Evolution in action: Habitat-transition leads to genome-streamlining in Methylophilaceae (Betaproteobacteriales)

Running title: Habitat-transition leads to genome-streamlining

Michaela M. Salcher[1,2]*, Daniel Schaefle[2,3], Melissa Kaspar[2], Stefan M. Neuenschwander[2], Rohit Ghai[1]

[1]Department of Aquatic Microbial Ecology, Institute of Hydrobiology, Biology Centre CAS, Na Sádkách 7, 37005 České Budějovice, Czech Republic

[2]Limnological Station, Institute of Plant and Microbial Biology, University of Zurich, Seestrasse 187, 8802 Kilchberg, Switzerland

[3]Institute of Medical Microbiology, University of Zurich, Gloriastrasse 28/30, 8006 Zurich, Switzerland

*corresponding author:
Michaela M. Salcher
Department of Aquatic Microbial Ecology, Institute of Hydrobiology, Biology Centre CAS
Na Sádkách 7, 37005 České Budějovice, Czech Republic
Phone: +420 387 775 836
Email: michaelasalcher@gmail.com

The authors declare that they have no competing interests.

Abstract

The most abundant aquatic microbes are small in cell and genome size. Genome-streamlining theory predicts gene loss caused by evolutionary selection driven by environmental factors, favouring superior competitors for limiting resources. However, evolutionary histories of such abundant, genome-streamlined microbes remain largely unknown. Here we reconstruct the series of steps in the evolution of some of the most abundant genome-streamlined microbes in freshwaters ('*Ca*. Methylopumilus') and oceans (marine lineage OM43). A broad genomic spectrum is visible in the family Methylophilaceae (Betaproteobacteriales), from sediment microbes with medium-sized genomes (2-3 Mbp genome size), an occasionally blooming pelagic intermediate (1.7 Mbp), and the most reduced pelagic forms (1.3 Mbp). We show that a habitat transition from freshwater sediment to the relatively oligotrophic pelagial was accompanied by progressive gene loss and adaptive gains. Gene loss has mainly affected functions not necessarily required or advantageous in the pelagial or are encoded by redundant pathways. Likewise, we identified genes providing adaptations to oligotrophic conditions that have been transmitted horizontally from pelagic freshwater microbes. Remarkably, the secondary transition from the pelagial of lakes to the oceans required only slight modifications, i.e., adaptations to higher salinity, gained via horizontal gene transfer from indigenous microbes. Our study provides first genomic evidence of genome-reduction taking place during habitat transitions. In this regard, the family Methylophilaceae is an exceptional model for tracing the evolutionary history of genome-streamlining as such a collection of evolutionarily related microbes from different habitats is practically unknown for other similarly abundant microbes (e.g., '*Ca*. Pelagibacterales', '*Ca*. Nanopelagicales').

Keywords

Genome-streamlining, Genome reduction, Horizontal gene transfer, Evolutionary selection, Habitat transition, Comparative Genomics, Bacteria

Introduction

Marine and freshwater pelagic habitats are numerically dominated by very small microbes (cell volumes <0.1 μm^3) that seem to be perfectly adapted to nutrient-poor (oligotrophic) conditions by successfully competing for dissolved organic matter and nutrients at low nM concentrations due to higher surface-to-volume ratios and superior transport systems [1]. Small-sized cells also enjoy other benefits such as reduced replication costs and mortality rates by size selective protistan predators [2]. The genomes of such oligotrophs are characterized by being very small (streamlined, <1.5 Mbp) with highly conserved core genomes and few pseudogenes, compacted intergenic spacers, reduced numbers of paralogs, and a low genomic GC content [3, 4]. While genetic drift has been proposed as the evolutionary mechanism behind the reduced genomes of symbionts, parasites and commensals, selection driven by environmental factors has been suggested as the primary driving force in the case of free-living oligotrophs [3]. The most abundant organisms on earth, bacteria of the marine SAR11 lineage ('*Ca*. Pelagibacter ubique', Alphaproteobacteria) serve as models for genome streamlining in the oceans [5] and their freshwater sister lineage LD12 is also known to be of similarly small size [6, 7]. Other examples of aquatic microbes with small cells and reduced genomes can be found among Actinobacteria (marine '*Ca*. Actinomarina minuta' [8], freshwater '*Ca*. Nanopelagicales' [9, 10], freshwater luna1 lineage [11, 12]), Thaumarchaeota (marine '*Ca.* Nitrosopelagicus brevis')[13], and Betaproteobacteriales (freshwater '*Ca*. Methylopumilus planktonicus' [14], marine OM43 lineage [15, 16]).

The latter are methylotrophs that are specialized in using reduced one-carbon (C_1) compounds like methanol, methylamine and formaldehyde as sole energy and carbon sources by means of a modular system of different pathways for their oxidation, demethylation and assimilation [17]. The family Methylophilaceae (Betaproteobacteriales) is among the most important methylotrophs playing a key role in the carbon cycle of aquatic habitats [17, 18]. Four genera are so far validly described (*Methylotenera*, *Methylobacillus,*

Methylophilus, Methylovorus) that mainly inhabit the sediment of freshwater lakes [19-22]. Axenic strains have been also isolated from the pelagial of lakes ('*Ca*. Methylopumilus')[14] and oceans (lineage OM43) [15, 16, 23]. Freshwater '*Ca*. Methylopumilus planktonicus' are ubiquitous and highly abundant in lakes [24] with distinct maxima during diatom and/or cyanobacterial blooms [14, 25, 26], indicating that C_1 substrates supporting their growth are released from primary producers. Members of the coastal marine OM43 lineage display similar temporal patterns with highest numbers during phytoplankton blooms [27-29].

In this work, we analysed the evolutionary history of the family Methylophilaceae by comparative genomics. While sediment dwellers have a larger cell and genome size, pelagic lineages are genome-streamlined. We hypothesize that the evolutionary origin of the family can be traced back to freshwater sediments, from where these microbes emerged to colonize the plankton of lakes and eventually also crossing the freshwater-marine boundary. The transition from sediments to the pelagial resulted in a pronounced genome reduction and adaptive gene loss has mainly affected functions that are not necessarily required or advantageous in the pelagial or are encoded in redundant pathways. Likewise, genes providing adaptations to oligotrophic conditions might have been transmitted horizontally from indigenous pelagic microbes.

Material and Methods

Isolation of planktonic freshwater methylotrophs.

Novel strains of '*Ca*. Methylopumilus' and other Methylophilaceae were isolated from the pelagial of Lake Zurich (CH), Římov Reservoir (CZ), and Lake Medard (CZ). Dilution-to-extinction using filtered (0.2 µm) and autoclaved water amended with vitamins and amino acids as a medium was used for Lake Zurich[14]. A full-cycle isolation approach [30] was employed for samples from Římov Reservoir and Lake Medard, with filtered water samples (0.45 µm filters), being diluted 1:10 with Artificial Lake Water (ALW [31]) containing vitamins

(0.593 µM thiamine, 0.08 µM niacin, 0.000074 µM cobalamine, 0.005 µM para-amino benzoic acid, 0.074 µM pyridoxine, 0.081 µM pantothenic acid, 0.004 µM biotin, 0.004 µM folic acid, 0.555 µM myo-inosito, 0.01 µM riboflavin), 30 µM $LaCl_3$, 1 mM methanol and 0.1 mM methylamine and incubated for 1-2 days at *in situ* temperatures. This step resulted in a preadaptation of methylotrophs only (C_1 compounds as sole carbon source) without causing a shift in the assemblage of '*Ca*. Methylopumilus' as these microbes displays slow growth with doubling times of approx. two days. Thereafter, a dilution to extinction technique was employed [14] with approx. 1 cell per cultivation well in 24-well-plates. Plates were incubated for 4-6 weeks at *in situ* temperature and growth in individual wells was checked microscopically and by PCR and Sanger sequencing of 16S rRNA genes.

Whole-genome sequencing, assembly, and functional annotation.

Thirty-eight pure cultures of '*Ca*. Methylopumilus sp.' and three *Methylophilus* sp. were grown in 400 ml ALW medium supplemented with vitamins, $LaCl_3$, methanol and methylamine for 6-8 weeks, pelleted by centrifugation, and DNA was isolated with a MagAttract® HMW DNA Kit (Qiagen). 550-bp libraries were constructed with the KAPA Hyper Prep Kit (Roche) and paired-end sequences (2 x 250-bp) were generated on an Illumina MiSeq instrument with a 500-cycle MiSeq Reagent v2 kit (Illumina). Library preparation and sequencing was done at the Genetic Diversity Center Zurich (GDC). Raw reads were quality trimmed with trimmomatic [32], assembled with SPAdes [33] and subsequently mapped to the resulting assemblies with Geneious 9 (www.geneious.com) in order to identify potential assembly errors. Assembly usually resulted in 1-2 large contigs with overlapping ends that mostly could be circularized *in silico*. In the case of non-overlapping contigs, genomes were closed by designing specific primers for PCR and Sanger sequencing. Moreover, regions containing low coverage (≤10 fold), ambiguities, or anomalies in the mapping were verified by designing specific primers for PCR and Sanger sequencing to produce high-quality reference genomes. Gene prediction was done with PROKKA [34] and annotation was done with an in-house pipeline [10] based on BLAST

searches to NCBI NR, COG [35], TIGRFAM [36] and KEGG databases [37]. Metabolic pathways were inferred from KEGG [37] and MetaCyc [38] and manually examined for completeness. Pathways involved in methylotrophy were identified by collecting 1016 reference protein sequences from published genomes of methylotrophs [14, 17, 39-45] and for the sake of completeness, also pathways not common to Methylophilaceae were included (e.g., methane oxidation [46-48], methylovory [49, 50]). These proteins were classified into 25 modules representing distinct (or sometimes alternative) biochemical transformations relevant to a methylotrophic lifestyle (e.g. M01-methanol oxidation, M02-pyrroloquinoline quinone biosynthesis, etc.; a complete list is provided in Supplementary Table S5). Protein sequences were clustered at 90% identity and 90% coverage with cd-hit [51] and the clusters were aligned using muscle [52]. The alignments were converted to HMMs (Hidden Markov Models) using the hmmbuild program in the HMMER3 package [53]. The program hmmsearch was used to scan complete genomes using these HMMs using e-value cut-off of 1e-3. The entire set of HMMs is available as Supplementary Data Set.

Fragment recruitment from metagenomes.

Publicly available metagenomes gained from freshwater sediments (n=131), the pelagial of lakes (n=345), rivers (n=43), estuaries, brackish and coastal oceanic sites (n=53) as well as open oceans (n=201) were used for fragment recruitment (see Table S2 for sampling sites and SRR accessions). rRNA sequences in genomes were identified with barrnap (http://www.vicbioinformatics.com/software.barrnap.shtml) and masked to avoid biases, and metagenomic reads were queried against the genomes using BLASTN [54] (cut-offs: length ≥50 bp, identity ≥95%, e-value ≤1e-5). These hits were used to compute RPKG values (number of reads recruited per kb of genome per Gb of metagenome), which provide a normalized value that is comparable across different genomes and metagenomes.

Comparative genomic analyses.

All publicly available genomes of high quality (>95% completeness, <20 scaffolds) affiliated with the family Methylophilaceae (Table S1, n=37) were downloaded from NCBI and re-

157 annotated for comparative analyses. Average nucleotide identities (ANI [55]) and average
158 amino acid identities (AAI [56]) were calculated to discriminate different species and genera.
159 Phylogenomic trees based on conserved concatenated protein sequences (351,312 amino
160 acid sites from 878 proteins for all Methylophilaceae, Fig. 1; 337,501 amino acid sites from
161 983 proteins for all '*Ca*. Methylopumilus' spp., Fig. S1) was generated with FastTree [57] (100
162 bootstraps) after alignment with kalign [58]. *Methyloversatilis* sp. RAC08 (NZ_CP016448) and
163 '*Ca*. Methylosemipumilus turicensis' MMS-10A-171 (NZ_LN794158) served as outgroup for
164 the trees displayed in Fig. 1 and Fig. S1, respectively. The core- and pangenome of the family
165 was computed using all-vs-all comparisons of all proteins for each genome using BLASTP
166 (≥50% identity and ≥50% coverage cut-offs to define an ortholog). Paralogs in each genome
167 were identified with BLASTP (cut-offs: ≥80% coverage, ≥70% similarity, ≥50% identity).
168 Closest relatives for proteins putatively transferred horizontally were identified with BLASTP
169 against the NCBI Protein Reference Sequences database (cut-off: *E* values ≤1e-5). Trees for
170 individual proteins or concatenated proteins for specific pathways were constructed with
171 RAxML (GAMMA BLOSUM62 model [59]) after alignment with MAFFT v7.388 [60].

172 *Availability of data*

173 All genomes have been submitted to NCBI under BioProject XXX, BioSamples XY-XYX
174 (Please note: Submission is in progress, accession numbers will be provided as soon as
175 possible). The entire set of HMMs related to methylotrophic functions is available as
176 Supplementary Data Set.

177

178 **Results and discussion**

179 *Phylogenomics and global occurrence of Methylophilaceae*

180 Currently, 31 Methylophilaceae genomes of high quality (i.e., >99% completeness, <20
181 scaffolds) are publicly available, mostly from axenic isolates from freshwater sediments (Fig.
182 1, Table S1). We additionally sequenced the genomes of 41 strains of planktonic freshwater

183 strains affiliated with '*Ca*. Methylopumilus planktonicus' (38 strains) and *Methylophilus* sp. (3
184 strains). These microbes were isolated from the pelagial of three different freshwater habitats
185 (Lake Zurich, CH; Římov Reservoir, CZ; Lake Medard, CZ) by dilution-to-extinction [14, 30].
186 All novel genomes are of very high quality, i.e., they are complete, with one circular
187 chromosome (Table S1). The 39 strains classified as '*Ca*. M. planktonicus' by 16S rRNA gene
188 sequences analysis (99.94-100% sequence identity), constitute at least three different species
189 according to average nucleotide and amino acid identity (ANI and AAI [61]) (Fig. S1-S4). We
190 tentatively name these three taxa '*Ca*. Methylopumilus rimovensis' (two strains isolated from
191 Římov Reservoir), '*Ca*. Methylopumilus universalis' (29 strains from Lake Zurich and Římov
192 Reservoir) and the originally described '*Ca*. Methylopumilus planktonicus' (eight strains from
193 Lake Zurich; Fig. 1, Fig. S1)[14]. AAI values also suggest that '*Ca*. M. turicensis' is a different
194 genus (62% AAI with '*Ca*. Methylopumilus', Fig. S3) that we tentatively rename to '*Ca*.
195 Methylosemipumilus turicensis'. This reclassification is in line with the recently released
196 Genome Taxonomy Database (GTDB [62]). Moreover, the genus *Methylotenera* might be split
197 in different genera and the GTDB suggests a reclassification of several strains to the genus
198 *Methylophilus*. Our analysis notes a polyphyletic pattern of *Methylotenera* with three different
199 genera (*Methylotenera*-1, *Methylotenera*-2, and *Methylotenera*-3; Fig. 1a, Fig. S3, >70% AAI).
200 However, further work is necessary to clarify the formal naming of these strains as AAI values
201 are inconclusive and the proposed hard cut-off of 65% AAI for genus delineation are not met
202 for most members of Methylophilaceae. Three novel pelagic *Methylophilus* sp. isolates (MMS-
203 M-34, MMS-M-37, MMS-M-51) constitute a novel species that we tentatively named *M.*
204 *medardicus*, with closest hits to isolates from freshwater sediment. These strains might
205 originate from the same clone, as they were gained from the same sample from Lake Medard
206 and were 100% identical in in their genome sequence. *M. mediardicus* seem to be not
207 abundant in the pelagial of lakes, as indicated by recruitments from 345 different pelagic
208 freshwater metagenomic datasets, however they could be readily detected in relatively high
209 proportions in sediment metagenomes (Fig. 1b, Table S2). Sediments also appear to be the
210 main habitat of other *Methylophilus* and *Methylotenera*. The three strains isolated from marine

systems, that were referred to as OM43-lineage [15, 16, 23], form two different genera based on AAI (Fig. S3). However, none appear to be abundant in the open ocean (Fig. 1b), and only strain HTCC2181 could be detected in estuarine/coastal metagenomes, although lineage OM43 has been repeatedly reported in coastal oceans by CARD-FISH, where they can reach up to 4% or 0.8 x 10^5 cells ml^{-1} during phytoplankton blooms [28, 29]. It is thus likely that other, more abundant strains of OM43 still await isolation. '*Ca*. Methylopumilus spp.' on the other hand, were found in moderate proportions in estuarine/coastal systems, but their main habitat is clearly the pelagial of lakes, where they are highly abundant (Fig. 1b), as previously reported based on CARD-FISH analyses [14, 63], 16S rRNA gene amplicon sequencing [24, 64-66], and metagenomics [67, 68]. All '*Ca*. Methylopumilus', especially '*Ca*. M. rimovensis' were also prevalent in rivers (Fig. 1b).

Genome-streamlining in pelagic strains

The genomes of pelagic freshwater '*Ca*. Methylopumilus sp.' (n=39) and the marine OM43 lineage (n=3) are characterized by very small sizes (1.26-1.36 Mbp) and a low genomic GC content (35.3-37.7%) (Table S1, Fig. 2, Fig. S5). '*Ca*. Methylosemipumilus turicensis' MMS-10A-171 has a slightly larger genome (1.75 Mbp) with higher GC content (44.5%), while all other Methylophilaceae have genome sizes >2.37 Mbp (max. 3.25 Mbp) and a higher GC content (41.9-55.7%, average 47.3%). A highly significant relationship between genome size and GC content, length of intergenic spacers, coding density, mean CDS length, number of overlapping CDS, paralogs, and numbers of genes involved in sensing of the environment (i.e., histidine kinases and sigma factors) was evident (Fig. 2). All these features have been proposed to be relevant for genome-streamlining [3] with freshwater '*Ca*. Methylopumilus' and marine OM43 displaying the most reduced forms and '*Ca*. Methylosemipumilus turicensis' presenting an intermediate state (Table S1). Moreover, we observed a negative relationship between genomic GC content and stop-codon usage of TAA instead of TAG, as well as a preferred amino acid usage of lysine instead of arginine (Fig. 2), both suggested to be involved in nitrogen limitation [4]. Furthermore, amino acids with less nitrogen and sulphur and

238 more carbon atoms were favourably encoded by the genome-streamlined microbes (Fig. 2,
239 S5, S6).

240 *Adaptive gene loss during habitat transition from the sediment to the pelagial*

241 The core genome of the family Methylophilaceae consists of 664 protein families (4.3%
242 of the pangenome) and an open pangenome of >15,000 protein families, while the
243 streamlined genomes of '*Ca*. Methylopumilus' have a highly conserved core (48%, Fig. S7).
244 By contrast, sediment Methylophilaceae have a larger pangenome with a more modular
245 assortment featuring several redundant pathways for methylotrophic functions [18] and a
246 large fraction of proteins overlapping with '*Ca*. M. turicensis', indicating a high evolutionary
247 relatedness (Fig. S7). It appears that both extant pelagic and sediment Methylophilaceae
248 shared a common sediment-dwelling methylotrophic ancestor. While one lineage
249 (*Methylotenera* and *Methylophilus*) retains the ancestral character (large genomes) of the
250 common ancestor, the other lineage diversified towards a pelagic lifestyle ('*Ca*. M. turicensis',
251 '*Ca*. Methylopumilus' and OM43). Remarkably, '*Ca*. M. turicensis' appears to constitute an
252 early diverging lineage that displays somewhat mixed characteristics of both sediment and
253 truly pelagic forms ('*Ca*. Methylopumilus' and OM43), not only in its phylogenetic position, but
254 also in genomic characteristics. Monitoring data from Lake Zurich showed consistently high
255 cell densities of '*Ca*. Methylopumilus', while '*Ca*. M. turicensis' were mostly below detection
256 limits except for a 3-month phase in one year where they reached high numbers in the
257 hypolimnion [14]. Moreover, also fragment recruitment from freshwater metagenomes
258 showed a global occurrence of '*Ca*. Methylopumilus' in very high relative proportions, while
259 '*Ca*. M. turicensis' was less prevalent (Fig. 1b). This hints again at the somewhat transitional
260 character of '*Ca*. M. turicensis' (occasional 'bloomer') that is not as perfectly adapted to the
261 pelagial as '*Ca*. Methylopumilus'.

262 We tested the hypothesis of adaptive gene loss driven by evolutionary selection during
263 the transition from sediment to the pelagial by comparative genomics of metabolic modules of
264 *Methylophilus*, *Methylotenera*, '*Ca*. M. turicensis', '*Ca*. Methylopumilus' and marine OM43

strains. *Methylobacillus* and *Methylovorus* were excluded as they seem too distantly related and also not very abundant in lake sediments (Fig. 1b, Table S2). The most pronounced differences in the genetic make-up of sediment vs. pelagic strains were detected in motility and chemotaxis (Figs. 3, 4), with all *Methylophilus* and all but two *Methylotenera* strains having flagella and type IV pili, while the planktonic strains have lost mobility and also greatly reduced the number of two-component regulatory systems and sigma factors. A large number of membrane transporters for inorganic compounds was detected exclusively in sediment Methylophilaceae, while this number is reduced in '*Ca*. M. turicensis' and even more in '*Ca*. Methylopumilus' and OM43 (Fig. 4, Table S3). Moreover, *Methylophilus* and *Methylotenera* encode multiple pathways for nitrogen acquisition, with transporters for ammonia, nitrate, nitrite, taurine, cyanate or urea, and pathways for urea or cyanate utilization (Fig. 3, Table S4)[69]. '*Ca*. M. turicensis', on the other hand, has only transporters for nitrate/taurine and ammonia (Amt family), and '*Ca*. Methylopumilus' and marine OM43 only carry ammonia transporters. All sediment Methylophilaceae further possess genes for assimilatory nitrate reduction to ammonia, some for dissimilatory nitrate reduction to nitrous oxide or detoxification of nitric oxide (quinol type of *norB*) [70] and *Methylotenera mobilis* 13 is a complete denitrifier [69, 71, 72], while none of the pelagic strains have any genes involved in nitrate reduction (Figs. 3, 4, Table S4). Ammonia is the main microbial nitrogen source in the epilimnion of lakes and oceans, while nitrate and other compounds like urea, taurine, or cyanate are more abundant in deeper, oxygenated layers and the sediment [14, 73, 74]. Therefore, an adaptation to ammonium uptake might be advantageous for pelagic microbes.

Furthermore, a high diversity of pathways involved in sulfur metabolism was detected in Methylophilaceae, with the genome-streamlined strains representing the most reduced forms again. All *Methylophilus*, *Methylotenera,* '*Ca*. Methylosemipumilus turicensis' and one strain of '*Ca*. Methylopumilus rimovensis' encode ABC transporters for sulfate uptake, and a sulfate permease was annotated in OM43 and several sediment Methylophilaceae, while the majority of '*Ca*. Methylopumilus' lack these transporters (Fig. 3, 4). Canonical assimilatory sulfate reduction seems to be incomplete in most Methylophilaceae, as adenylyl sulfate kinase *cysC*

was annotated only in a few strains (Table S4). Thus, the mode of sulfite generation remains unclear, with unknown APS kinases or other links from APS to sulfite. *Methylophilus rhizosphaera* encodes genes for dissimilatory sulfate reduction and most sediment strains possess ABC transporters for alkanesulfonates, most likely transporting methylsulfonate, that can be oxidised to sulfite by methanesulfonate monooxygenases generating formaldehyde as by-product. Dimethyl sulfide (DMS) seems to be a source for sulfur and formaldehyde as well, as dimethylsulfoxide and dimethyl monooxygenases are present in several sediment Methylophilaceae, but absent in all pelagic strains. It is thus still unclear how '*Ca*. Methylopumilus' fuel their sulfur demand, especially as they grow in a defined medium containing sulfate (200 µM $MgSO_4$; 160 µM $CaSO_4$) and vitamins as sole sulfur sources.

All Methylophilaceae have complete pathways for the biosynthesis of amino acids and vitamins, with the exception of cobalamin (vitamin B12) that was lacking in the pelagic lineage ('*Ca*. M. turicensis', '*Ca*. Methylopumilus', marine OM43), while either the complete biosynthesis or the salvage pathway was present in the sediment isolates (Figs. 3, 4, Table S4). However, putative cobalamin transporters were annotated in all isolates.

The methylcitric acid (MCA) cycle for oxidising propionate via methylcitrate to pyruvate is present in *Methylotenera*, *Methylophilus*, '*Ca*. M. turicensis', and the marine OM43, but absent in all '*Ca*. Methylopumilus' strains, suggesting it has been selectively lost in these organisms. All genes were arranged in a highly conserved fashion, with the exception of '*Ca*. M. turicensis' having a bifunctional aconitate hydratase 2/2-methylisocitrate dehydratase (*prpD/acnB*) gene and *acnB* genes being located in different genomic regions in OM43 and '*Ca*. M. turicensis' (possessing two copies), however, with high synteny of flanking genes (Fig. S8). Phylogenetic analysis of the MCA gene cluster resulted in genus-specific branching, and notably, the MCA pathway of OM43 is most closely related to that of '*Ca*. M. turicensis' (Fig. S8), suggesting it was retained in the OM43 lineage after divergence from a common ancestor of OM43 and '*Ca*. M. turicensis'.

Genome-streamlining leading to a loss of redundant methylotrophic pathways

Some of the sediment dwellers seem to be facultative methylotrophs, as ABC transporters for amino acids were annotated (Figs. 3, 4, Table S3). *Methylotenera versatilis* 301 additionally encodes a fructose-specific phosphotransferase system (PTS) and a 1-phosphofructokinase, as well as transporters for putrescine uptake and the subsequent pathway for its degradation. '*Ca*. M. turicensis' might also be a facultative methylotroph, as it possesses a PTS system for cellobiose, while this (as well as amino acid transporters) is lacking in all '*Ca*. Methylopumilus' and OM43 strains, making them obligate methylotrophs. These observations suggest that the ancestor of both pelagic and sediment lineages was also a facultative methylotroph and that obligate methylotrophy emerged only in the truly pelagic strains.

Remarkably, also pathways involved in methylotrophy were reduced in the course of genome streamlining with the sediment dwelling *Methylophilus* and *Methylotenera* having the most complete modules for C_1 compound oxidation, demethylation and assimilation (Fig. 3, 4, Table S4). They also encode multiple types of methanol dehydrogenases (up to five different types in single strains), while the pelagic forms possess only XoxF4-1 (Fig. S9). Moreover, the latter encode neither traditional methylamine-dehydrogenases nor the N-methylglutamate (NMG) pathway for methylamine oxidation. Thus, the mode of methylamine uptake is still unclear, although it has been experimentally demonstrated that some pelagic strains can utilize this C_1 substrate [14, 75]. However, also nearly half of the sediment strains lack these well-described pathways in a patchy manner only partly reflected by phylogeny, therefore it is likely that methylamine utilization is not a common feature within Methylophilaceae, or that alternative routes of its oxidation still await discovery [69]. Formaldehyde oxidation can be achieved via three alternative routes, and only four *Methylophilus* strains encode all of them, i.e., all others lack a formaldehyde-dehydrogenase. All *Methylophilus* and *Methylotenera* as well as '*Ca*. M. turicensis' carry genes for the tetrahydromethanopterin (H_4MPT) pathway, but none of the '*Ca*. Methylopumilus' and OM43 strains. Therefore, the only route for formaldehyde oxidation in these genome-streamlined microbes is the tetrahydrofolate (H_4F) pathway which includes the spontaneous reaction of formaldehyde to H_4F and is thought to be

348 relatively slow [14, 17]. The ribulose monophosphate (RuMP) cycle for formaldehyde
349 assimilation/oxidation and formate oxidation via formate dehydrogenases was annotated in all
350 Methylophilaceae, while none of them possess other potential methylotrophic modules such
351 as the serine cycle, the ethylmalonyl-CoA-pathway for glyoxylate regeneration, a glyoxylate
352 shunt, nor the Calvin-Benson-Bassham cycle for CO_2 assimilation, as already previously noted
353 to be lacking in Methylophilaceae [17]. Thus, the core methylotrophic modules in
354 Methylophilaceae contain methanol oxidation via XoxF methanol dehydrogenases,
355 formaldehyde oxidation via the H_4F pathway, the RuMP cycle, and formate oxidation (Fig. 3,
356 Table S4)[17, 69, 76]. The majority of genes encoding these pathways were organized in
357 operon structures or found in close vicinity to each other with high synteny and
358 phylogenetically reflecting the overall phylogeny of the family (Fig. S10, S11).

359 *Photoheterotrophy as adaptation to oligotrophic pelagic conditions.*

360 Rhodopsins are light-driven proton pumps producing ATP that fuel e.g., membrane
361 transporters [77] and play important roles during carbon starvation [78] in oligotrophic aquatic
362 environments. Therefore, the acquisition of rhodopsins are proposed to be powerful
363 adaptations to the pelagial. '*Ca*. Methylosemipumilus turicensis' acquired a proteorhodopsin
364 and the complete pathway for retinal biosynthesis via horizontal gene transfer (HGT) from the
365 abundant freshwater microbe *Polynucleobacter cosmopolitanus* (81% amino acid similarity,
366 Fig. 5, Fig. S12a). Interestingly, '*Ca*. Methylopumilus spp.' carry, in addition to a
367 proteorhodopsin highly similar to '*Ca*. M. turicensis' (78.3-79.5% similarity), a second
368 rhodopsin gene inserted between the proteorhodopsin and the retinal biosynthesis cluster
369 (Fig. 5). This xantho-like rhodopsin was most likely gained from rare freshwater
370 Betaproteobacteria (*Janthinobacterium lividum, Massilia psychrophila;* 49.4-53.5% similarity,
371 Fig. S12b), however, the binding of a carotenoid antenna seems unlikely due to the
372 replacement of a glycine with tryptophan in position 156, suggesting it also functions as a
373 proteorhodopsin (Fig. S12c)[79, 80]. Both rhodopsins are tuned to green light, which is
374 common in freshwaters [81] and possess the canonical DxxxK retinal binding motif in helix-7

that is characteristic of proton pumping rhodopsins [82]. The marine OM43 lineage only carry the xantho-like rhodopsin (59.0-63.9% similarity to '*Ca*. Methylopumilus'). It is unclear if the proteorhodopsin was never present in the marine lineage or was lost subsequently, and if so, the reasons for a secondary loss remain enigmatic as two rhodopsins would provide an even better adaptation to oligotrophic waters than one.

The second transition from freshwater pelagial to the marine realm is characterized by adaptations to a salty environment

The second habitat transition across the freshwater-marine boundary does not appear to involve genome streamlining, as genomes of pelagic freshwater and marine methylotrophs are of similar small size and low GC content (Figs. 1, 2). We hypothesize that this transition had less impact on the lifestyle (purely planktonic, oligotrophic) but required specific adaptations to the marine realm that were mainly acquired by HGT, and as suggested by the long branches in the phylogenetic tree (Fig. 1), multiple, rapid changes in existing genes. Besides the MCA pathway and the proteorhodopsin, no major rearrangements or reductions in general metabolic pathways were detected in marine OM43 in comparison to freshwater '*Ca*. Methylopumilus'. However, several adaptations to higher salt concentrations could be identified (Figs. 3, 4). Salinity is one of the most important obstacle in freshwater-marine colonization, and successful transitions have occurred rarely during the evolution of Proteobacteria [83, 84]. Main adaptations to higher salinities involve genes for osmoregulation and inorganic ion metabolism that might have been acquired from the indigenous community by HGT. For example, genes regulating the Na^+-dependent respiratory chain (Na^+-translocating NADH:quinone oxidoreductase, NQR, Fig. S13) have been transmitted from the marine *Roseobacter* lineage to strain HTCC2181 [16, 84]. The NQR system provides energy by generating a sodium motif force, yet, the sodium pumping might also be an adaptation to enhanced salinities [85]. All other Methylophilaceae possess the energetically more efficient H^+-translocating type (NDH), which works better under low salinity conditions and is thus common in freshwater microbes [85].

Ectoine, a compatible solute along with glycine betaine, helps organisms survive extreme osmotic stress by acting as an osmolyte [86]. Ectoine is synthesized from L-aspartate 4-semialdehyde, the central intermediate in the synthesis of amino acids of the aspartate family. Two marine OM43 strains (KB13 and MBRS-H7) encode this pathway followed by sodium:proline symporter *putP* arranged in high synteny and protein similarity with marine/hypersaline sediment microbes, thus it is likely that both components were gained via HGT (Fig. S14). A second copy of the *putP* symporter was common to all Methylophilaceae (data not shown). Also a dipeptide/tripeptide permease (DtpD) unique for the marine OM43 lineage seems to be transferred horizontally, either from marine Bacteroidetes or sediment-dwelling *Sulfurifustis* (Gammaproteobacteria, Fig. S15). Other putative membrane compounds involved in sodium transport in marine OM43 include a sodium:alanine symporter (AlsT, Fig. S16a), a sodium:acetate symporter (ActP, Fig. S16b), a sodium:dicarboxylate symporter (GltT, Fig. S16c), a sodium:proton antiporter (NhaP, Fig. S16d), and another putative sodium:proton antiporter (NhaE-like, Fig. S16e). Although also several other Methylophilaceae carry some of these sodium transporters, they are only distantly related to OM43, thus they might be acquired horizontally. Conversely, ActP and GltT of OM43 are most closely related to three '*Ca*. M. universalis' strains and the two '*Ca*. M. rimovensis' strains, respectively (Fig. S16b, S16c). Both symporters are related to microbes from freshwater and marine habitats, hinting to some yet unknown lineages related to both OM43 and '*Ca*. Methylopumilus' most likely thriving in the freshwater-marine transition zone.

Conclusions

Our study provides first genomic evidence that the ancestors of genome-streamlined pelagic Methylophilaceae can be traced back to sediments with two habitat transitions occurring in the evolutionary history of the family. The first from sediments to the pelagial is characterized by pronounced genome reduction driven by selection pressure for relatively more oligotrophic environmental conditions. This adaptive gene loss has mainly affected functions that (i) are not necessarily required in the pelagial (e.g., motility, chemotaxis), (ii) are

not advantageous for survival in an oligotrophic habitat (e.g., low substrate affinity transporters), and (iii) are encoded in redundant pathways (e.g., formaldehyde oxidation). Likewise, (iv) genes providing adaptations to oligotrophic conditions have been transmitted horizontally from indigenous pelagic microbes (e.g., rhodopsins). The second habitat transition across the freshwater-marine boundary did not result in further genome-streamlining, but is characterized by adaptations to higher salinities acquired by HGT. '*Ca*. M. turicensis' was identified as transitional taxon, retaining multiple ancestral characters while also gaining adaptations to the pelagial. In this regard, the family Methylophilaceae is an exceptional model for tracing the evolutionary history of genome-streamlining as such a collection of evolutionarily related microbes from different habitats is practically unknown for other similarly abundant genome-streamlined microbes (e.g., '*Ca*. Pelagibacterales', '*Ca*. Nanopelagicales').

Acknowledgements

We thank the team of the Genetic Diversity Center Zurich (GDC) for providing sequencing facilities and help with library preparation. Thomas Posch and Eugen Loher are acknowledged for help in sampling of Lake Zurich, Petr Znachor, Pavel Rychtecký and Jiří Nedoma for help in sampling of Řimov Reservoir and Lake Medard. MMS was supported by the research grant 19-23469S (Grant Agency of the Czech Republic). RG was supported by the research grant 17-04828S (Grant Agency of the Czech Republic). Sampling for the isolation of novel strains from Lake Zurich was supported by the SNF D-A-CH project 310030E-160603/1 awarded to Thomas Posch.

Authors' contributions

MMS conceived the project, isolated and sequenced the strains, analysed the data and wrote the manuscript. DS, MK and SMN sequenced the strains and contributed to data analyses. RG developed programs for analysis and contributed to data analyses. All authors helped to interpret the results and contributed to writing the manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. Button DK. Biochemical basis for whole-cell uptake kinetics: specific affinity, oligotrophic capacity, and the meaning of the Michaelis constant. *Appl Environ Microbiol* 1991; **57:** 2033-2038.
2. Pernthaler J. Predation on prokaryotes in the water column and its ecological implications. *Nat Rev Microbiol* 2005; **3:** 537-546.
3. Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for microbial ecology. *ISME J* 2014; **8:** 1553–1565.
4. Luo H, Thompson LR, Stingl U, Hughes AL. Selection maintains low genomic GC content in marine SAR11 lineages. *Mol Biol Evol* 2015; **32:** 2738-2748.
5. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D *et al.* Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 2005; **309:** 1242-1245.
6. Salcher MM, Pernthaler J, Posch T. Seasonal bloom dynamics and ecophysiology of the freshwater sister clade of SAR11 bacteria 'that rule the waves' (LD12). *ISME J* 2011; **5:** 1242-1252.
7. Eiler A, Mondav R, Sinclair L, Fernandez-Vidal L, Scofield DG, Schwientek P *et al.* Tuning fresh: radiation through rewiring of central metabolism in streamlined bacteria. *ISME J* 2016; **10:** 1902-1914.
8. Ghai R, Mizuno CM, Picazo A, Camacho A, Rodriguez-Valera F. Metagenomics uncovers a new group of low GC and ultra-small marine Actinobacteria. *Sci Rep* 2013; **3**.
9. Newton RJ, Jones SE, Helmus MR, McMahon KD. Phylogenetic ecology of the freshwater *Actinobacteria* acI lineage. *Appl Environ Microbiol* 2007; **73:** 7169-7176.
10. Neuenschwander SM, Ghai R, Pernthaler J, Salcher MM. Microdiversification in genome-streamlined ubiquitous freshwater Actinobacteria. *ISME J* 2018; **12:** 185.
11. Hahn MW, Schmidt J, Taipale SJ, Doolittle WF, Koll U. Rhodoluna lacicola gen. nov., sp. nov., a planktonic freshwater bacterium with stream-lined genome. *Int J Syst Evol Microbiol* 2014; **64:** 3254-3263.
12. Kang I, Lee K, Yang S-J, Choi A, Kang D, Lee YK *et al.* Genome sequence of "*Candidatus* Aquiluna" sp. strain IMCC13023, a marine member of the Actinobacteria isolated from an Arctic fjord. *J Bacteriol* 2012; **194:** 3550-3551.
13. Santoro AE, Dupont CL, Richter RA, Craig MT, Carini P, McIlvin MR *et al.* Genomic and proteomic characterization of "Candidatus Nitrosopelagicus brevis": An ammonia-oxidizing archaeon from the open ocean. *Proc Natl Acad Sci USA* 2015; **112:** 1173-1178.
14. Salcher MM, Neuenschwander SM, Posch T, Pernthaler J. The ecology of pelagic freshwater methylotrophs assessed by a high-resolution monitoring and isolation campaign. *ISME J* 2015; **9:** 2442-2453.
15. Giovannoni SJ, Hayakawa DH, Tripp HJ, Stingl U, Givan SA, Cho JC *et al.* The small genome of an abundant coastal ocean methylotroph. *Environ Microbiol* 2008; **10:** 1771-1782.
16. Jimenez-Infante F, Ngugi DK, Vinu M, Alam I, Kamau AA, Blom J *et al.* Comprehensive genomic analyses of the OM43 clade, including a novel species from the Red Sea, indicate ecotype differentiation among marine methylotrophs. *Appl Environ Microbiol* 2016; **82:** 1215-1226.
17. Chistoserdova L. Modularity of methylotrophy, revisited. *Environ Microbiol* 2011; **13:** 2603-2622.

18. Chistoserdova L. Methylotrophs in natural habitats: current insights through metagenomics. *Appl Microbiol Biotechnol* 2015; **99:** 5763-5779.
19. Kalyuzhnaya MG, Bowerman S, Lara JC, Lidstrom ME, Chistoserdova L. *Methylotenera mobilis* gen. nov., sp. nov., an obligately methylamine-utilizing bacterium within the family *Methylophilaceae*. *Int J Syst Evol Microbiol* 2006; **56:** 2819-2823.
20. Govorukhina NI, Trotsenko YA. Methylovorus, a new genus of restricted facultatively methylotrophic bacteria. *Int J Syst Bacteriol* 1991; **41:** 158-162.
21. Yordy JR, Weaver TL. Methylobacillus: a new genus of obligately methylotrophic bacteria. *Int J Syst Bacteriol* 1977; **27:** 247-255.
22. Jenkins O, Byrom D, Jones D. *Methylophilus* - a new genus of methanol-utilizing bacteria. *Int J Syst Bacteriol* 1987; **37:** 446-448.
23. Huggett M, Hayakawa D, Rappe M. Genome sequence of strain HIMB624, a cultured representative from the OM43 clade of marine Betaproteobacteria. *Standards in Genomic Sciences* 2012; **6**.
24. Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. A guide to the natural history of freshwater lake bacteria. *Microbiol Mol Biol R* 2011; **75:** 14-49.
25. Woodhouse JN, Kinsela AS, Collins RN, Bowling LC, Honeyman GL, Holliday JK *et al.* Microbial communities reflect temporal changes in cyanobacterial composition in a shallow ephemeral freshwater lake. *ISME J* 2016; **10:** 1337–1351.
26. Li J, Zhang J, Liu L, Fan Y, Li L, Yang Y *et al.* Annual periodicity in planktonic bacterial and archaeal community composition of eutrophic Lake Taihu. *Sci Rep* 2015; **5:** 15488.
27. Ramachandran A, Walsh DA. Investigation of XoxF methanol dehydrogenases reveals new methylotrophic bacteria in pelagic marine and freshwater ecosystems. *FEMS Microbiol Ecol* 2015; **91:** fiv105.
28. Morris RM, Longnecker K, Giovannoni SJ. Pirellula and OM43 are among the dominant lineages identified in an Oregon coast diatom bloom. *Environ Microbiol* 2006; **8:** 1361-1370.
29. Sekar R, Fuchs BM, Amann R, Pernthaler J. Flow sorting of marine bacterioplankton after fluorescence in situ hybridization. *Appl Environ Microbiol* 2004; **70:** 6210-6219.
30. Salcher MM, Šimek K. Isolation and cultivation of planktonic freshwater microbes is essential for a comprehensive understanding of their ecology. *Aquat Microb Ecol* 2016; **77:** 183-196.
31. Zotina T, Köster O, Jüttner F. Photoheterotrophy and light-dependent uptake of organic and organic nitrogenous compounds by *Planktothrix rubescens* under low irradiance. *Freshwater Biol* 2003; **48:** 1859-1872.
32. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; **30:** 2114-2120.
33. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012; **19:** 455-477.
34. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014; **30:** 2068-2069.
35. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS *et al.* The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucl Acid Res* 2001; **29:** 22-28.
36. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen Ian T *et al.* TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucl Acid Res* 2001; **29:** 41-43.
37. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol* 2016; **428:** 726-731.
38. Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M *et al.* The MetaCyc database of metabolic pathways and enzymes. *Nucl Acid Res* 2017; **46:** D633-D639.
39. Chistoserdova L, Lapidus A, Han C, Goodwin L, Saunders L, Brettin T *et al.* Genome of *Methylobacillus flagellatus*, molecular basis for obligate methylotrophy, and polyphyletic origin of methylotrophy. *J Bacteriol* 2007; **189:** 4020-4027.

40. Lapidus A, Clum A, LaButti K, Kaluzhnaya MG, Lim S, Beck DAC *et al.* Genomes of three methylotrophs from a single niche reveal the genetic and metabolic divergence of the *Methylophilaceae*. *J Bacteriol* 2011; **193:** 3757-3764.
41. Vuilleumier S, Chistoserdova L, Lee M-C, Bringel F, Lajus A, Zhou Y *et al.* Methylobacterium Genome Sequences: A Reference Blueprint to Investigate Microbial Metabolism of C1 Compounds from Natural and Industrial Sources. *PLOS ONE* 2009; **4:** e5584.
42. Good N, Lamb A, Beck D, Martinez-Gomez N, Kalyuzhnaya M. C1-Pathways in Methyloversatilis universalis FAM5: Genome Wide Gene Expression and Mutagenesis Studies. *Microorganisms* 2015; **3:** 175.
43. Kalyuzhnaya MG, Hristova KR, Lidstrom ME, Chistoserdova L. Characterization of a Novel Methanol Dehydrogenase in Representatives of Burkholderiales: Implications for Environmental Detection of Methylotrophy and Evidence for Convergent Evolution. *J Bacteriol* 2008; **190:** 3817-3823.
44. Brautaset T, Jakobsen ØM, Flickinger MC, Valla S, Ellingsen TE. Plasmid-Dependent Methylotrophy in Thermotolerant Bacillus methanolicus. *J Bacteriol* 2004; **186:** 1229-1238.
45. Vorholt JA, Kalyuzhnaya MG, Hagemeier CH, Lidstrom ME, Chistoserdova L. MtdC, a Novel Class of Methylene Tetrahydromethanopterin Dehydrogenases. *J Bacteriol* 2005; **187:** 6069-6074.
46. Ward N, Larsen Ø, Sakwa J, Bruseth L, Khouri H, Durkin AS *et al.* Genomic Insights into Methanotrophy: The Complete Genome Sequence of Methylococcus capsulatus (Bath). *PLoS Biol* 2004; **2:** e303.
47. Hou S, Makarova KS, Saw JH, Senin P, Ly BV, Zhou Z *et al.* Complete genome sequence of the extremely acidophilic methanotroph isolate V4, Methylacidiphilum infernorum, a representative of the bacterial phylum Verrucomicrobia. *Biology Direct* 2008; **3:** 26.
48. Wu ML, Wessels HJCT, Pol A, Op den Camp HJM, Jetten MSM, van Niftrik L *et al.* XoxF-type methanol dehydrogenase from the anaerobic methanotroph "*Candidatus* Methylomirabilis oxyfera". *Appl Environ Microbiol* 2015; **81:** 1442-1451.
49. Sun J, Steindler L, Thrash JC, Halsey KH, Smith DP, Carter AE *et al.* One carbon metabolism in SAR11 pelagic marine bacteria. *PLoS ONE* 2011; **6:** e23973.
50. Denef VJ, Mueller RS, Chiang E, Liebig JR, Vanderploeg HA. Chloroflexi CL500-11 populations that predominate deep-lake hypolimnion bacterioplankton rely on nitrogen-rich dissolved organic matter metabolism and C1 compound oxidation. *Appl Environ Microbiol* 2016; **82:** 1423-1432.
51. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006; **22:** 1658-1659.
52. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl Acid Res* 2004; **32:** 1792-1797.
53. Eddy SR. Accelerated Profile HMM Searches. *PLOS Computational Biology* 2011; **7:** e1002195.
54. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acid Res* 1997; **25:** 3389-3402.
55. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 2007; **57:** 81-91.
56. Rodriguez-R LM, Konstantinidis KT. Bypassing cultivation to identify bacterial species. *ASM Microbe Magazine* 2014; **9:** 111-118.
57. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLOS ONE* 2010; **5:** e9490.
58. Lassmann T, Sonnhammer EL. Kalign – an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* 2005; **6:** 298.
59. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014; **30:** 1312-1313.

60. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 2013; **30:** 772-780.
61. Konstantinidis KT, Rossello-Mora R, Amann R. Uncultivated microbes in need of their own taxonomy. *ISME J* 2017; **11:** 2399-2406.
62. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology* 2018.
63. Salcher MM, Pernthaler J, Frater N, Posch T. Vertical and longitudinal distribution patterns of differnt bacterioplankton populations in a canyon-shaped, deep prealpine lake. *Limnol Oceanogr* 2011; **56:** 2027-2039.
64. Bock C, Salcher M, Jensen M, Pandey RV, Boenigk J. Synchrony of eukaryotic and prokaryotic planktonic communities in three seasonally sampled Austrian lakes. *Front Microbiol* 2018; **9**.
65. Linz AM, Crary BC, Shade A, Owens S, Gilbert JA, Knight R *et al.* Bacterial community composition and dynamics spanning five years in freshwater bog lakes. *mSphere* 2017; **2**.
66. Okazaki Y, Nakano S-i. Vertical partitioning of freshwater bacterioplankton community in a deep mesotrophic lake with a fully oxygenated hypolimnion (Lake Biwa, Japan). *Environ Microbiol Rep* 2016; **8:** 780-788.
67. Cabello-Yeves PJ, Zemskaya TI, Rosselli R, Coutinho FH, Zakharenko AS, Blinov VV *et al.* Genomes of novel microbial lineages assembled from the sub-ice waters of Lake Baikal. *Appl Environ Microbiol* 2018; **84:** e02132-02117.
68. Bendall ML, Stevens SLR, Chan L-K, Malfatti S, Schwientek P, Tremblay J *et al.* Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J* 2016; **10:** 1589-1601.
69. Beck DAC, McTaggart TL, Setboonsarng U, Vorobev A, Kalyuzhnaya MG, Ivanova N *et al.* The expanded diversity of *Methylophilaceae* from Lake Washington through cultivation and genomic sequencing of novel ecotypes. *PLoS ONE* 2014; **9:** e102458.
70. Hendriks J, Oubrie A, Castresana J, Urbani A, Gemeinhardt S, Saraste M. Nitric oxide reductases in bacteria. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 2000; **1459:** 266-273.
71. Mustakhimov I, Kalyuzhnaya MG, Lidstrom ME, Chistoserdova L. Insights into denitrification in *Methylotenera mobilis* from denitrification pathway and methanol metabolism mutants. *J Bacteriol* 2013; **195:** 2207-2211.
72. Kalyuzhnaya MG, Lapidus A, Ivanova N, Copeland AC, McHardy AC, Szeto E *et al.* High-resolution metagenomics targets specific functional types in complex microbial communities. *Nature Biotechnology* 2008; **26:** 1029-1034.
73. Kitzinger K, Padilla CC, Marchant HK, Hach PF, Herbold CW, Kidane AT *et al.* Cyanate and urea are substrates for nitrification by Thaumarchaeota in the marine environment. *Nat Microbiol* 2019; **4:** 234-243.
74. Wetzel R: *Limnology. Lake and River Ecosystems*. 3rd edn: Elsevier Academic Press; 2001.
75. Halsey KH, Carter AE, Giovannoni SJ. Synergistic metabolism of a broad range of C_1 compounds in the marine methylotrophic bacterium HTCC2181. *Environ Microbiol* 2011; **14:** 630-640.
76. Chistoserdova L. Methylotrophy in a lake: from metagenomics to single-organism physiology. *Appl Environ Microbiol* 2011; **77:** 4705-4711.
77. Pinhassi J, DeLong EF, Béjà O, González JM, Pedrós-Alió C. Marine bacterial and archaeal ion-pumping rhodopsins: Genetic diversity, physiology, and ecology. *Microbiol Mol Biol R* 2016; **80:** 929-954.
78. Steindler L, Schwalbach MS, Smith DP, Chan F, Giovannoni SJ. Energy starved *Candidatus* Pelagibacter ubique substitutes light-mediated ATP production for endogenous carbon respiration. *PLoS ONE* 2011; **6:** e19725.

657 79. Luecke H, Schobert B, Stagno J, Imasheva ES, Wang JM, Balashov SP *et al.* Crystallographic
658 structure of xanthorhodopsin, the light-driven proton pump with a dual chromophore.
659 *Proceedings of the National Academy of Sciences* 2008; **105:** 16561-16565.
660 80. Imasheva ES, Balashov SP, Choi AR, Jung K-H, Lanyi JK. Reconstitution of Gloeobacter
661 violaceus Rhodopsin with a Light-Harvesting Carotenoid Antenna. *Biochemistry* 2009; **48:**
662 10948-10955.
663 81. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* The *Sorcerer* II
664 Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS*
665 *Biol* 2007; **5:** e77.
666 82. Béjà O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP *et al.* Bacterial Rhodopsin:
667 Evidence for a New Type of Phototrophy in the Sea. *Science* 2000; **289:** 1902-1906.
668 83. Logares R, Bråte J, Bertilsson S, Clasen JL, Shalchian-Tabrizi K, Rengefors K. Infrequent
669 marine-freshwater transitions in the microbial world. *Trends Microbiol* 2009; **17:** 414-422.
670 84. Walsh DA, Lafontaine J, Grossart H-P: On the eco-evolutionary relationships of fresh and salt
671 water bacteria and the role of gene transfer in their adaptation. In *Lateral Gene Transfer in*
672 *Evolution.* Edited by Gophna U: Springer New York; 2013: 55-77
673 85. Zhang H, Yoshizawa S, Sun Y, Huang Y, Chu X, González JM *et al.* Repeated evolutionary
674 transitions of flavobacteria from marine to non-marine habitats. *Environ Microbiol* 2019; **0**.
675 86. Roberts MF. Organic compatible solutes of halotolerant and halophilic microorganisms.
676 *Saline Systems* 2005; **1:** 5.

677

678

679 ## Figure legends:

680 **Figure 1: Phylogeny of Methylophilaceae and their occurrence in different**
681 **environments.** (a) Phylogenomic tree based on 878 common concatenated proteins
682 (351,312 amino acid sites) with *Methyloversatilis* sp. RAC08 as outgroup. The 39 complete
683 genomes of '*Ca*. Methylopumilus sp.' are collapsed to the species level, see Fig. S1 for a
684 complete tree. Different genera (70% AAI cut-off, Fig. S2) are marked by grey boxes.
685 Isolation sources of strains are indicated by different colours and incomplete genomes
686 consisting of <17 contigs (estimated completeness >99%) are marked with asterisks.
687 Bootstrap values (100 repetitions) are indicated at the nodes, the scale bar at the bottom
688 indicates 20% sequence divergence. The genome sizes for all strains are shown with circles
689 of proportional size and GC content is depicted within each circle. (b) Fragment recruitment
690 of public metagenomes from freshwater sediments (n=131), lake pelagial (n=345), rivers
691 (n=43), estuaries and coastal oceans (n=53), and open oceans (n=201). Maximum RPKG
692 values (number of reads recruited per kb of genome per Gb of metagenome) for each
693 ecosystem are shown for each genome.

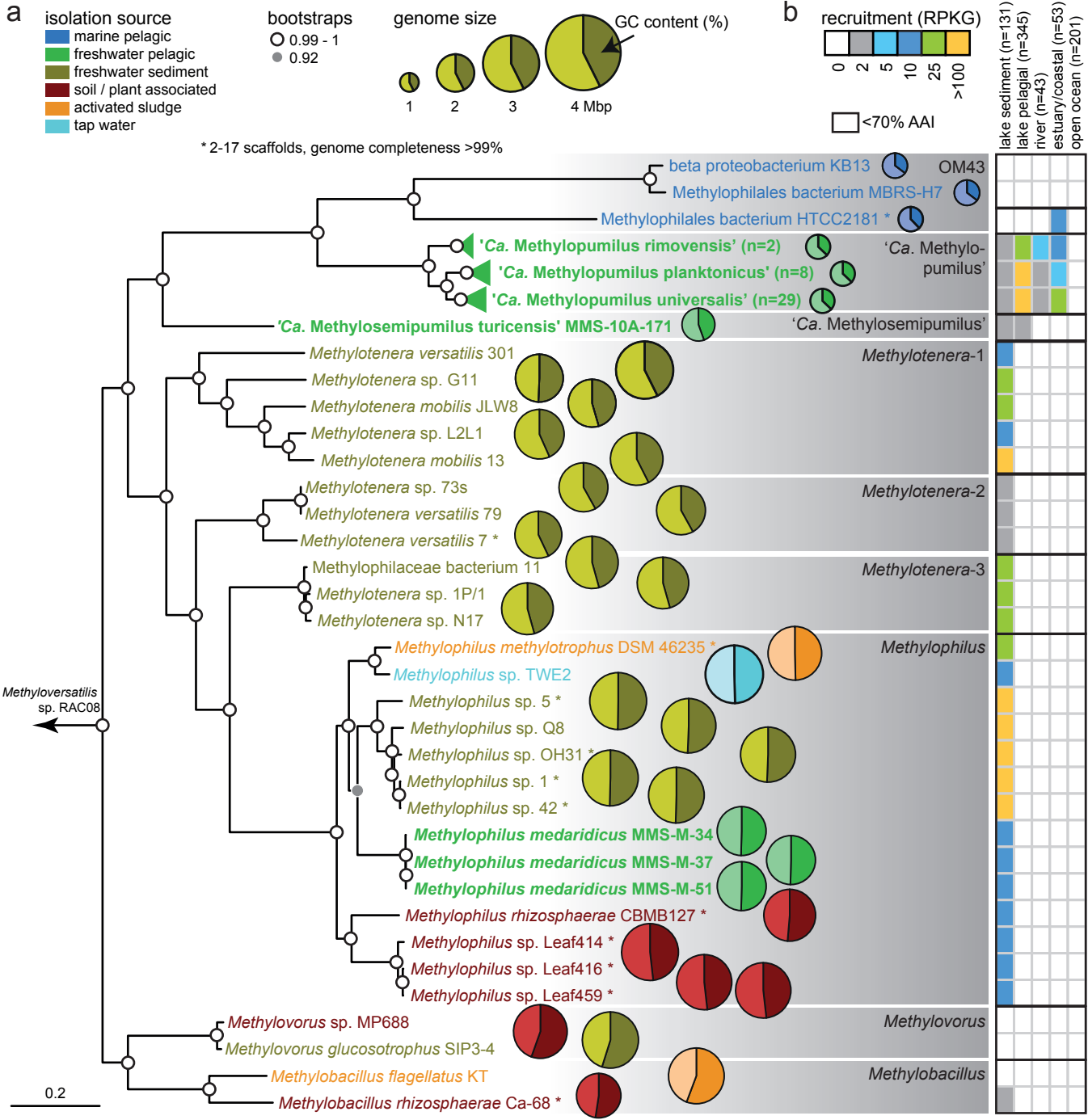
694 **Figure 2: Genome streamlining in Methylophilaceae.** Significant relationships between
695 genome sizes (Mbp) and genomic GC content (%), lengths of intergenic spacers (bp), coding
696 density (%), mean CDS length (bp), overlapping CDS (%), number of paralogs, histidine
697 kinases and sigma factors and significant relationships between genomic GC content (%)
698 and TAA stop codon, TAG stop codon, lysine, and arginine usage (%) and C, N, and S
699 atoms per amino acid (mol%).

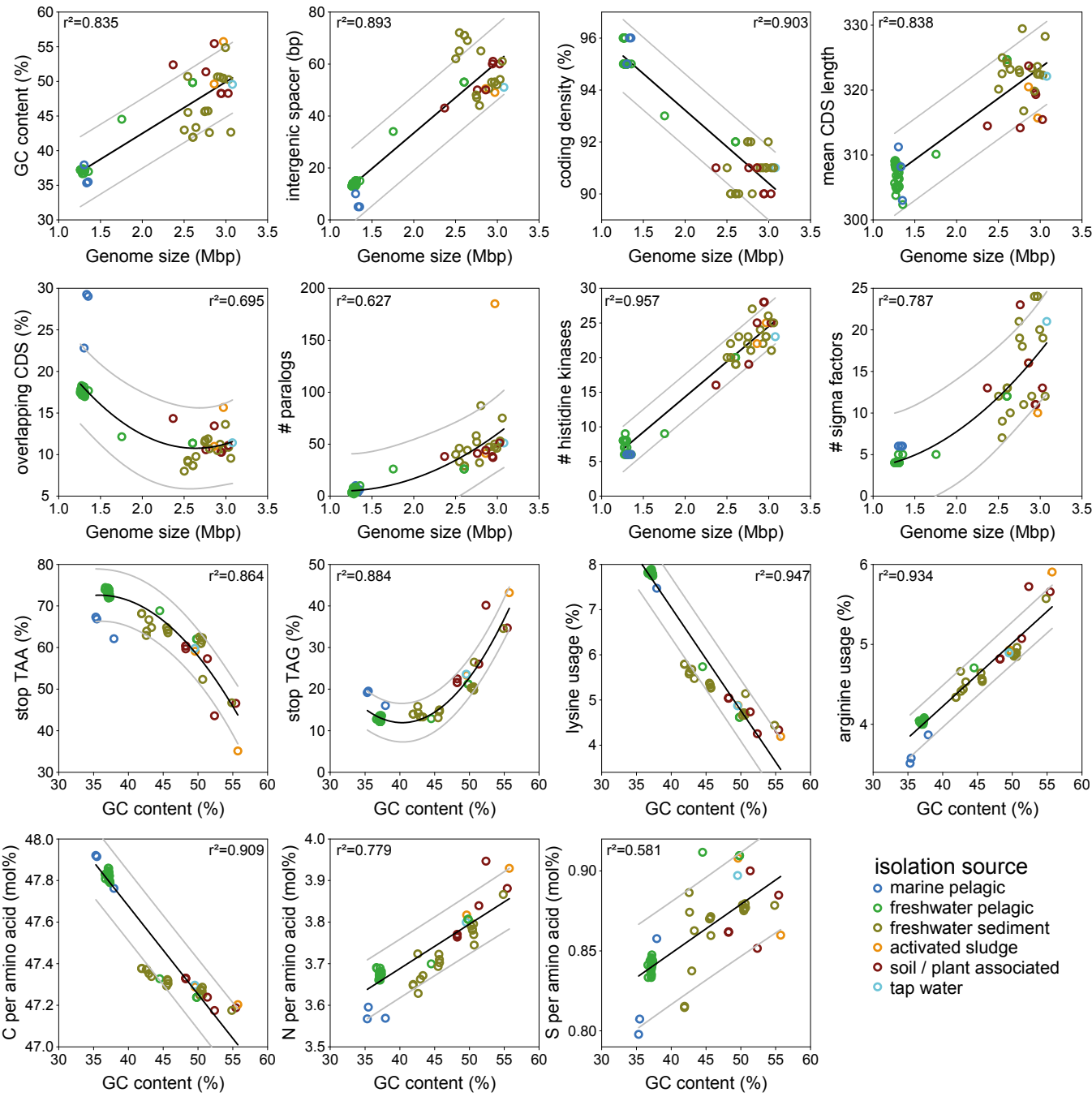
700 **Fig. 3: Metabolic modules in Methylophilaceae.** Presence of selected metabolic modules
701 in Methylophilaceae strains. For details on phylogenomic tree see Fig. 1 and Fig. S1, for
702 details on pathways see Table S4. T4SS: type 4 secretion system; MOX-PQQ: methanol
703 oxidation and pyrroloquinoline quinone biosynthesis; XoxF/MxaF: methanol dehydrogenase
704 XoxF/MxaF; NMG: N-methyloglutamate pathway; MADH: methylamine dehydrogenase;
705 FaDH: formaldehyde dehydrogenase; H4F-folD: H_4-linked formaldehyde oxidation, folD form;

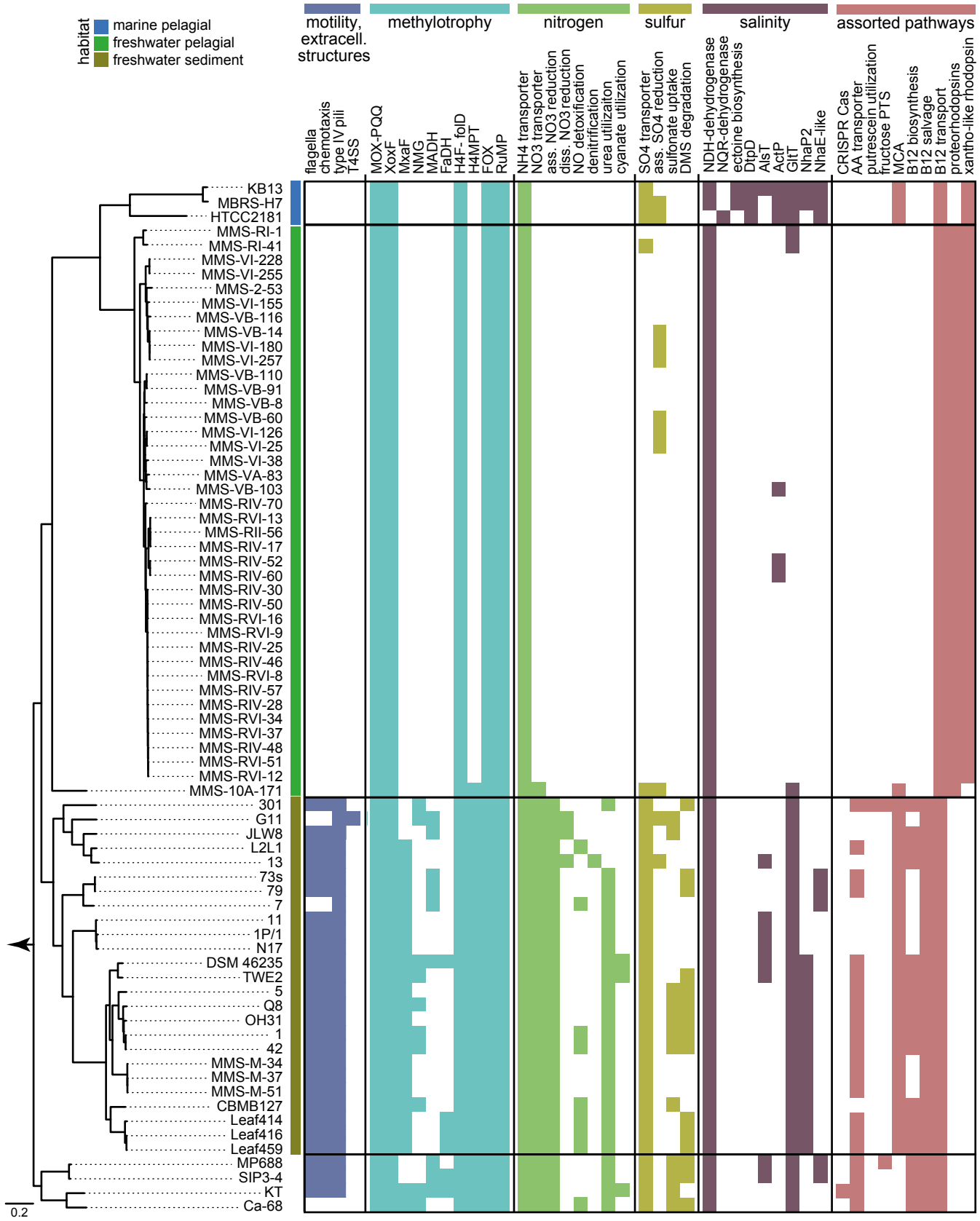
H4MPT: H_4MPT-linked formaldehyde oxidation; FOX: formate oxidation; RuMP: ribulose monophosphate cycle; ass. NO3 reduction: assimilatory nitrate reduction; diss. NO3 reduction: dissimilatory nitrate reduction; ass. SO4 reduction: assimilatory sulfate reduction; NDH-dehydrogenase: H^+-translocating NADH:quinone oxidoreductase; NQR-dehydrogenase: Na^+-translocating NADH:quinone oxidoreductase; DtpD: dipeptide/tripeptide permease DtpD; AlsT: sodium:alanine symporter AlsT; ActP: sodium:acetate symporter ActP; GltT: sodium:dicarboxylate symporter GltT; NhaP2: sodium:proton antiporter NhaP2; NhaE-like: sodium:proton antiporter NhaE; AA transporter: amino acid transporters; fructose PTS: fructose-specific phosophotransferase system; MCA: methylcitric acid cycle.

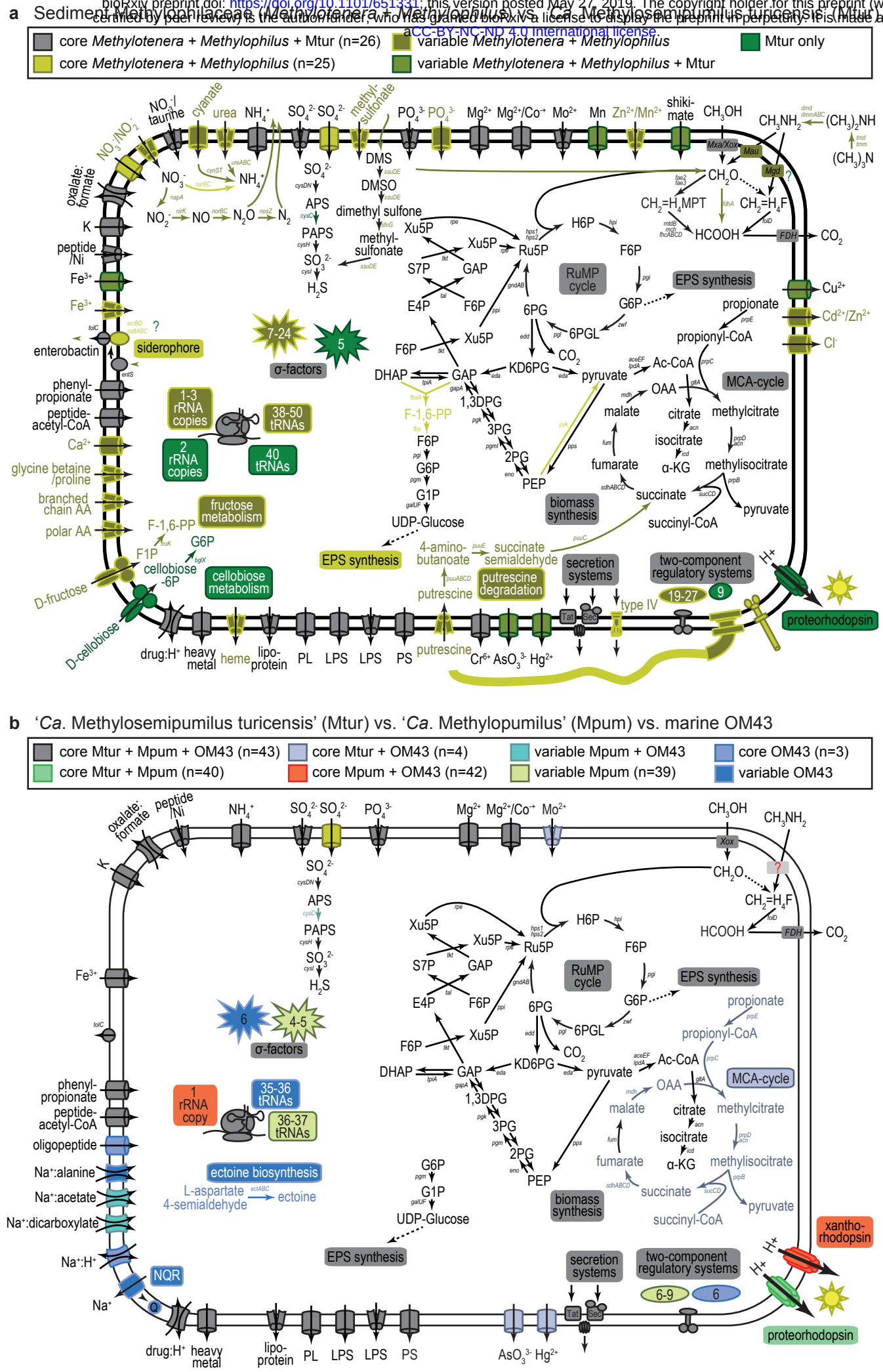
Figure 4: Comparative metabolic maps of different taxa of Methylophilaceae. (a) Comparison of the core metabolism in sediment Methylophilaceae (*Methylotenera* and *Methylophilus*) vs. '*Ca*. Methylosemipumilus turicensis' (Mtur). (b) Comparison of the core metabolism in '*Ca*. Methylosemipumilus turicensis' (Mtur) vs. '*Ca*. Methylopumilus' (Mpum) vs. marine OM43. For details on pathways see Table S4.

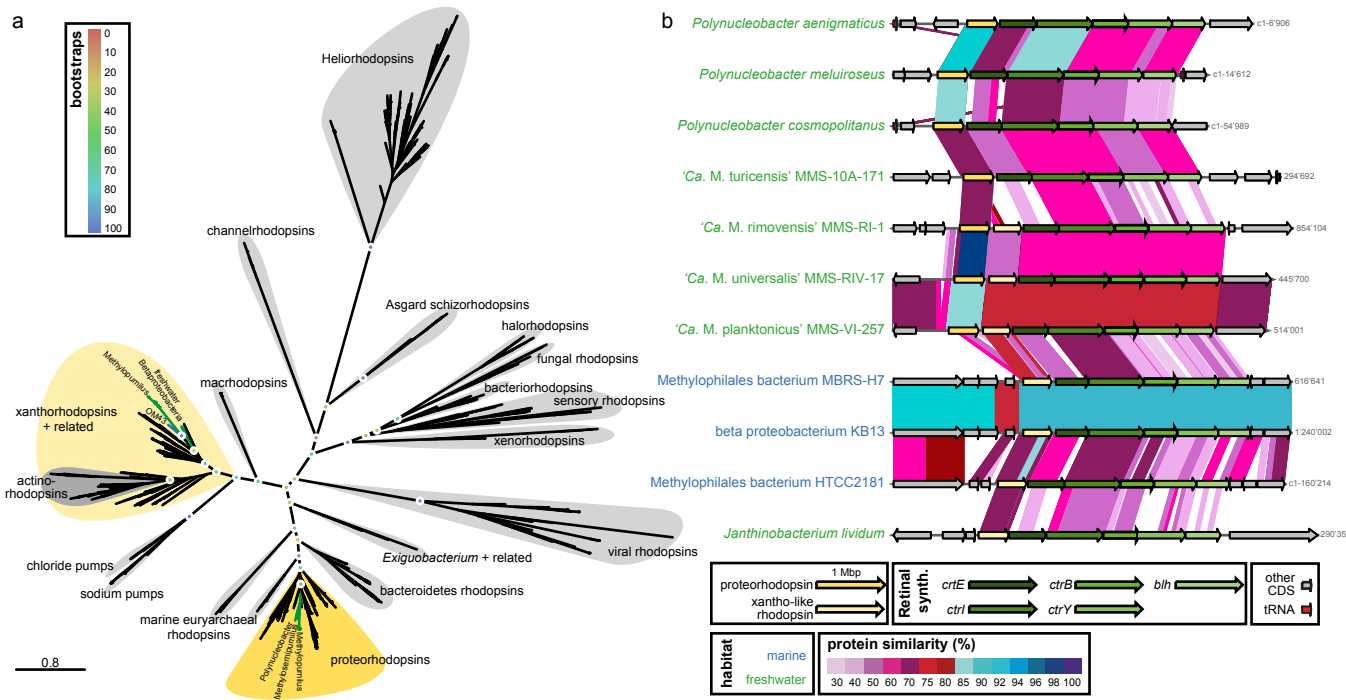
Figure 5: Horizontal gene transfers of two different rhodopsins. (a) Phylogenetic tree (RAxML, 100 bootstraps) of different rhodopsin types. See Fig. S12 for details of closely related proteo- and xantho-like rhodopsins of Methylophilaceae. (b) Arrangement and protein similarity of rhodopsins and the retinal biosynthesis gene cluster in '*Ca*. Methylopumilus spp.', '*Ca*. Methylosemipumilus turicensis', marine OM43 and other freshwater microbes with closely related rhodopsin types.


a
isolation source
marine pelagic
freshwater pelagic
freshwater sediment
soil / plant associated
activated sludge
tap water
bootstraps
0.99 - 1
0.92
genome size
GC content (%)
1
2
3
4 Mbp
b
recruitment (RPKG)
0
2
5
10
25
>100
<70% AAI
lake sediment (n=131)
lake pelagial (n=345)
river (n=43)
estuary/coastal (n=53)
open ocean (n=201)
* 2-17 scaffolds, genome completeness >99%
beta proteobacterium KB13
Methylophilales bacterium MBRS-H7
Methylophilales bacterium HTCC2181 *
OM43
'Ca. Methylopumilus rimovensis' (n=2)
'Ca. Methylopumilus planktonicus' (n=8)
'Ca. Methylopumilus universalis' (n=29)
'Ca. Methylopumilus'
'Ca. Methylosemipumilus turicensis' MMS-10A-171
'Ca. Methylosemipumilus'
Methylotenera versatilis 301
Methylotenera sp. G11
Methylotenera mobilis JLW8
Methylotenera sp. L2L1
Methylotenera mobilis 13
Methylotenera-1
Methylotenera sp. 73s
Methylotenera versatilis 79
Methylotenera versatilis 7 *
Methylotenera-2
Methylophilaceae bacterium 11
Methylotenera sp. 1P/1
Methylotenera sp. N17
Methylotenera-3
Methylophilus methylotrophus DSM 46235 *
Methylophilus sp. TWE2
Methylophilus sp. 5 *
Methylophilus sp. Q8
Methylophilus sp. OH31 *
Methylophilus sp. 1 *
Methylophilus sp. 42 *
Methylophilus medaridicus MMS-M-34
Methylophilus medaridicus MMS-M-37
Methylophilus medaridicus MMS-M-51
Methylophilus rhizosphaerae CBMB127 *
Methylophilus sp. Leaf414 *
Methylophilus sp. Leaf416 *
Methylophilus sp. Leaf459 *
Methylophilus
Methyloversatilis sp. RAC08
Methylovorus sp. MP688
Methylovorus glucosotrophus SIP3-4
Methylovorus
Methylobacillus flagellatus KT
Methylobacillus rhizosphaerae Ca-68 *
Methylobacillus
0.2


r²=0.835
GC content (%)
Genome size (Mbp)
r²=0.893
intergenic spacer (bp)
Genome size (Mbp)
r²=0.903
coding density (%)
Genome size (Mbp)
r²=0.838
mean CDS length
Genome size (Mbp)
r²=0.695
overlapping CDS (%)
Genome size (Mbp)
r²=0.627
# paralogs
Genome size (Mbp)
r²=0.957
# histidine kinases
Genome size (Mbp)
r²=0.787
# sigma factors
Genome size (Mbp)
r²=0.864
stop TAA (%)
GC content (%)
r²=0.884
stop TAG (%)
GC content (%)
r²=0.947
lysine usage (%)
GC content (%)
r²=0.934
arginine usage (%)
GC content (%)
r²=0.909
C per amino acid (mol%)
GC content (%)
r²=0.779
N per amino acid (mol%)
GC content (%)
r²=0.581
S per amino acid (mol%)
GC content (%)
isolation source
marine pelagic
freshwater pelagic
freshwater sediment
activated sludge
soil / plant associated
tap water


habitat
marine pelagial
freshwater pelagial
freshwater sediment
motility, extracell. structures
methylotrophy
nitrogen
sulfur
salinity
assorted pathways
flagella
chemotaxis
type IV pili
T4SS
MOX-PQQ
XoxF
MxaF
NMG
MADH
FaDH
H4F- folD
H4MPT
FOX
RuMP
NH4 transporter
NO3 transporter
ass. NO3 reduction
diss. NO3 reduction
NO detoxification
denitrification
urea utilizaiton
cyanate utilization
SO4 transporter
ass. SO4 reduction
sulfonate uptake
DMS degradation
NDH-dehydrogenase
NQR-dehydrogenase
ectoine biosynthesis
DtpD
AlsT
ActP
GltT
NhaP2
NhaE-like
CRISPR Cas
AA transporter
putrescein utilization
fructose PTS
MCA
B12 biosynthesis
B12 salvage
B12 transport
proteorhodopsins
xantho-like rhodopsin
KB13
MBRS-H7
HTCC2181
MMS-RI-1
MMS-RI-41
MMS-VI-228
MMS-VI-255
MMS-2-53
MMS-VI-155
MMS-VB-116
MMS-VB-14
MMS-VI-180
MMS-VI-257
MMS-VB-110
MMS-VB-91
MMS-VB-8
MMS-VB-60
MMS-VI-126
MMS-VI-25
MMS-VI-38
MMS-VA-83
MMS-VB-103
MMS-RIV-70
MMS-RVI-13
MMS-RII-56
MMS-RIV-17
MMS-RIV-52
MMS-RIV-60
MMS-RIV-30
MMS-RIV-50
MMS-RVI-16
MMS-RVI-9
MMS-RIV-25
MMS-RIV-46
MMS-RVI-8
MMS-RIV-57
MMS-RIV-28
MMS-RVI-34
MMS-RVI-37
MMS-RIV-48
MMS-RVI-51
MMS-RVI-12
MMS-10A-171
301
G11
JLW8
L2L1
13
73s
79
7
11
1P/1
N17
DSM 46235
TWE2
5
Q8
OH31
1
42
MMS-M-34
MMS-M-37
MMS-M-51
CBMB127
Leaf414
Leaf416
Leaf459
MP688
SIP3-4
KT
Ca-68
0.2




a
bootstraps
0
10
20
30
40
50
60
70
80
90
100
Heliorhodopsins
channelrhodopsins
Asgard schizorhodopsins
halorhodopsins
fungal rhodopsins
bacteriorhodopsins
sensory rhodopsins
xenorhodopsins
macrhodopsins
xanthorhodopsins
+ related
actino-
rhodopsins
chloride pumps
sodium pumps
marine euryarchaeal
rhodopsins
proteorhodopsins
viral rhodopsins
Exiguobacterium + related
bacteroidetes rhodopsins
0.8
b
Polynucleobacter aenigmaticus
Polynucleobacter meluiroseus
Polynucleobacter cosmopolitanus
'Ca. M. turicensis' MMS-10A-171
'Ca. M. rimovensis' MMS-RI-1
'Ca. M. universalis' MMS-RIV-17
'Ca. M. planktonicus' MMS-VI-257
Methylophilales bacterium MBRS-H7
beta proteobacterium KB13
Methylophilales bacterium HTCC2181
Janthinobacterium lividum
c1-6'906
c1-14'612
c1-54'989
294'692
854'104
445'700
514'001
616'641
1'240'002
c1-160'214
290'35
1 Mbp
proteorhodopsin
xantho-like
rhodopsin
Retinal
synth.
crtE
crtI
ctrB
ctrY
blh
other
CDS
tRNA
habitat
marine
freshwater
protein similarity (%)
30 40 50 60 70 75 80 85 90 92 94 96 98 100