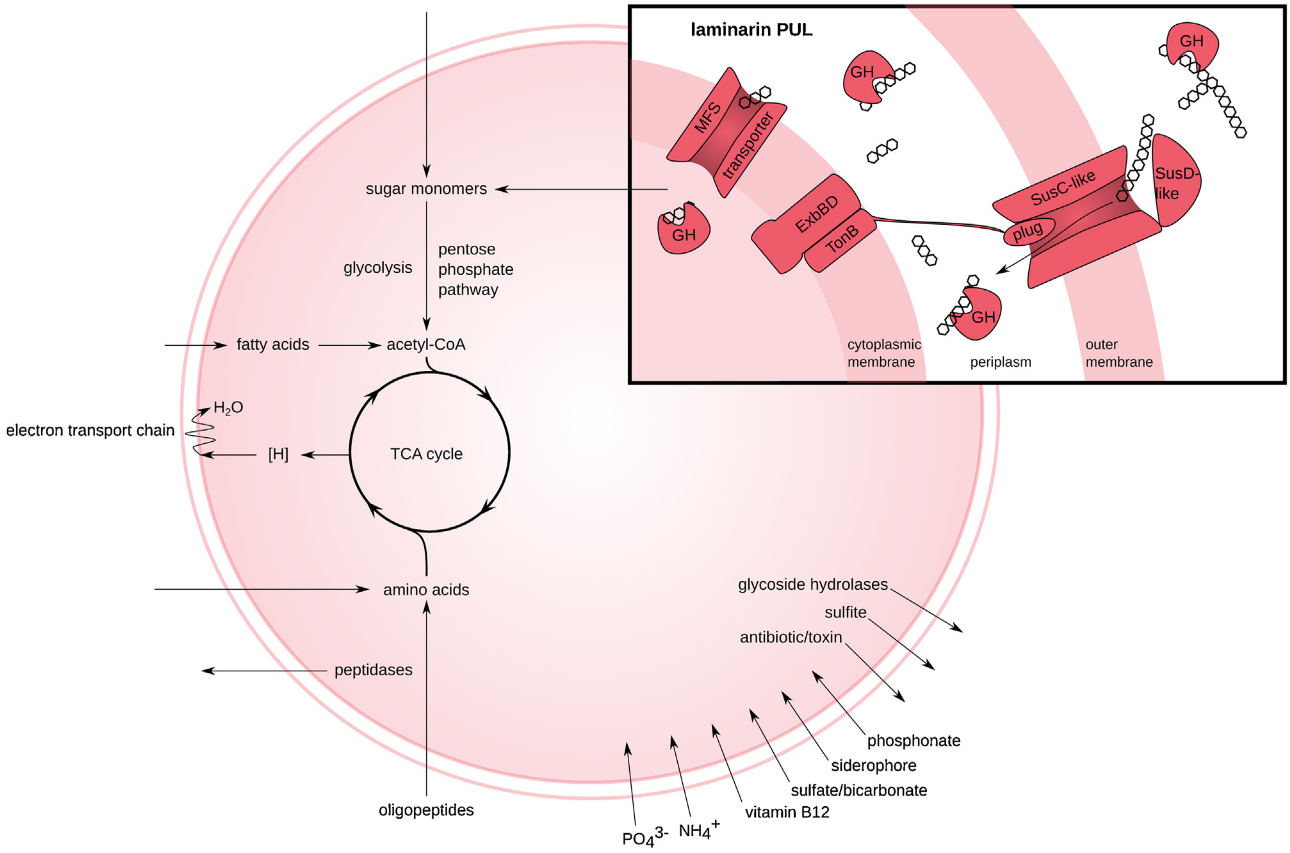

Fig. 5. General summary of predicted metabolic potential of *Ca.* Prosiliicoccus vernus. Energy conservation occurs via consumption of low molecular weight organic molecules such as small peptides and glycans, and the abundant polysaccharide laminarin. Other sugar monomers than glucose could not be definitively shown to be utilised. Light energy can also be conserved via the use of a proteorhodopsin. Otherwise *Ca.* Prosiliicoccus vernus is an obligate aerobe which uses Na^+ translocating NADH-quinone reductase as part of its respiratory chain.

Transport

Ion transporters are well represented in the reassembled *Ca.* Pv MAG, with transporters for copper, magnesium, zinc, sodium, potassium, cobalt, manganese, iron (including probable siderophore carrying TonB-dependent transporters), and bicarbonate/sulfate, in addition to the aforementioned ammonium transporter.

Also present are transporters belonging to the *gld* family found in many flavobacteria, and also to the type IX secretion system. Other genes with predicted export function include homologues of multidrug exporters and macrolide exporters, and a fluoride efflux transporter. Import functions include di/tripeptide transporters, an oligopeptide permease, the aforementioned fatty acid transporter, various TonB dependent transporters including SusC-like homologues, and ABC and MFS type transporters. These transporters can have diverse functions that are not readily predicted with regard to their substrate but are expected to include transporters of cobalamin, which *Ca.* Pv appears to be auxotrophic for. One interesting SusC/D-like pair in this genome is not co-located with any CAZymes, and thus does not have an inferred function in polysaccharide transport. Instead it sits in close proximity to a collection of ribosomal protein genes, a transcription elongation factor, and an outer membrane protein assembly factor. The function of this pair remains indeterminate.

CAZyme profile and predicted polysaccharide utilisation

Annotation of CAZymes and transporters indicated the presence of a single canonical polysaccharide utilisation structure. This PUL includes a SusC/D-like pair, and otherwise contains just a pair of glycoside hydrolases of the GH16 and GH3 families (Fig. 6), in close proximity to genes for ATP-dependent 6-phosphofructokinase, glyceraldehydephosphate dehydrogenase, and glucosephosphate isomerase (all involved in glycolysis). The GH16 enzyme is predicted to be extracellular, while the GH3 is predicted to be either periplasmic or extracellular (Supplementary Table S10). The most probable putative substrate for this PUL is laminarin, as most characterised GH3 family enzymes have β-glucosidase activity, and GH16 family members, while having more widely varying described functions, are known to include β-1,3-glucanase activity that would act on laminarin. Additional support for this comes from GH3- and GH16-containing PULs that have been experimentally confirmed to be laminarin-specific [23]. Identity between protein sequences of the GH16 in *Ca.* Prosiliicoccus vernus and its homologue in the experimentally confirmed PUL from *Gramella forsetii* KT0803^T [23] was 40%, and identity between GH3s was 58%. The fact that the proteins belong to the same families, and share similar neighbouring gene functions such as the SusC/D-like pair, support the prediction of laminarin as the substrate of this PUL, despite the fact that specific functional domains could not be assigned to the sequences. On the same contig, some 30 kbp removed from the SusC/D-like pair, the reassembled MAG contains a pair of GH17 enzymes, also known to be active on laminarin and to be part of laminarin active PULs [8]. The *Ca.* Pv MAG Prosiliicoccus_vernus_Helgoland_20100420 has a more complete assembly of this region of the genome, and possesses a cluster including three GH17 family genes (two cytoplasmic, the other undetermined), together with one GH30 and one GH2 gene (no consensus on localisation), and an MFS family glucan transporter that is localised to the cytoplasmic membrane (Supplementary Table S10). This is identical in gene content and order to that found in

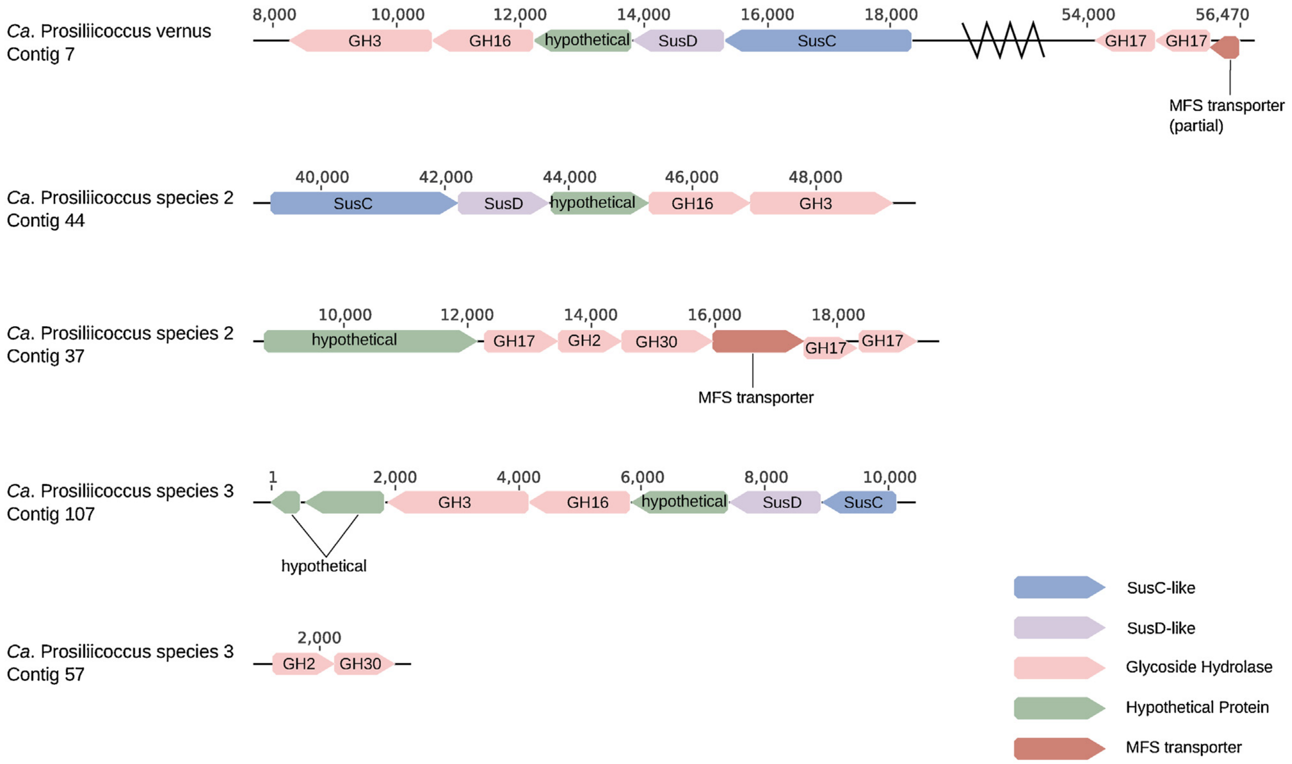

Fig. 6. Putative laminarin degrading PUL-like structures in reassembled *Ca.* Prosiliicoccus MAGs, showing conserved gene order among the three species. The two types of structure seen here, the SusC/D-like pair with GH16 and GH3, and the GH17/GH30 collections, are both similar to laminarin degrading PULs known in other North Sea species. Numbers above genes indicate position on contig in base pairs.

the *Ca.* P2 reassembled MAG contig 37 (Fig. 6). *Ca.* Pv also has a gene pair comprising a GH20 β-hexosaminidase and a GH2 exo-β-glucosaminidase. Glucosamine is typically found in chitin and chitosan, two other abundant marine polysaccharides. There is no gene for any N-acetylglucosamine transporter in close proximity to these genes in any of the *Ca.* Pv MAGs however.

Annotation of reassembled Ca. *P2 and P3 MAGs and partial inference of metabolic potential*

The reassembled *Ca.* P2 MAG contains 1765 predicted genes, with 690 of those predicted to be hypothetical, and 32 tRNAs, of which only tRNA genes for isoleucine and tryptophan are absent. The predicted metabolic capabilities are similar to that for *Ca.* Pv, with only the absence of cysteine and lysine degradation, and fatty acid metabolism being noteworthy. With respect to CAZyme profile, *Ca.* P2 has the same 3 CAZyme collections mentioned for *Ca.* Pv, namely the GH3, GH16, SusC/D-like pair and GH17, GH30 collections putatively involved in laminarin degradation shown in Fig. 6, and the GH2, GH20 pair plausibly involved in chitin utilisation.

The reassembled MAG of *Ca.* P3 contains 1306 predicted genes, including 536 hypothetical genes and 21 tRNAs. As with *Ca.* P2, the predicted metabolic pathways are similar to those for *Ca.* Pv, although with fewer complete pathways and more absent predicted pathways, as is to be expected from a less complete genome. The CAZyme and PUL profile of *Ca.* P3 differs from that of the other two species. While it has the same SusC/D-like pair, GH16, GH3 PUL and at least a hint of the second laminarin degrading cluster contained on a short contig with a GH30 and GH2 gene (Fig. 6), it also possesses a gene cluster containing a GH29 family enzyme and two sulfatases (one family S1_19 endo-carageenan function, the other family S1_25, unknown function), a PUL with SusC/D-like pair and two GH86 family β-porphyranases/β-agarases putatively degrading porphyran or agar, and a PUL that consists of a SusC/D-like pair, three predicted sulfatase genes (families S1_36 and S1_16 - unknown function, and family S1_11 – heparan/mucin function), a GH128 family gene, and two hypothetical proteins (Supplementary Fig. S8). The implication from these gene collections, despite the fact that they may not be completely assembled here, is that there exists a distinct substrate niche for this species that includes sulfated polysaccharides which are not available to the two other *Ca.* Prosiliicoccus species.

Discussion

The novel candidate genus *Candidatus* Prosiliicoccus, family *Flavobacteriaceae*, order *Flavobacteriales*, class *Flavobacteriia*, phylum *Bacteroidetes* presented here comprises three distinct species detected in metagenomic datasets deriving from samples collected during spring blooms in the North Sea at the island of Helgoland. The genus *Ca.* Prosiliicoccus comprises apparently obligate aerobic heterotrophs, which react to phytoplankton blooms. One species is more abundant during blooms than the other two, and near complete metagenome assembled genomes could be recovered that describe this population. Based on these and associated data presented here, we formally describe the candidate species *Candidatus* Prosiliicoccus vernus, according to the standards outlined by Konstantinidis et al. [26,27].

Candidatus Prosiliicoccus vernus is capable of rapid growth to the extent that, based on FISH counts, population doubling times can approach and at times even fall below one day. Because of this growth, this species can react swiftly to phytoplankton blooms and transiently make up between 5–20% of the total free-living bacterioplankton population. Rapid growth may be in part facilitated by the smaller genome and concomitant small cell size when compared to other known *Flavobacteriia. Ca.* Pv most likely relies on a combination of protein, small peptide, and free amino acids as a primary source of carbon and nitrogen. Additionally it is likely that it consumes some form of the polysaccharide laminarin, a storage polysaccharide produced by diatoms and brown algae. Laminarin is released in large amounts during spring blooms at Helgoland, as a result of the high abundance of diatoms during these periods [50]. The use of laminarin by *Ca.* Prosiliicoccus vernus is therefore likely to be substantial. Global abundance patterns indicate this species is restricted to temperate and polar latitudes, but appears to respond to phytoplankton bloom events in many locations where they are known to occur. This implies a commonality between phytoplankton blooms that can be exploited by *Ca.* Prosiliicoccus populations, although the nature of this commonality is not yet determined.

What is unfortunately still unclear from our data, is what precisely permits *Ca.* Prosiliicoccus vernus populations, and indeed perhaps only three strains of this species, to respond so strongly to the increases in algal abundance. This is a particularly challenging question to answer, given that the recovered gene content of the three *Ca.* Prosiliicoccus species is generally similar (corroborated by the high amino acid identity between the species), and that the temporal abundance pattern is similar for all three species despite the vast disparities in cell numbers. It is unsurprising that three species evidently capable of consuming laminarin and protein would respond to some extent to the massive increases in these substrates in the water column that occur as a result of phytoplankton blooms, thus some other mechanism not here readily determinable is necessary to explain the size of *Ca.* Pv populations during the spring at Helgoland. The existence of homologues of nutrient uptake systems (specifically the phosphate and ammonia transporters, and the polyphosphate kinase and exopolyphosphatase) in *Ca.* P2 suggests that presence of these gene functions alone is also unlikely to be allowing the high growth rates of *Ca.* Pv populations.

Compared to the closely related genera *Ulvibacter, Aequorivita, Altibacter* and "*Cochleicola*", and the genome of Unidentified eubacterium SCB49, *Candidatus* Prosiliicoccus vernus has several notable differences. Firstly the genome is much reduced in size, at 1.9 Mbp compared to the typical 3–4 Mbp genomes of close relatives. Cells are also coccoid rather than rod shaped, which sets them apart not only from close relatives, but also from other free living *Flavobacteriia* known to be abundant in the North Sea such as members of the genera *Formosa* and *Polaribacter.* It has been suggested that reduced cell size and coccoid morphology can have adaptive benefits in evading grazers [40], which might aid the rapid growth capacity of *Ca.* Pv. The putatively reduced capacity for degradation of complex polysaccharides, when compared to many *Flavobacteriia,* including others known to respond to phytoplankton blooms in the North Sea, is an additional distinguishing feature of *Ca.* Pv.

The multiple lines of evidence presented here, including divergence of 16S rRNA gene sequences, divergence of conserved single copy genes, genomic distinctness in terms of ANI and AAI, and the recovery and description of a near complete metagenome assembled genome, all support the description of a novel species and genus within the *Flavobacteriaceae*.

Description of Candidatus *Prosiliicoccus*

Candidatus Prosiliicoccus (Pro.si.li.i.coc'cus. L. v. *prosilio, prosilire, prosilui* to leap, jump, rush, spring forth; N.L. mas. n. *coccus* from Gr. mas. n. *kokkos* grain, seed; N.L. mas. n. *Prosiliicoccus*).

Members of the genus *Ca.* Prosiliicoccus are predicted to be obligate aerobes, with currently no indication of fermentation or anaerobic respiration. They are heterotrophic, marine surface water dwelling bacteria, capable of using glycans and proteins as primary sources of organic matter. The three species have been detected in seawater sampled in the North Sea during spring phytoplankton blooms via metagenomic assembly and fluorescence in situ hybridisation. Cells are coccoid, and may be detected using FISH probe ULV995 [49]. High quality 16S rRNA gene sequences share approximately 94% identity with closely related genera *Ulvibacter* and *Gilvibacter.* G + C content for all three species is between 36% and 37%. The genus *Ca.* Prosiliicoccus belongs to the family *Flavobacteriaceae*, order *Flavobacteriales*, class *Flavobacteriia*, and phylum *Bacteroidetes*. Type species is *Candidatus* Prosiliicoccus vernus.

Description of Candidatus *Prosiliicoccus vernus*

Ca. Prosiliicoccus vernus (ver'nus. L. mas. adj. v*ernus* pertaining to spring, vernal).

Genome annotation allows prediction of consumption of protein, peptides, and amino acids, as well as putatively the polysaccharide laminarin. The spectrum of glycans putatively available to *Candidatus* Prosiliicoccus vernus is restricted, with laminarin appearing to be the most significant polysaccharide, while it may also consume chitin. Genome size is small relative to described *Flavobacteriia,* at an estimated 1.9 Mbp. Fluorescence microscopy reveals the cells to be coccoid, with diameter ranging between 0.5 and 1 μm. Bloom-forming behaviour is observed in *Candidatus* Prosiliicoccus vernus during and immediately after phytoplankton blooms in the North Sea.

Type material is the metagenome assembled genome 'Prosiliicoccus_vernus_reassembled_20100413' submitted to ENA in project PRJEB28156, and also to the Digital Protologue database under TaxoNumber CA00022. Together the data presented here fulfil all of the criteria required for description of uncultivated prokaryotic taxa outlined by Konstantinidis et al. [27].

Digital Protologue for *Candidatus* Prosiliicoccus vernus

Taxonumber	CA00022
Species name	Prosiliicoccus vernus
Genus name	Prosiliicoccus
Specific epithet	vernus
Genus etymology	Prosiliicoccus (Pro.si.li.i.coc'cus. L. v. prosilio, prosilire, prosilui to leap, jump, rush, spring forth N.L. mas. n. coccus (from Gr. mas. n. kokkos grain, seed N.L. mas. n. Prosiliicoccus)
Genus status	gen. nov.
Species etymology	vernus (ver'nus. L. mas. adj. vernus pertaining to spring, vernal)
Species status	sp. nov.
Authors	Francis TB, Krüger K, Fuchs BM, Teeling H, Amann RI
Title	*Candidatus* Prosiliicoccus vernus, a spring phytoplankton bloom associated member of the *Flavobacteriaceae*
Journal	Systematic and Applied Microbiology
Corresponding author	Amann RI
E-mail of the corresponding author	ramann@mpi-bremen.de
Submitter	
E-mail of the submitter	tfrancis@mpi-bremen.de
Designation of the type MAG	Prosiliicoccus_vernus_reassembled_20100413
Metagenome accession number MAG	PRJEB28156
Genome size	1927
GC mol%	36.88
Country of origin	DEU
Region of origin	Helgoland
Source of sample	ENVO_00002150 coastal seawater
Sampling date	2010-04-13
Geographic location	Helgoland
Latitude	54° 11’ 17.9988” N
Longitude	7° 54’ 0” E
Relationship to O2	Aerobe
Energy metabolism	Chemoorganotroph
Assembly	1 sample
Sequencing technology	Illumina HiSeq 2500
Binning software used	CONCOCT, anvi'o v3
Assembly software used	SPAdes v3.10
Habitat	ENVO_00002150 coastal seawater
Biotic relationship	Free-living

Acknowledgements

The authors acknowledge the support of the Max Planck Society. TBF, BMF, HT, and RIA acknowledge funding from German Research Foundation (DFG) project FOR 2406 – 'Proteogenomics of Marine Polysaccharide Utilisation (POMPU)'. Metagenome sequencing was conducted in the framework of the COGITO project (Contract No. DE-AC02-05CH11231) by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, and is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. The authors acknowledge the contribution of Tijana Glavina del Rio as project leader of metagenome sequencing at JGI on the COGITO project. We also thank Aharon Oren for his expertise and input in matters of nomenclature, and Ivaylo Kostadinov of GFBio for support with deposition of sequence data.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.syapm.2018.08.007.

References

[1] Alneberg, J., Bjarnason, B.S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., Quince, C. (2014) Binning metagenomic contigs by coverage and composition. Nat. Methods 11, 1144–1146.

[2] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

[3] Andrews, S. 2010 FastQC: a Quality Control Tool for High Throughput Sequence Data.

[4] Baek, K., Jo, H., Choi, A., Kang, I., Cho, J.C. (2014) *Ulvibacter marinus* sp. nov., isolated from coastal seawater. Int. J. Syst. Evol. Microbiol. 64, 2041–2046.

[5] Barbeyron, T., Brillet-Guéguen, L., Carré, W., Carrière, C., Caron, C., Czjzek, M., Hoebeke, M., Michel, G. (2016) Matching the diversity of sulfated biomolecules: creation of a classification database for sulfatases reflecting their substrate specificity. PLoS One 11, e0164846.

[6] Barbeyron, T., Thomas, F., Barbe, V., Teeling, H., Schenowitz, C., Dossat, C., Goesmann, A., Leblanc, C., Glöckner, F.O., Czjzek, M., Amann, R., Michel, G. (2016) Habitat and taxon as driving forces of carbohydrate catabolism in marine heterotrophic bacteria: example of the model algae-associated bacterium *Zobellia galactanivorans* DsijT. Environ. Microbiol. 18, 4610–4627.

[7] Barrero-Canosa, J., Moraru, C., Zeugner, L., Fuchs, B.M., Amann, R. (2017) Direct-geneFISH: a simplified protocol for the simultaneous detection and quantification of genes and rRNA in microorganisms. Environ. Microbiol. 19, 70–82.

[8] Becker, S., Scheffel, A., Polz, M.F., Hehemann, J.H. (2017) Accurate quantification of laminarin in marine organic matter with enzymes from marine microbes. Appl. Environ. Microbiol. 83, e03389–16.

[9] Bennke, C.M., Krüger, K., Kappelmann, L., Huang, S., Gobet, A., Schüler, M., Barbe, V., Fuchs, B.M., Michel, G., Teeling, H., Amann, R.I. (2016) Polysaccharide utilisation loci of Bacteroidetes from two contrasting open ocean sites in the North Atlantic. Environ. Microbiol. 18, 4456–4470.

[10] Buchan, A., LeCleir, G.R., Gulvik, C.A., González, J.M. (2014) Master recyclers: features and functions of bacteria associated with phytoplankton blooms. Nat. Rev. Microbiol. 12, 686–698.

[11] Chafee, M., Fernàndez-Guerra, A., Buttigieg, P.L., Gerdts, G., Eren, A.M., Teeling, H., Amann, R.I. (2017) Recurrent patterns of microdiversity in a temperate coastal marine environment. ISME J. 12, 237–252.

[12] Choi, T.H., Lee, H.K., Lee, K., Cho, J.C. (2007) *Ulvibacter antarcticus* sp. nov., isolated from Antarctic coastal seawater. Int. J. Syst. Evol. Microbiol. 57, 2922–2925.

[13] Deorowicz, S., Debudaj-Grabysz, A., Gudyś, A. (2016) FAMSA: fast and accurate multiple sequence alignment of huge protein families. Sci. Rep. 6, 33964.

[14] Diepenbroek, M., Glöckner, F., Grobe, P., Güntsch, A., Huber, R., König-Ries, B., Kostadinov, I., Nieschulze, J., Seeger, B., Tolksdorf, R., Triebel, D. (2014) Towards an Integrated Biodiversity and Ecological Research Data Management and Archiving Platform: The German Federation for the Curation of Biological Data (GFBio). In: Plödereder, E, Grunske, L, Schneider, E, Ull, D, (Eds.). Informatik 2014–Big Data Komplexität meistern. GI-Edition: Lecture Notes in Informatics (LNI) – Proceedings. GI edn. vol. 232. Köllen Verlag, Bonn, pp. 1711–1724.

[15] Eddy, S.R. (2011) Accelerated profile HMM searches. PLoS Comput. Biol. 7, e1002195.

[16] Eren, A.M., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., Delmont, T.O. (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3, e1319.

[17] Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A. (2015) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44, D279–D285.

[18] González, J.M., Fernández-Gómez, B., Fernàndez-Guerra, A., Gómez-Consarnau, L., Sánchez, O., Coll-Lladó, M., González, J.M., Fernández-Gómez, B., Fernàndez-Guerra, A., Gómez-Consarnau, L., Sánchez, O., Coll-Lladó, M., del Campo, J., Escudero, L., Rodríguez-Martínez, R., Alonso-Sáez, L., Latasa, M., Paulsen, I., Nedashkovskaya, O., Lekunberri, I., Pinhassi, J., Pedrós-Alió, C. (2008) Genome analysis of the proteorhodopsin-containing marine bacterium *Polaribacter* sp. MED152 (Flavobacteria). Proc. Natl. Acad. Sci. U. S. A. 105, 8724–8729.

[19] Grondin, J.M., Tamura, K., Déjean, G., Abbott, D.W., Brumer, H. (2017) Polysaccharide Utilization Loci: fuelling microbial communities. J. Bacteriol. 1, 00860–16.

[20] Haft, D.H., Selengut, J.D., Richter, R.A., Harkins, D., Basu, M.K., Beck, E. (2012) TIGRFAMs and genome properties in 2013. Nucleic Acids Res. 41, D387–D395.

[21] Hahnke, R.L., Bennke, C.M., Fuchs, B.M., Mann, A.J., Rhiel, E., Teeling, H., Amann, R., Harder, J. (2015) Dilution cultivation of marine heterotrophic bacteria abundant after a spring phytoplankton bloom in the North Sea. Environ. Microbiol. 17, 3515–3526.

[22] 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.

[23] Kabisch, A., Otto, A., König, S., Becher, D., Albrecht, D., Schüler, M., Teeling, H., Amann, R.I., Schweder, T. (2014) Functional characterization of polysaccharide utilization loci in the marine Bacteroidetes '*Gramella forsetii*' KT0803. ISME J. 8, 1492–1502.

[24] Karp, P.D., Latendresse, M., Paley, S.M., Krummenacker, M., Ong, Q.D., Billington, R., Kothari, A., Weaver, D., Lee, T., Subhraveti, P. (2015) Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. Brief. Bioinf. 17, 877–890.

[25] Khan, S.T., Nakagawa, Y., Harayama, S. (2007) *Sediminibacter furfurosus* gen. nov., sp. nov. and *Gilvibacter sediminis* gen. nov., sp. nov., novel members of the family Flavobacteriaceae. Int. J. Syst. Evol. Microbiol. 57, 265–269.

[26] Konstantinidis, K.T., Rosselló-Móra, R. (2015) Classifying the uncultivated microbial majority: a place for metagenomic data in the *Candidatus* proposal. Syst. Appl. Microbiol. 38, 223–230.

[27] Konstantinidis, K.T., Rosselló-Móra, R., Amann, R. (2017) Uncultivated microbes in need of their own taxonomy. ISME J. 11, 2399–2406.

[28] Lagkouvardos, I., Joseph, D., Kapfhammer, M., Giritli, S., Horn, M., Haller, D., Clavel, T. (2016) IMNGS: a comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. Sci. Rep. 6, 33721.

[29] Letunic, I., Bork, P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44 (W1), W242–W245.

[30] Lindh, M.V., Sjöstedt, J., Andersson, A.F., Baltar, F., Hugerth, L.W., Lundin, D., Muthusamy, S., Legrand, C., Pinhassi, J. (2015) Disentangling seasonal bacterioplankton population dynamics by high-frequency sampling. Environ. Microbiol. 17, 2459–2476.

[31] Lucas, J., Wichels, A., Teeling, H., Chafee, M., Scharfe, M., Gerdts, G. (2015) Annual dynamics of North Sea bacterioplankton: seasonal variability superimposes short-term variation. FEMS Microbiol. Ecol. 91, fiv099.

[32] Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhu, K., Buchner, A., Lai, T., Steppi, S., Jobb, G., Förster, W., Brettske, I., Gerber, S., Ginhart, A.W., Gross, O., Grumann, S., Hermann, S., Jost, R., König, A., Liss, T., Lüßmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., Schleifer, K.H. (2004) ARB: a software environment for sequence data. Nucleic Acids Res. 32, 1363–1371.

[33] Martens, E.C., Koropatkin, N.M., Smith, T.J., Gordon, J.I. (2009) Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm. J. Biol. Chem. 284, 24673–24677.

[34] Mende, D.R., Sunagawa, S., Zeller, G., Bork, P. (2013) Accurate and universal delineation of prokaryotic species. Nat. Methods 10, 881–884.

[35] Nedashkovskaya, O.I., Kim, S.B., Han, S.K., Rhee, M.S., Lysenko, A.M., Falsen, E., Frolova, G.M., Mikhailov, V.V., Bae, K.S. (2004) *Ulvibacter litoralis* gen. nov., sp. nov., a novel member of the family *Flavobacteriaceae* isolated from the green alga *Ulva fenestrata*. Int. J. Syst. Evol. Microbiol. 54, 119–123.

[36] Needham, D.M., Fuhrman, J.A. (2016) Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. Nat. Microbiol. 1, 16005.

[37] Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P.A. (2017) metaSPAdes: a new versatile metagenomic assembler. Genome Res. 27, 824–834.

[38] 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.

[39] 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. 25, 1043–1055.

[40] Pernthaler, J. (2005) Predation on prokaryotes in the water column and its ecological implications. Nat. Rev. Microbiol. 3, 537.

[41] Pruesse, E., Peplies, J., Glöckner, F.O. (2012) SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 28, 1823–1829.

[42] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O. (2013) The SILVA ribosomal RNA gene database project:

improved data processing and web-based tools. Nucleic Acids Res. 41, D590–D596.

[43] Quince, C., Delmont, T.O., Raguideau, S., Alneberg, J., Darling, A.E., Collins, G., Eren, A.M. (2017) DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome Biol. 18, 181.

[44] Rawlings, N.D., Barrett, A.J., Bateman, A. (2011) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 40, D343–D350.

[45] Rodriguez-R, L.M., Konstantinidis, K.T. (2016) The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ, Preprints e1900v1.

[46] Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069.

[47] Soo, R.M., Skennerton, C.T., Sekiguchi, Y., Imelfort, M., Paech, S.J., Dennis, P.G., Steen, J.A., Parks, D.H., Tyson, G.W., Hugenholtz, P. (2014) An expanded genomic representation of the phylum *Cyanobacteria*. Genome Biol. Evol. 6, 1031–1045.

[48] Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.

[49] Teeling, H., Fuchs, B.M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C.M., Kassabgy, M., Huang, S., Mann, A.J., Waldmann, J., Weber, M., Klindworth, A., Otto, A., Lange, J., Bernhardt, J., Reinsch, C., Hecker, M., Peplies, J., Bockelmann, F.D., Callies, U., Gerdts, G., Wichels, A., Wiltshire, K.H., Glöckner, F.O., Schweder, T., Amann, R. (2012) Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. Science 336, 608–611.

[50] Teeling, H., Fuchs, B.M., Bennke, C.M., Krüger, K., Chafee, M., Kappelmann, L., Reintjes, G., Waldmann, J., Quast, C., Glöckner, F.O., Lucas, J., Wichels, A., Gerdts, G., Wiltshire, K.H., Amann, R.I. (2016) Recurring patterns in bacterioplankton dynamics during coastal spring algae blooms. eLife 5, e11888.

[51] Terrapon, N., Lombard, V., Drula, E., Coutinho, P.M., Henrissat, B. 2017 The CAZy database/the carbohydrate-active enzyme (CAZy) database: principles and usage guidelines. In: A Practical Guide to Using Glycomics Databases, Springer, Tokyo, pp. , 117–131.

[52] Terrapon, N., Lombard, V., Drula, É., Lapébie, P., Al-Masaudi, S., Gilbert, H.J., Henrissat, B. (2017) PULDB: the expanded database of Polysaccharide Utilization Loci. Nucleic Acids Res. 46, D677–D683.

[53] Terrapon, N., Lombard, V., Gilbert, H.J., Henrissat, B. (2014) Automatic prediction of polysaccharide utilization loci in *Bacteroidetes* species. Bioinformatics 31, 647–655.

[54] Thiele, S., Fuchs, B.M., Amann, R. 2011 Treatise on Water Science, Elsevier, Oxford, pp. , 171–189.

[55] Thiele, S., Fuchs, B.M., Ramaiah, N., Amann, R. (2012) Microbial community response during the iron fertilization experiment LOHAFEX. Appl. Environ. Microbiol. 78, 8803–8812.

[56] Thomas, F., Hehemann, J.H., Rebuffet, E., Czjzek, M., Michel, G. (2011) Environmental and gut *Bacteroidetes*: the food connection. Front. Microbiol. 2, 93.

[57] Xing, P., Hahnke, R.L., Unfried, F., Markert, S., Huang, S., Barbeyron, T., Harder, J., Becher, D., Schweder, T., Glöckner, F.O., Amann, R.I., Teeling, H. (2014) Niches of two polysaccharide-degrading *Polaribacter* isolates from the North Sea during a spring diatom bloom. ISME J. 9, 1410–1422.

[58] Yamaguchi, T., Kawakami, S., Hatamoto, M., Imachi, H., Takahashi, M., Araki, N., Yamaguchi, T., Kubota, K. (2015) In situ DNA-hybridization chain reaction (HCR): a facilitated in situ HCR system for the detection of environmental microorganisms. Environ. Microbiol. 17, 2532–2541.

[59] Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K.H., Whitman, W.B., Euzéby, J., Amann, R., Rosselló-Móra, R. (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat. Rev. Microbiol. 12, 635–645.

[60] Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J.R., Amaral-Zettler, L., Gilbert, J.A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., Birren, B.W., Blaser, M.J., Bonazzi, V., Booth, T., Bork, P., Bushman, F.D., Buttigieg, P.L., Chain, P.S.G., Charlson, E., Costello, E.K., Huot-Creasy, H., Dawyndt, P., DeSantis, T., Fierer, N., Fuhrman, J.A., Gallery, R.E., Gevers, D., Gibbs, R.A., Gil, I.S., Gonzalez, A., Gordon, J.I., Guralnick, R., Hankeln, W., Highlander, S., Hugenholtz, P., Jansson, J., Kau, A.L., Kelley, S.T., Kennedy, J., Knights, D., Koren, O., Kuczynski, J., Kyrpides, N., Larsen, R., Lauber, C.L., Legg, T., Ley, R.E., Lozupone, C.A., Ludwig, W., Lyons, D., Maguire, E., Methé, B.A., Meyer, F., Muegge, B., Nakielny, S., Nelson, K.E., Nemergut, D., Neufeld, J.D., Newbold, L.K., Oliver, A.E., Pace, N.R., Palanisamy, G., Peplies, J., Petrosino, J., Proctor, L., Pruesse, E., Quast, C., Raes, J., Ratnasingham, S., Ravel, J., Relman, D.A., Assunta-Sansone, S., Schloss, P.D., Schriml, L., Sinha, R., Smith, M.I., Sodergren, E., Spor, A., Stombaugh, J., Tiedje, J.M., Ward, D.V., Weinstock, G.M., Wendel, D., White, O., Whiteley, A., Wilke, A., Wortman, J.R., Yatsunenko, T., Glöckner, F.O. (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat. Biotechnol. 29, 415–420.

[61] Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., Xu, Y. (2012) dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 40, W445–W451.

[62] Yu, C.S., Chen, Y.C., Lu, C.H., Hwang, J.K. (2006) Prediction of protein subcellular localization. Proteins Struct. Funct. Bioinf. 64, 643–651.

[63] Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S.C., Ester, M., Foster, L.J., Brinkman, F.S. (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 26, 1608–1615.