Fishing for biodiversity: novel methanopterin-linked C_1 transfer genes deduced from the Sargasso Sea metagenome

Research Article

Marina G. Kalyuzhnaya[1], Olivier Nercessian[1], Alla Lapidus[2], and Ludmila Chistoserdova[1]

[1]Department of Chemical Engineering, University of Washington, Seattle WA 98195

[2]Joint Genome Institute

Corresponding author: Ludmila Chistoserdova. Mailing address: Department of Chemical Engineering, University of Washington, Box 352125, Seattle, WA 98195-2125. Phone: 206-543-6682, FAX: 206-616-5721, E-mail: milachis@u.washington.edu.

Key words: Sargasso Sea, metagenome, tetrahydromethanopterin, formaldehyde oxidation

Running title: Methanopterin-linked C_1 transfer genes in the Sargasso Sea

Abstract

The recently generated database of microbial genes from an oligotrophic environment populated by a calculated 1,800 of major phylotypes (the Sargasso Sea metagenome) presents a great source for expanding local databases of genes indicative of a specific function. In this paper we analyze the Sargasso Sea metagenome in terms of the presence of methanopterin-linked C_1 transfer genes that are signature for methylotrophy. We conclude that more than 10 phylotypes possessing genes of interest are present in this environment, and a few of these are relatively abundant species. The sequences representative of the major phylotypes do not appear to belong to any known microbial group capable of methanopterin-linked C_1 transfer. Instead, they separate from all known sequences on phylogenetic trees, pointing towards their affiliation with a novel microbial phylum. These data imply a broader distribution of methanopterin-linked functions in the microbial world than previously known.

Introduction

Biodiversity is a widely used term in modern biological science (Irigoien, Huisman andHarris, 2004; Nee, 2004), including microbiology, employed to describe the variety of species or phylotypes in a given environment, sometimes accompanied by attempts to deduce functional significance of those phylotypes (Pace, 1997; Rodriguez-Valera, 2002). The description of biodiversity in the microbial world is complicated by one well-recognized obstacle, of most extant microbial species being unknown, due to either their unculturability, or to the limited sampling (Staley and Konopka, 1985; Pace, 1997; Rappe and Giovannoni, 2003). It has been becoming more apparent, however, that true understanding of prokaryotic diversity and ultimately of prokaryotic evolution can only be gained based on the knowledge encompassing a much wider variety of microbes. It is also becoming more apparent that describing prokaryotic diversity based only on 16S rRNA data has severe limitations (Rodriguez-Valera, 2002). The emerging field of environmental genomics opens a window of opportunity for not only detecting a variety of uncultured microbes in a given environment, but also for analyzing the variety and the diversity of physiological activities, metabolic pathways and fitness/ survival strategies used by microbes, as deduced from environmental sequences. However daunting the task might be, the work on describing microbial communities to a great depth of coverage has been recently pioneered, using the approach of shot-gun environmental sequencing (Tyson et al., 2004; Venter et al., 2004). While the community described by Tyson et al. presents a model for highly specialized communities with limited biodiversity, the environment sampled by Venter et al. presents a model for communities with a great breadth of diversity, which will likely be found in other natural environments, such as fresh waters and soils. In fact, the database of sequences generated in the latter work has increased the number of

sequences in the non-redundant protein database by an order of magnitude, presenting, on one hand, a great challenge to the community striving to achieve reliable annotation for every gene, on another hand, a great opportunity to use this database as a free resource for expanding databases of functional genes and searching for missing links in limited databases, to gain a better understanding of how genes encoding specific physiological functions might have evolved.

A set of genes and enzymes involved in tetrahydromethanopterin (H_4MPT)-linked C_1 transfer reactions has been recognized recently as one of the major methylotrophy metabolic modules, the formaldehyde oxidation (FOX) module (Chistoserdova et al., 1988; Vorholt et al., 1999; Chistoserdova et al., 2003). Some of the genes and enzymes involved in this module are similar to the genes and enzymes involved in methanogenesis and sulfate reduction by anaerobic archaea, implying a commonality in the evolutions for the two bioconversions (Chistoserdova et al., 2004). More recently, genes involved in the FOX module have been also recognized in the genomes of the Planctomycetes (Glökner et al., 2003; Chistoserdova et al., 2004), implying that these bacteria may be active in metabolizing formaldehyde. Uncovering the phylogenetic divergence of the components of the FOX module present in Proteobacteria, Planctomycetes and methanogenic and sulphate-reducing Archaea has cast doubts on the previously favored hypothesis of lateral transfer of respective genes between archaeal methanogens and proteobacterial methylotrophs, the hypothesis mostly based on the presumably limited distribution of these genes (Chistoserdova et al., 2004). However, to reconstruct the evolution of these genes in both Archaea and Bacteria with more precision, more divergent sequences need to be included in the analyses. We employed the FOX metabolic module in this work to pursue two major objectives: (1) to test if shot-gun environmental databases, such as the Sargasso Sea

metagenome are amenable to functional predictions, specifically for sequences not closely related to sequences from known microbial groups, and (2) to test if such databases present useful sources of novel sequences with predictable functions, thus carrying a potential for discovering novel microbial groups and linking them to a function in the environment.

Materials and methods

Analysis of the SSM database

Sequences of genes and polypeptides potentially involved into the FOX module were determined in the SSM database (http://www.ncbi.nlm.nih.gov/BLAST/Genome/EnvirSamplesBlast.html), via BLAST searches using 17 protein queries, as shown in Table 1. Sequences representative of both Proteobacteria and Planctomycetes were used as queries. Candidate homologs then were used as queries against the non-redundant database at NCBI (http://www.ncbi.nlm.nih.gov/) as well as against our proprietary databases of respective subsets of manually identified and curated genes, to ensure that indeed they were closest homologs to the genes involved in H_4MPT-linked C_1 transfers. For homologs that have passed this test, affiliations with contigs and scaffolds were determined, and respective contigs and scaffolds analyzed. Scaffolds bearing FOX genes were first matched against the database of scaffolds associated with particular organisms (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide&cmd=Search&term=CH004436:CH004736[PACC]), as determined by Venter et al. (2004), but none of them was present in this database. The scaffolds of interest were then retrieved from the database of scaffolds not associated with any particular organism (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide&cmd=Search&term=CH004737:CH236877[PACC]), and every contig in respective scaffolds was manually analyzed, via

BLASTP analysis of translated polypeptides against non-redundant database, to determine how closely these sequences were related to known sequences (http://www.ncbi.nlm.nih.gov/BLAST/). To search for primary C_1 oxidation genes in the SSM database, polypeptide sequences of the most conserved proteins, the alpha subunit of particulate methane monooxygenase (MMO, PmoA), the alpha subunit of soluble MMO (MmoX), the alpha subunit of methanol dehydrogenase (MxaF), and the beta subunit of methylamine dehydrogenase (MauB) representative of different groups of proteobacterial methylotrophs were used as BLAST queries, and identity of top hits determined. Resulting from this analysis, PmoA top hits were more related to the homologous AmoA protein (Holmes et al., 1995) than to PmoA, no significant similarity was found with MmoX, MxaF top hits showed closer relationship to the MxaF homolog (also known as XoxF) not involved in methanol oxidation (Chistoserdova and Lidstrom, 1997), and MauB top hits were more closely related to uncharacterized proteins from *Burkholderia cepacia, Ralstonia eutropha, Rubrivivax gelatinosus* and *Novosphingobium* species, bacteria that have not been characterized as methylamine utilizers, but showed only 31% identity with MauB polypeptides. To search for the presence of 16S rDNA sequences similar to the ones for known methylotrophic groups, including the groups previously identified in the Sargasso Sea (Sieburth et al., 1987; 1993), sequences representative of the following genera were used in BLASTN analyses: *Methylobacterium, Hyphomicrobium, Methylosynus, Methylocystis* (α-proteobacteria), *Methylophaga*, *Methylobacillus* (β-proteobacteria), *Methylomicrobium, Methylomonas, Methylobacter, Methylosarcina, Methylococcus*, *Methylosphaera* (γ-Proteobacteria). 16s rDNA sequences representative of the mayor groups of Planctomycetes (*Gemmata, Planctomyces, Pirellula* and *Isosphaera*) were also used in direct BLAST analyses.

Species richness analysis

For species richness analysis, a cut-off of 97% for rRNA genes was used (Stackebrandt and Goebel, 1994) to conclude on the presence of known species, and a cut-off of approximately 95% at the protein level (approximately 94% at the DNA level, Venter et al., 2004) was used to define phylotypes.

Genomic analysis of Methylobacillus flagellatus

The genome of *M. flagellatus*, a β-proteobacterial methylotroph, has been recently sequenced by the Joint Genome Institute under the DOE funding, and first draft of the genome is publicly available (http://genome.jgi-psf.org/draft_microbes/metfl/metfl.home.html). To expand the databases of β-proteobacterial FOX genes, we identified these genes in the genome, via BLAST searches using polypeptide sequences of *Methylobacterium extorquens* (Chistoserdova et al., 2003) as queries. The genes listed in Table 1 were identified in two gene clusters in the genome, ordered as follows: *fhcC-D-A-B-mptG-mtdB-orfY-mch-orf5-orf7-tal-hps-hpi-fae-orf17-orf1-orf9-pabB-orf21*, and *orf20-orf19-orf22*. Three additional homologs of *fae* were also identified that were not parts of the FOX gene clusters. One of them was highly similar to the *fae* gene in the FOX cluster, while two others were similar to *fae2* and *fae3* genes of unknown function previously identified in *M. extorquens* (Chistoserdova et al., 2004).

Expression of fae3

To test if *fae3* could fulfill the function of *fae*, *fae3* from *M. flagellatus* was cloned into the expression vector pCM80 (Marx and Lidstrom, 2001), under a strong promoter for expression in *M. extorquens*, and introduced into the *fae* mutant of *M. extorquens* that is negative for growth on methanol and sensitive to methanol vapors (Vorholt et al., 2000). Fae3 was not able to restore wild type growth of the mutant, or to alleviate its sensitivity to methanol.

Phylogenetic analysis

Translated amino acid sequences were aligned using the ClustalW program (Thompson *et al*., 1994). For phylogenetic analysis the Phylip package (Felsenstein, 2003) was used. Distance and parsimony methods were employed; 100 bootstrap analyses were performed. Concatenated polypeptide sequences were produces as follows. Separate polypeptide sequences (or truncated sequences) were aligned using ClustalW program, and all the sequences were truncated to be of the same length. Then the truncated polypeptide sequences of each organism were fused together (in order Orf20-Orf19-Orf22-Orf17-Fae-Orf7-Orf5-Mch-MptG), and the concatenated sequences re-aligned using ClustalW program, and the alignments manually curated. The respective polypeptide sequences were retrieved from the respective genomic databases as described above for *M. flagellatus* and as described in Chistoserdova et al. (2004).

RESULTS

Methanopterin-linked formaldehyde oxidation genes in the Sargasso Sea metagenome

A total of 17 conserved genes have been identified that are involved in H_4MPT-linked reactions as well as cofactor (H_4MPT or methanofuran analog) biosynthesis (Chistoserdova et al, 2003; unpublished results). We used the respective 17 polypeptide queries (Table 1), to detect the FOX metabolic module in the Sargasso Sea metagenome (SSM). Homologs for all 17 polypeptideses were recognizable in the metagenome, their numbers varying between only two for *orf21* (one of the least conserved genes in the module) and 24 for *fae* (one of the most conserved genes in the module). However, of the 24 *fae* homologs, only 11 appeared to be "true" *fae*, based on phylogenetic analysis (see below). 22 homologs were detected for FhcA, but only 11 to 15 homologs were detected for the genes encoding three other subunits of

formyltransferase/hydrolase enzyme (Pomper et al., 2002). The most accurate estimate of the number of organisms possessing the FOX module would probably be based on the number of homologs of the *mch* gene, as no *mch* duplication has been observed in the known bacterial genomes encoding the FOX module enzymes, or in archaeal genomes encoding homologous functions (Reeve, 1997; unpublished results). Besides, *mch* is a highly conserved, highly specific gene, thus it should be recognized with high confidence. 12 *mch* homologs were identified in the SSM, six as parts of genomic scaffolds containing other FOX genes, and 6 translated from singleton reads. This number is likely to be the most accurate lowest estimate for the number of organiss in the Sargasso Sea possessing the FOX module.

Species richness

We analyzed the sequence divergence of the newly identified homologs in each of the 17 gene groups, by direct DNA/DNA (not shown) and protein/protein BLAST analyses (Table 1). Sequences representative of α, β, and γ Proteobacteria, planctomycetes and methanogenic archaea were included in these analyses, as references. Significant gene divergence was found in every gene group analyzed, pointing towards great species richness (Table 1). We exemplify the outcomes of such analyses in Table 2, for Mch homologs. It has been argued before that *mch* might be one of the best targets in the H_4MPT-linked C_1 transfer pathway for phylogenetic comparisons (Reeve et al., 1997, Chistoserdova et al., 2004), thus it should also be a good target for assessing phylotype diversity. Of the 12 homologs of *mch* identified in the metagenome, only 2 were similar enough to represent one single phylotype, assuming a cutoff for functional genes at 94% at the DNA level (Venter et al., 2004) and approximately 95% at the protein level. The sequences representing the unique 11 Mch phylotypes formed a total of 6 distinct groups, with a

cut-off of only 71% for the major group (Group 1), based on protein-protein alignments (Table 2). We have recently generated a large database of Mch sequences from Proteobacteria belonging to the α, β, and γ groups (Kalyuzhnaya, Lidstrom and Chistoserdova, 2004). Mch sequences are also available for a number of methanogenic and sulphate-reducing archaea (http://www.ncbi.nlm.nih.gov/), as well as two representatives of Planctomycetes (Glöckner et al., 2003; http://www.tigr.org/). Representatives of all 6 groups in general showed low identity with the known Mch sequences. For four of these groups, identities with proteobacterial sequences were higher than with planctomycete or archaeal sequences (but not exceeding 61%), for one of the groups, identities were higher with planctomycete sequences (but not exceeding 46%), and for one remaining group, identities were slightly higher with archaeal sequences (up to 43%). We assume that all the homologs do encode Mch enzymes, based on the previous data on remarkably similar properties of bacterial and archaeal Mch enzymes, while bacterial and archaeal Mch polypeptides only share 36 to 39% sequence identity (Vaupel, Vorholt and Thauer, 1988; Pomper et al., 1999). Meaningful phylogenetic analysis of the new Mch sequences was not possible, due to both the partial nature of most sequences, and to their great divergence resulting in low "taxonomical signal". However, we were able to observe a few specific trends. In both parsimony and distance analyses, sequences of the main group (Group 1) separated with high bootstrap support from all known sequences, sequences of group 3 clustered with planctomycete sequences with moderate bootstrap support, but positions of branches representing groups 2, 4, 5 and 6 could not be resolved (data not shown). Pair-wise comparisons and phylogenetic analyses for the remaining 16 groups of sequences produced similar results, and followed trends similar to the ones described for Mch sequences. For most polypeptides, up to six distinct sequence groups were identified based on protein-protein alignments. Most of these sequences formed the major

group, separating into a novel branch on phylogenetic trees. In cases of Mch, Mtd, FhcA, FhcC, MptG, Fae, Orf5, Orf9, Orf20, and OrfY, a group was identified clustering with planctomycete sequences. In most cases, sequences not belonging to the major group or clustering with planctomycetes were affiliated with the bacterial sequences, but poorly resolved from the known sequences (data not shown).

The group of *fae* homologs identified in this study requires a special discussion. The function of *fae* located in the FOX gene cluster, in reaction between formaldehyde and H_4MPT, has been originally described in *M. extorquens* AM1 (Vorholt et al., 2000), and later a large database of *fae* genes from bacteria has been built (Kalyuzhnaya, Lidstrom and Chistoserdova, 2004). However, two more homologs of *fae* have been identified in the genome of *M. extorquens* AM1, designated as *fae2* and *fae3* (Chistoserdova et al., 2004). A close homolog for *fae3* has been also identified in the genome of *Pirellula* sp. Strain 1 (Glöckner et al., 2003). In this work, we identified homologs for both *fae2* and *fae3* in the recently sequenced genome of *M. flagellatus* KT (see Materials and Methods). The new sequences were compared to all known groups of Fae homologs. Of the *24* Fae homologs identified in the SSM, only 11 clustered with true Fae, based on pair-wise gene and protein comparisons and phylogenetic analysis (data not shown). Of the remaining 13, 12 clustered with *fae3*. Such abundance of *fae3* in the SSM prompted us to test if it able to fulfill the function of *fae*, in catalysis of reaction between H_4MPT and formaldehyde (see Materials and Methods). We obtained negative results suggesting a different function for *fae3*. One remaining Fae homolog clustered with the Fae-like polypeptides translated from the genomes of *Burkholderia cepacia* (http://www.sanger.ac.uk/Projects/B_cenocepacia/) and *Ersinia pestis* (http://www.sanger.ac.uk/Projects/Y_pestis/). These sequences represent the fourth group of

bacterial *fae*-like genes (tentatively designated as *fae4*), and their function also remains unknown. While homologs of *fae3* and *fae4* were not clustered with other genes in the FOX module, many of the homologs of true *fae* were found clustered with other FOX genes ("FOX islands", Table 1).

More detail should also be given on the nalysis of *mtd* homologs. Two homologs have been characterized in bacteria, *mtdA* and *mtdB*, encoding methylene-H_4MPT dehydrogenase enzymes with overlapping specificities (Vorholt et al., 1998; Hagemeier et al., 2000). *mtdB* genes have been identified in all proteobacteria possessing the FOX module, resulting in a rich MtdB database, while *mtdA* genes seem to be less widespread in proteobacteria (Kalyuzhnaya, Lidstrom, and Chistoserdova, 2004; unpublished results). In Planctomycetes, however, only *mtdA* homologs are found, and these are clustered with *fae* homologs in the genomes (Glöckner et al., 2003; unpublished results). Of the 12 Mtd homlogs identified in the SSM (Table 1), two showed higher similarity to MtdA. One of the genes was paired with a *fae* homolog, as is typical of planctomycetes, and the translated polypeptide revealed high level of similarity to planctomycete MtdA enzymes. The second MtdA homolog was translated from a singleton read, and showed high similarity to MtdA from *Methylococcus capsulatus*, a γ-proteobacterial methylotroph. The remaining Mtd homologs showed higher similarity to MtdB proteins, and these were further separated into four major groups, based on sequence similarities (data not shown).

Species abundance

The abundance of species or phylotypes in the metagenome can be roughly estimated from the depth of sequence coverage, which, in turn is correlated with the size of the assembled

scaffolds (Venter et al., 2004). We identified 12 scaffolds larger that 3 kb containing the FOX islands (Table 3). Of these, 5 were larger than 10 kb, the largest scaffold being 70,015 kb in size (not counting gaps). For comparison, the largest scaffold for the SAR11 clade, the phylotype implied to account for up to 50% of the microbial community in the Sargasso Sea, based on FISH estimates (Morris et al., 2002), was only 21,000 kb in size in the SSM. The largest scaffold for another abundant marine bacterium, *Prochlorococcus* (Dufense et al., 2003; Rocap et al., 2003) was about 45,000 kb in size. From these comparisons, we can estimate that the major FOX-containing phylotype (represented by SCF 2223320) must be at least as abundant, and possibly more abundant in the site than *Prochlorococcus* or SAR11. Te other major phylotypes are less abundant but soundly present in the site. Most of the identified FOX genes, however, were present on singleton sequencing reads or as parts of small scaffolds, implying limited depth of coverage for the respective phylotypes.

The major phylotypes

The phylotype represented by SCF2223320 is likely to be the major phylotype in the Sargasso Sea containing the FOX module. Genes for 15 out of the 17 protein queries were found on this major scaffold, and one other gene is likely present in the remaining gap between the contigs (Fig. 1). The FOX gene order in this scaffold was remarkably conserved with the gene order in FOX islands known for a number of proteobacteria (Chistoserdova et al., 1998; Marx et al., 2004; unpublished results), exemplified in Fig.1 by the gene cluster in *M. extorquens*. Two other major scaffolds (SCF2229052 and SCF2208802) showed gene clustering that seemed to be identical to the one found for SCF2223320 (Fig. 1). Phylogenetically, genes found on the major scaffolds fell into the dominant, multimember gene sub-groups in every gene group represented

by a sufficient number of sequences (Group 1 for Mch), pointing towards the dominance of one single group of microbes possessing the FOX module, in terms of both abundance and species richness. No 16S rRNA genes were found on any of the scaffolds containing the FOX gene clusters, thus we attempted assessing the phylogenetic position of these microbes, based on the analysis of the genes surrounding the FOX genes. We analyzed a total of 50 non-FOX genes present on the scaffold SCF2223320 and a total of 22 genes on the scaffold SCF2163205. In general, most of them showed low identity with bacterial gene counterparts, on the order of 35-40%, and the identity values were close for representatives of various groups of Proteobacteria, Gram-positive bacteria, or deeply-branching bacterial groups such as *Deinococcus* and *Aquifex*, but were in general lower with archaeal and eukaryotic counterparts. From these comparisons we conclude that the major FOX-containing phylotypes in the Sargasso Sea must represent novel bacteria whose identity remains unknown. This conclusion was further supported by phylogenetic analysis involving FOX polypeptides, of the two novel dominant phylotypes (SCF2223320 and SCF2229052), and counterpart polypeptides representative of Proteobacteria, Planctomycetes, and methanogenic Archaea. To enhance the phylogenetic signal, we used concatenated sequences of 9 FOX polypeptides conserved in all the groups involved in the analysis (See Materials and Methods). In these analyses, the sequences representative of the two novel phylotypes branched separately from the control groups and were positioned between planctomycete and proteobacterial groups (Fig. 2). The only scaffolds that we were able to tentatively place into a phylogenetic context were SCF2162480 and SCF2162604. Besides *fae* and *mtdA*, SCF2162480 contained 10 non-FOX genes, 6 of which showed highest similarity to known planctomycete genes, with identities 22-73% at the amino acid level, while SCF2162604 contained two genes in addition to *mptG*, with highest hits to non-FOX planctomycete genes (49

and 67% identity, respectively, at the amino acid level). This data implies that one of the dominant FOX-containing phylotypes in the SSM must be a planctomycete.

Methylotrophy in the Sargasso Sea

We searched for other traditional methylotrophy genes in the SSM, to test for the presence of genes indicative of the primary methylotrophic substrate, via BLAST analysis (see Materials and Methods). Tests for the presence of genes for particulate or soluble methane monooxygenases or methanol dehydrogenase were negative, implying that the FOX module encoded in the SSM was unlikely a part of methane or methanol metabolism. Tests for methylamine dehydrogenase were also negative. We directly searched the SSM database using 16S rDNA sequences representative of known groups of methylotrophs (see Materials and Methods), including the ones previously isolated from the Sargasso Sea (Sieburth 1997; 1993), or identified via culture-independent approaches (Giovannoni and Rappe, 2000), and did not find any sequences that were more than 89% similar to the sequences of known methylotroph species, thus demonstrating again that the FOX genes identified in the SSM did not belong to any known group of methylotrophs. 16S rDNA queries representing the major groups of Planctomycetes were also used to search the database, and the sequences with most similarity showed only 81-88% identity to the queries, implying that the planctomycete represented by SCF2162480 and SCF 2162604 must be a part of a novel, deeply branching group in this division.

Discussion

In this work we took an advantage of the recently generated, extensive environmental gene database (the SSM) to question the presence, abundance, and diversity of one specific

metabolic module, the H_4MPT-linked formaldehyde oxidation (FOX) module. This module only recently emerged as the main methylotrophy module (Chistoserdova et al., 1998; Vorholt et al., 1999), but more recently it has been also found in bacteria not traditionally considered methylotrophs (Glöckner et al., 2003; Chistoserdova et al., 2004; Marx et al., 2004). However, the physiological significance of this module beyond methylotrophs remained largely unexplored, and the understanding of its evolution uncertain. We demonstrate here that more than 10 novel phylotypes containing the FOX module were present in the Sargasso Sea sample, a few of them abundant species. For comparison, only 37 types of RubisCo genes have been detected in the same sample (Venter et al., 2004), given that the Calvin-Benson-Bassham cycle must be the major process in cyanobacteria abundantly present in the site. Comparison of the scaffold sizes for the newly discovered species possessing the FOX module with the scaffold sizes for some abundant marine species, such as *Prochlorococcus* and SAR11 implies that these novel FOX-containing species must be as abundant in the sea. Of the novel FOX-possessing phylotypes, one seems to be a deeply-branching planctomycete species, likely representing a new genus within this division, while the remaining phylotypes do not fall within any known groups of bacteria. Although methylotrophs are clearly present in the site, based on previous enrichment studies (Sieburth 1987; 1993) as well as culture-independent detection (Giovannoni and Rappe, 2000), they must be present in numbers too low to be represented the database generated by Venter et al. (2004). Based on the absence of the respective primary oxidation genes in the database, the FOX module genes identified in this study do not seem to be associated with methane or methanol oxidation. The exact function of these genes will remain uncertain till these organisms could be isolated, cultivated and studied, but such a sound presence of the FOX module in the Sargasso Sea sample implies its physiological significance. One possible function

of this module may be in formaldehyde oxidation/detoxification. The presence of formaldehyde has been documented in the Sargasso Sea before, and connected to methane oxidation (Johnson, Davis and Sieburth, 1983; Eberhardt an Sieburth, 1985). However the low abundance of methanotrophs, as judged from the absence of their characteristic genes from the metagenome argues against methane oxidation as the main source of formaldehyde. Humic acids, on another hand, present a more prominent source of formaldehyde in marine waters, as they are known to spontaneously break down under light, yielding formaldehyde (Kieber, Zhou and Mopper, 1990). The microbes possessing the FOX module may also be involved in metabolism of methylamine that is a break down product of decaying marine eukaryotes, via the N-methylglutamate pathway (Jones and Bellion, 1991), genes for which remain unknown.

One important outcome of this study is the identification of a novel phylogenetic branch of microbes, presumably bacteria, possessing the H_4MPT-linked C_1 transfer functions. We have recently analyzed the possible scenarios for the evolution of the FOX module (Chistoserdova et al., 2004), and were not able to discriminate between the two most probable scenarios, of this module being present in the last universal common ancestor (LUCA) or emerging in Planctomycetes. The position of the novel group on phylogenetic trees argues in favor of the module's presence in the LUCA, and thus implies its great antiquity. Overall, our data argue for a broader distribution of the FOX module than previously thought.

Acknowledgements

We acknowledge support from the Microbial Observatories Program funded by the National Science Foundation. The Joint Genome Institute is acknowledged for sequencing the genome of *M. flagellatus*, and the Institute for Genomic Research is acknowledged for early release of the

genomic data for *Methylococcus capsulatus* and *Gemmata obscuriglobus* (all projects funded by the Department of Energy).

References

Chistoserdova, L., J.A.Vorholt, R.K. Thauer, and M.E. Lidstrom. 1988. C_1 transfer enzymes and coenzymes linking methylotrophic bacteria and methanogenic Archaea. Science 281:99-102.

Chistoserdova, L., S.-w. Chen, A. Lapidus, and M.E. Lidstrom. 2003. Methylotrophy in *Methylobacterium extorquens* AM1 from a genomic point of view. J. Bacteriol. 185:2980-2987.

Chistoserdova, L., C. Jenkins, M.G. Kalyuzhnaya, C.J. Marx, A. Lapidus, J.A. Vorholt, J.T. Staley, and M.E. Lidstrom. 2004. The enigmatic planctomycetes may hold a key to the origins of methanogenesis and methylotrophy. Mol. Biol. Evol. 21:1234-1241.

Dufresne, A., M. Salanoubat, F. Partensky et al. (22 co-authors). 2003. Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. Proc. Natl. Acad. Sci. USA. 100:10020-10025.

Eberhardt, M.A., and J.M. Sieburth. 1985. A colorimetric procedure for the determination of aldehydes in seawater and in cultures of methylotrophic bacteria. Mar. Chem. 17:199-212.

Felsenstein, J. (2003). Inferring Phylogenies. Sinauer Associates, Inc., Sunderland, Massachusetts.

Giovannoni, S., and M.S. Rappe. 2000. Evolution, diversity, and molecular ecology of marine prokaryotes. In Microbial ecology of the oceans. D.L. Kirchman, Ed., Wiley, New York, pp. 47-84.

Glöckner, F.O., M. Kube, M. Bauer et al. (14 co-authors). Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. 2003. Proc. Natl. Acad. Sci. USA. 100:8298-8303.

Hagemeier, C.H., L. Chistoserdova, M.E. Lidstrom, R.K. Thauer, and J.A. Vorholt. 2000. Characterization of a second methylene tetrahydromethanopterin dehydrogenase from *Methylobacterium extorquens* AM1. Eur. J. Biochem. 267:3762-3769.

Chistoserdova and Lidstrom. 1997. Molecular and mutational analysis of a DNA region separating two methylotrophy gene clusters in Methylobacterium extorquens AM1. Microbiol. 143:1729-1736.

Holmes, A.J., A. Costello, M.E. Lidstrom, and J.C. Murrell. 1995. Evidence that particulate methane monooxygenase and ammonia monooxygenase may be evolutionarily related. FEMS Microbiol. Lett. 132:203-208.

Irigoien, X., J. Huisman, and R.P. Harris. 2004. Global biodiversity patterns of marine phytoplankton and zooplankton. Nature 429:863-867.

Johnson, K.M., P.G. Davis and J.M. Sieburth. 1983. Diel variation of TCO_2 in the upper layer of oceanic waters reflects microbial composition, variation and possibly methane cycling. Mar. Biol. 77:1-10.

Jones J.G., and E. Bellion. 1991. In vivo ^{13}C and ^{15}N NMR studies of methylamine metabolism in *Pseudomonas* species MA. J. Biol. Chem. 266:11705-11713.

Kalyuzhnaya, M.G., M.E. Lidstrom, and L. Chistoserdova. 2004. Utility of environmental probes targeting ancient enzymes: methylotroph detection in Lake Washington. Microb. Ecol. In Press.

Kieber, R.J., X. Zhou, and K. Mopper. 1990. Formation of carbonyl compounds from UV-induced photodegradation of humic substances in natural waters: fate of riverine carbon in the sea. Limnol. Oceanogr. 35:1503-1515.

Marx, C.J., and M.E. Lisdtrom. 2001. Development of improved versatile broad-host-range vectors for use in methylotrophs amd other Gram-negative bacteria. Microbiol. 147:2065-2075.

Marx, C.J., J.A. Miller, L. Chistoserdova, and M.E. Lidstrom. 2004. Multiple formaldehyde oxidation/detoxification pathways in *Burkholderia fungorum* LB400. J, Bacteriol. 186:2173-2178.

Morris, R.M., Rappe, M.S., Connon, S.A., Vergin, K.L., Siebold, W.A. Carlson, C.A., and Giovannoni S.J. 2002. SAR11 clade dominates ocean surface bacterioplankton communities. Nature 420:806-10.

Nee, S. 2004. More than meets the eye. Nature 429:804-805.

Pace, N. A molecular view of microbial diversity and the biosphere. 1997. Science 276:734-740.

Pomper, B.K., O. Saurel, A. Milton, and J.A. Vorholt. 2002. Generation of formate by the formyltransferase/hydrolase complex (Fhc) from *Methylobacterium extorquens* AM1. FEBS Lett. 523:133-137.

Pomper, B.K., J.A. Vorholt, L. Chistoserdova, M.E. Lidstrom, and R.K. Thauer. 1999. A methenyl tetrahydromethanopterin cyclohydrolase and a methenyl tetrahydrofolate cyclohydrolase in *Methylobacterium extorquens* AM1. Eur. J. Biochem. 261:475-80.

Rappe, M.S., and S.J. Giovannoni. 2003. The uncultured microbial majority. Annu. Rev. Microbiol. 57:369-394.

Reeve, J.N., J. Nölling, R. M. Morgan, and D.R. Smith. 1997. Methanogenesis: genes, genomes, and who's on first? J. Bacteriol. 179:5975-5986.

Rocap, G., F.W. Larimer, J. Lamerdin et al. (24 co-authors). 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424:1042-1047.

Rodriguez-Valera, F. 2002. Approaches to prokaryotic biodiversity: a population genetics perspective. Environ. Microbiol. 4:628-633.

Sieburth, J.M., P.W. Johnson, M.E. Eberhardt, M.E. Sieracki, M.E. Lidstrom, and D.C. Laux. 1987. The first methane-oxidizing bacterium from the upper mixing layer of the deep ocean: *Methylomonas pelagica* sp. Nov. Current Microbiol. 14:285-293.

Sieburth, J.M., P.W. Johnson, V.M. Church, and D.C. Laux. 1993. C_1 bacteria in the water column of Chesapeake Bay, USA. III. Immunologic relationships of the type species of marine monomethylamine- and methane-oxidizing bacteria to wild estuarine and oceanic cultures. Mar. Ecol. Progress Series. 95:91-102.

Stackebrandt, E., and B.M. Göebel. 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S sequence rRNA analysis in the present species definition in bacteriology. Int. J. Syst. Bacteriol. 44:846-849.

Staley, J.T., and A. Konopka. 1985. Measurement od in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu. Rev. Microbiol. 39:321-346.

Thompson J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight Matrix choice. Nucl. Acid. Res. 22:4673-4680.

Tyson, G.W., J. Chapman, P. Hugenholtz, E.E. Allen, R.J. Ram, P.M. Richardson, V.V. Solovyev, E.M. Rubin, D.S. Rokhsar, and J.F. Banfield. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37-43.

Vaupel, M., J.A. Vorholt, and R.K. Thauer. 1998. Overproduction and one-step purification of the N5,N10-methenyltetrahydromethanopterin cyclohydrolase (Mch) from the hyperthermophilic *Methanopyrus kandleri*. Extremophiles 2:15-22.

Venter, J.C., K. Remington, J. Heidelberg et Al. (23 co-authors). 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science. 304:66-74.

Vorholt, J.A., L. Chistoserdova, S.M. Stolyar, R.K. Thauer, and M.E. Lidstrom. 1999. Distribution of tetrahydromethanopterin-dependent enzymes in methylotrophic bacteria and phylogeny of methenyl tetrahydromethanopterin cyclohydrolases. J. Bacteriol. 181:5750-5757.

Vorholt, J.A., L. Chistoserdova, M.E. Lidstrom, and R.K. Thauer. 1998. The NADP-dependent methylene tetrahydromethanopterin dehydrogenase in *Methylobacterium extorquens* AM1. J. Bacteriol. 180:5351-5356.

Vorholt, J.A., C.J. Marx, M.E. Lidstrom, and R.K. Thauer. 2000. Novel formaldehyde-activating enzyme in *Methylobacterium extorquens* AM1 required for growth on methanol. J. Bacteriol. 182:6645-6650.

Figure legends

Fig. 1. Alignment of FOX islands in major scaffolds with the FOX island in *M. extorquens*. Open reading frames shown in dashed lines are implied to be present in gaps remaining between the contigs.

Fig. 2. A consensus phylogenetic tree showing positions of two major phylotypes present in the Sargasso Sea metagenome in relation to representatives of Proteobacteria, Planctomycetes and Archaea.

Table 1. H_4MPT-linked FOX genes in Sargasso Sea metagenome

Protein Query		Total	Clustered with FOX genes	Singletons	Clustered with non-FOX genes	Identity range, % (Amino acid)	Unique Phylotypes
Fae		24	8	7	9	28-100	19
	Fae	11	8	3	0	54-97	9
	Fae3	12	0	4	8	65-100	8
	Fae4	1	0	0	1	NA	1
Mtd		12	8	4	0	22-97	11
	MtdA	2	1	1	0	40	2
	MtdB	10	7	3	0	35-97	8
Mch		12	6	6	0	28-99	11
FhcA		22	9	12	1	37-99	21
FhcB		15	9	6	0	20-100	11
FhcC		11	5	5	1	24-97	10
FhcD		13	7	5	1	37-97	13
MptG		14	7	6	1	22-97	13
Orf5		7	5	2	0	26-70	5
Orf7		12	7	3	2	23-83	12

Orf9	10	3	7	0	26-93	10
Orf17	5	5	0	0	34-96	4
Orf19	11	7	4	0	25-71	11
Orf20	15	6	8	1	27-95	14
Orf21	2	1	1	0	30	2
Orf22	11	5	5	1	21-100	10
OrfY	6	3	3	0	24-31	6

Table 2. Divergence of *mch* genes the Sargasso Sea metagenome, based on protein-protein alignments

	Identity (%)				
	Within group	With other SSM groups	Proteo-bacteria	Plancto-mycetes	Archaea
Group 1 (5 sequences)	71-99	27-66	47-59	35-50	31-47
Group 2 (2 sequences)	83	37-63	51-61	38-45	35-40
Group 3 (2 sequences)	81	33-40	33-39	35-46	29-36
Group 4 (1 sequence)	NA	40-66	50-53	45-47	35-44
Group 5 (1 sequence)	NA	33-56	48-50	38-39	38
Group 6 (1 sequence)	NA	28-40	28-37	33-39	34-43

Table 3. Major FOX gene containing scaffolds

SCF	Size (kb)	FOX genes present*
2223320	70,015	*orf20-orf19-orf22-pabB-orf9-orf1-orf17-fae-orf7-orf5-mch-(orfY)-mtdB-mptG-fhcBADC*
2163205	15,399	*orf20-orf19-orf22*
2223402	15,297	*orf19-orf22*
2162480	13,629	*fae-mtdA*
2229052	10,815	*orf20-orf19-orf22-(pabB-orf9-orf1)-orf17-fae-orf7-orf5-mch-orfY-mtdB-mptG-fhcB*
2208802	8,110	*orf20-(orf19-orf22)-pabB-orf9-orf1-orf17-fae*
2178075	6,548	*mtdB-fhcBADC*
2207736	4,884	*fhcB(A)DC*
2214442	4,448	*fhcBA*
2162604	4,391	*mptG*
2211601	4,115	*mptG-fhcBA*
2167174	3,768	*orfY-mch-mptG*

*Genes in parentheses were implied to be present in the gaps between the contigs, from alignments with other scaffolds, see Fig. 1.

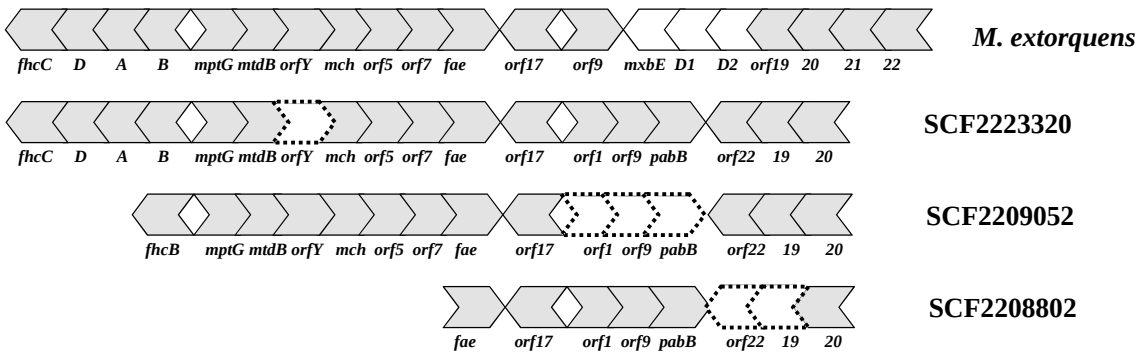
M. extorquens
fhcC D A B mptG mtdB orfY mch orf5 orf7 fae orf17 orf9 mxbE D1 D2 orf19 20 21 22
SCF2223320
fhcC D A B mptG mtdB orfY mch orf5 orf7 fae orf17 orf1 orf9 pabB orf22 19 20
SCF2209052
fhcB mptG mtdB orfY mch orf5 orf7 fae orf17 orf1 orf9 pabB orf22 19 20
SCF2208802
fae orf17 orf1 orf9 pabB orf22 19 20

Fig.1

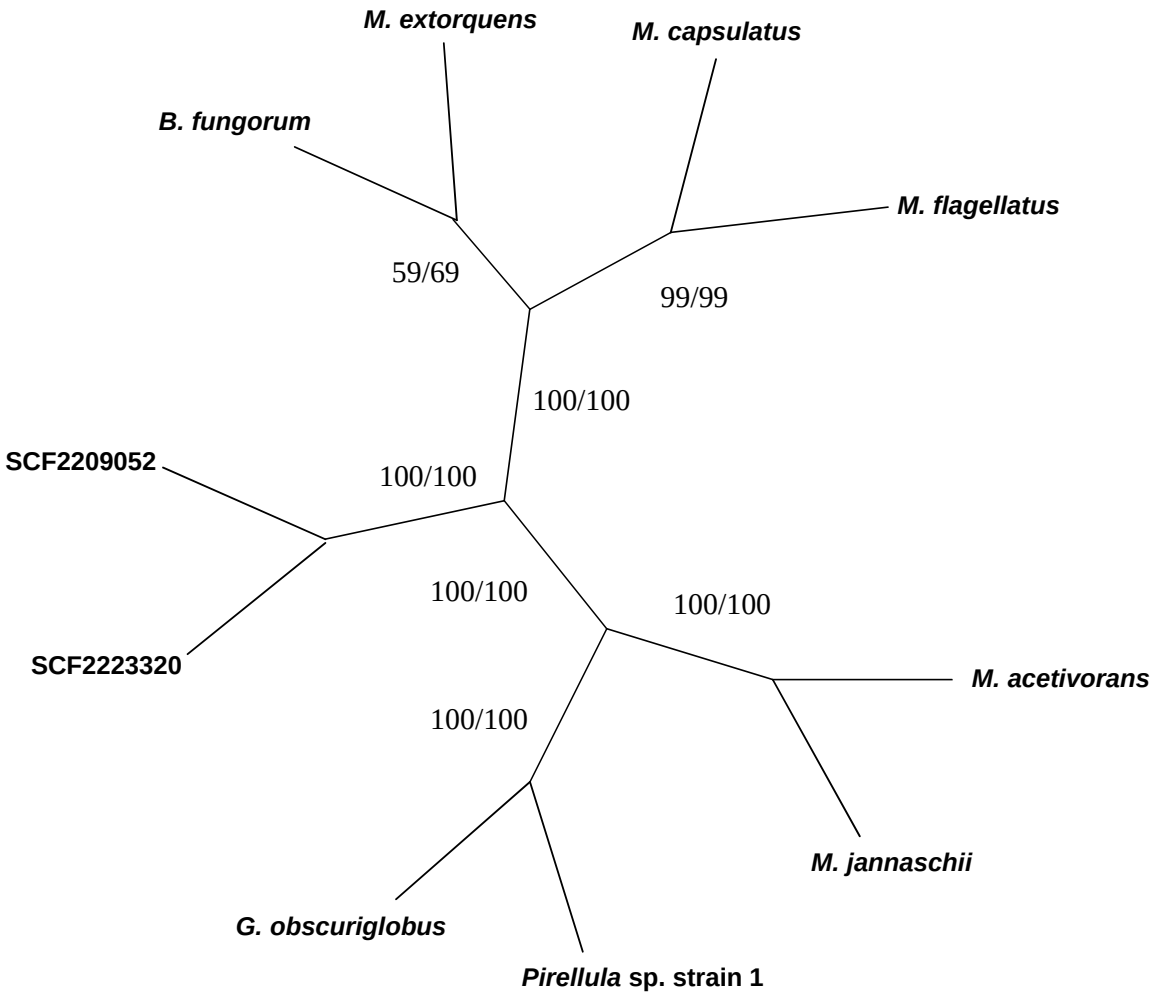
M. extorquens
M. capsulatus
B. fungorum
M. flagellatus
59/69
99/99
100/100
SCF2209052
100/100
100/100
100/100
SCF2223320
M. acetivorans
100/100
M. jannaschii
G. obscuriglobus
Pirellula sp. strain 1

Fig.2