High-resolution metagenomics targets major functional types in complex microbial communities

Marina G. Kalyuzhnaya[1], Alla Lapidus[3], Natalia Ivanova[3], Alex C. Copeland[3], Alice C. McHardy[4#], Ernest Szeto[5], Asaf Salamov[3], Igor V. Grigoriev[3], Dominic Suciu[6], Samuel R. Levine[2], Victor M. Markowitz[5], Isidore Rigoutsos[4], Susannah G. Tringe[3], David C. Bruce[7], Paul M. Richardson[3], Mary. E. Lidstrom[1,2] & Ludmila Chistoserdova[2]

Departments of [1]Microbiology and [2]Chemical Engineering, University of Washington, Seattle, WA 98195; [3]Production Genomics Facility, DOE Joint Genome Institute, 2800 Mitchell Drive, Bldg 400, Walnut Creek, CA 94596; [4]Bioinformatics and Pattern Discovery Group, IBM Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598; [5]Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mail Stop 50A-1148, Berkeley CA 94720; [6]Combimatrix Corporation, 6500 Harbour Heights Pkwy, Mukilteo, WA 98275; [7]DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, NM.

#Current address: Computational Genomics and Epidemiology Group, Max-Planck Institute for Computer Science, Campus E1 4, 66123 Saarbruecken, Germany

Most microbes in the biosphere remain uncultured and unknown[1]. Whole genome shotgun (WGS) sequencing of environmental DNA (metagenomics) allows glimpses into genetic and metabolic potentials of natural microbial communities[2-4]. However, in communities of high complexity metagenomics fail to link specific microbes to specific ecological functions. To overcome this limitation, we selectively targeted populations involved in oxidizing single-carbon (C_1) compounds in Lake Washington (Seattle, USA) by labeling their DNA via stable isotope probing (SIP), followed by WGS sequencing. Metagenome analysis demonstrated specific sequence enrichments in response to different C_1 substrates, highlighting ecological roles of individual phylotypes. We further demonstrated the utility of our approach by extracting a nearly complete genome of a novel methylotroph *Methylotenera mobilis*, reconstructing its metabolism and conducting genome-wide analyses. This approach allowing high-resolution genomic analysis of ecologically relevant species has the potential to be applied to a wide variety of ecosystems.

Methylotrophy, metabolism of organic compounds containing no carbon-carbon bonds (C_1 compounds), such as methane, methanol and methylated amines, is an important part of the global carbon cycle on Earth[5,6]. Identities of methylotrophs involved in utilization of specific C_1 substrates in a variety of environments have been previously assessed via both culture reliant[7] and culture independent methods[8]. The former provide important models for understanding the specific biochemical pathways enabling methylotrophy, and the latter provide insights into species richness within specific functional groups. However, while genomic data for some model methylotrophs are now available[9-11], these may not represent major players in specific functional guilds. At the same time, current methods for environmental detection provide little insight into the genomic structure of uncultivated methylotrophs.

Metagenomics or environmental genomics has recently become a powerful tool for collecting information on microbial communities, bypassing cultivation of individual species[1-4]. However, traditional metagenomic sequencing usually involves high cost and effort. Therefore only limited information can be gathered about highly complex natural communities, such as the ones inhabiting soils and lake sediments. As a proof of principle, we here utilized a strategy for targeting specific functional types within a community, via substrate-specific labeling of their DNA using Stable Isotope Probing (DNA-SIP) [12]. Focusing the sequencing effort on the labeled fraction of community DNA should result in higher sequence coverage for ecologically relevant species within a metagenome, directly linking them to an ecological function.

As our test model, we selected populations of microbes involved in methylotrophy in the sediment of a freshwater lake, Lake Washington in Seattle, WA, USA, an environment known for high rates of methane consumption[13]. Sediment samples from Lake Washington were exposed separately to ^{13}C-labeled methane, methanol, methylamine, formaldehyde, and formate, to target populations actively utilizing each of the C_1 compounds. Total DNA was extracted from each microcosm, and the ^{13}C-labeled fractions were separated from unlabeled DNA by isopycnic centrifugation (Fig. S1). ^{13}C-labeled DNA was used to construct five separate shotgun libraries, and these were sequenced at the JGI-PGF. 26 to 59 million base pairs (Mb) of sequence were produced from each microcosm, totaling 255 Mb. Sequences were assembled, automatically annotated, and loaded into the JGI's IMG/M system (Table 1 and Supplementary Methods), followed by manual analysis. Sequence coverage and degree of assembly depended on the sequencing effort applied and on the species richness and evenness of the enriched communities. Based on analysis of 16S rRNA gene sequences, community complexity was significantly reduced in microcosms exposed to each of the C_1 substrates compared to the complexity of the non-enriched community that we conservatively estimate to be over 5,000 species (Fig. 1, Table S1 and Supplementary Methods), and shifted toward specific functional guilds that included both *bona fide* methylotrophs (*Methylobacter tundripaludum, Methylomonas sp., Methylotenera mobilis, Methyloversatilis universalis, Ralstonia eutropha*) and organisms only distantly related to any cultivated species, implicating the latter in environmental cycling of C_1 compounds. The closest relatives of these included *Verrucomicrobia, Nitrospirae, Planctomycetes, Acidobacteria, Cyanobacteria,* and *Proteobacteria.* It is possible that some of these were not labeled by the primary substrate but by a labeled by-product, such as CO_2, as a result of cross-feeding. The 16S rRNA data were supported by data on phylogenetic profiling of each metagenomic dataset, based on top BLAST hit distribution patterns (not shown). From these analyses, the methylamine microcosm was one of the least complex in terms of species richness (Table S1) and most enriched in genes diagnostic for C_1 transforming capability (Table S2). It was dominated by a group of closely related strains identified as *M. mobilis*, represented by a novel obligate methylamine utilizer recently isolated from Lake Washington[14]. Based on 16S rRNA gene sequence coverage (up to 20X, Table S1), complete or nearly complete genomes of a few *M. mobilis* strains were predicted to be encoded in the methylamine microcosm metagenome. From traditional laboratory enrichments, *M. mobilis* does not appear to be a "weed" organism, as it is readily out-competed by other methylamine-utilizers (Table S3). However the incubation conditions used in this study must have favored *M. mobilis*, which appears to comprise less than 0.4% of the total bacterial population, based on random sequencing of amplified 16S rRNA genes[15]. A composite genome of *M. mobilis* totaling slightly over 11 Mb was extracted from the methylamine microcosm metagenome using the recently described compositional binning method, PhyloPythia[16] (genome statistics are shown in Tables 1 and S4). The quality of binning and the recovery of complete or almost complete genomes were validated by hybridizing DNA of a laboratory-cultivated *M. mobilis* to a custom DNA micoarray based on this composite genome (Supplementary Methods). We also validated genome completeness by examining the presence of various metabolic and housekeeping genes (Tables S5 and S6). In terms of central metabolism, we identified a complete set of genes for specific pathways enabling methylamine utilization in *M. mobilis*. Multiple copies for each gene were identified (3 to 15, Table S5), consistent with the composite genome being representative of a few closely related strains. In terms of the housekeeping functions, completeness of the genome was demonstrated by the presence of 181 tRNA genes corresponding to 36 tRNA acceptors for recognizing all 20 amino acids (not shown), and of a complete set of aminoacyl-tRNA transferases (Table S6). Standard sets of genes for DNA replication, transcription and translation were identified, and complete pathways were reconstructed for biosynthesis of all the amino acids and nucleotides and all the essential vitamins (Table S6).

We reconstructed the metabolism of *M. mobilis* and conducted genome-wide comparisons with the genome of *Methylobacillus flagellatus*, a methylotroph closely related to *M. mobilis*, of a similar genome size[11,14] (Figures 2 and S2 and Tables S4-6). *M. mobilis* from Lake Washington (Table S1) and *M. flagellatus* are 93 to 95% similar at the 16S rRNA gene sequence level and share most of the pathways enabling methylotrophy. However, they were found to be quite different in their genomic content, gene synteny and gene conservation. Reciprocal BLAST analyses revealed that only 57% of the proteins translated from the *M. flagellatus* chromosome had homologs in *M. mobilis* at a 50% cut-off, and only 62% of the proteins translated from the composite genome of *M. mobilis* had homologs in *M. flagellatus*. Focusing on some of the highly conserved genomic regions encoding methylotrophy functions, we uncovered examples of non-homologous replacements in common biochemical pathways as well as examples of homolog recruitment into novel/secondary functions. Two of the notable examples are illustrated in Fig. S3. (A) A gene for azurin, a specific electron acceptor from methylamine dehydrogenase (MADH) in *M. flagellatus*[11] is missing from the MADH gene cluster (and elsewhere in the composite genome) in *M. mobilis.* Instead, it is replaced by a cytochrome ($c_{551/552}$) gene that is so far unique to *M. mobilis*, demonstrating that in two closely related organisms, different strategies are employed for one of the key energy-generating pathways. (B) From a highly conserved gene cluster encoding reactions of tetrahydromethanopterin (H_4MPT)-linked formaldehyde oxidation, the *fae* gene is missing in *M. mobilis*. In its place, two novel genes are present, encoding a sensor histidine kinase and a response regulator. A homolog closely related to *fae* from *M. flagellatus* (85% amino acid identity) is instead part of a gene cluster predicted to be involved in chemotaxis, while a second, less similar homolog (60% amino acid identity), representing a novel phylogenetic subtype of *fae* (Fig. 3), is part of a predicted regulatory gene cluster. This conspicuous gene clustering suggests that Fae (Formaldehyde Activating Enzyme), whose known enzymatic function is to bind formaldehyde and convert it into methylene-H_4MPT[17], must have a second function, possibly as a sensor component of regulatory and/or signal transduction systems. This hypothesis is supported by experimental removal of the 'chemotaxis' gene cluster shown in Fig. 3, which did not affect chemotaxis of *M. mobilis* toward methylamine (not shown). Presence of *fae* homologs in genomes that do not encode H_4MPT-dependent C_1 transfer functions also supports this hypothesis (Supplementary Methods).

Global genome-genome comparisons between *M. mobilis* and *M. flagellatus* revealed that the conserved parts of the genomes encode central metabolism and housekeeping functions (methylotrophy, energy transduction, replication, transcription, translation, amino acid and vitamin biosynthesis), while the variable parts of the genomes encode auxiliary functions (transport, regulation, electron transfer, CRISPR, prophage, non-essential biochemical pathways). We were able to precisely map 63 indels of more than two genes on the chromosome of *M. flagellatus*, totaling approximately 1,070 kb, not present in the composite genome of *M. mobilis* (Table S7). The number and the size of indels could not be estimated with such precision for *M. mobilis* because of the composite nature of its genome (Table S8), but we were able to calculate that approximately 600 kb of sequences per genome were unique, when the genome size was estimated at approximately 2.5 Mb. One notable element missing from the composite genome of *M. mobilis* was the methanol dehydrogenase-encoding gene cluster thought to be highly conserved in most methylotrophs[7]. Conversely, some enzymes and pathways not present in *M. flagellatus* were identified in *M. mobilis,* such as the methylcitric acid cycle (Fig. S4). Comparisons of energy-generating electron transfer pathways encoded in the two genomes showed little overlap (Table S9), suggesting adaptation to significantly different life styles. For example, the presence of genes for the denitrification pathway suggested a propensity for *M. mobilis* to thrive in microaerobic environments, which was subsequently proven in experiments with cultivated *M. mobilis* (not shown), while *M. flagellatus* is known to be a strict aerobe[11]. The predicted denitrification capability of *M. mobilis* also suggests that C_1 and nitrogen cycling in Lake Washington sediment may be significantly interlinked.

Sequences of *M. mobilis* were also present in the metagenomes of microcosms incubated with methane, methanol, and formaldehyde (Tables S1 and S10), possibly as a result of cross-feeding on labeled formaldehyde that is an intermediate in the oxidation of methane and methanol. To test whether *M. mobilis* strains labeled by these substrates were metabolically different from *M. mobilis* strains in the methylamine microcosm, we conducted substrate-specific genome-genome comparisons, interrogating each dataset separately and all three datasets at once (to increase sequence coverage) with the *M.*

mobilis composite genome. In this way we detected a number of genes that were not present in the combined dataset for methane, methanol and formaldehyde microcosms, but were unique to the methylamine microcosm. Remarkably, the entire gene cluster encoding methylamine oxidation (*mauFBEAGLMNO*) was missing from the former, suggesting that methylamine oxidizing capability is “an acquired taste” and not an attribute of *M. mobilis* as a species and suggesting alternative primary substrates for some *M. mobilis* strains. In contrast, hits were found for the entire set of genes involved in the methylcitric acid cycle, pointing to its potential role as a central metabolic pathway (not shown). One proposed function for this cycle could be in utilizing propionate that is a product of de-methylation of a compound(s) typical of aquatic environments (such as dimethylsulfoniopropionate[18]). Another gene with a predicted function that was unique to the methylamine microcosm *M. mobilis* was the novel, divergent *fae* (Fig. 3 and S3), suggesting this novel *fae* and the surrounding genes may have a specialized function in the metabolism of methylamine. Conversely, specific metabolic traits were detected in methane and methanol microcosm *M. mobilis* that were not present in the methylamine microcosm strains. A remarkable feature of *M. mobilis* from the methanol microcosm was the presence of RuBisCO genes, suggesting these strains may be capable of autotrophy. *M. mobilis* from the methane microcosm featured nitrogenase genes, suggesting that some *M. mobilis* strains may be active in nitrogen fixation. We recently isolated a number of *M. mobilis* strains that reveal nutritional properties matching those predicted from the metagenomes. We are planning to completely sequence the genomes of three of these strains and compare them to each other and to the metagenome.

In addition to the *M. mobilis* composite genome, highly covered bacteriophage genomes were recovered from the methylamine sample. One of these (37 kb) was homologous to the genome of the *Bordetella* phage BPP1[19] (not shown), while others (approximately 10 kb) were distantly related to the genome of a marine bacteriophage PM2, the only member of the *Corticoviridae* family[20], and to a prophage found in the genome of *M. flagellatus*[11] (Fig. S5). Two of the contigs of the latter type were found to contain overlapping sequences at the ends. These were trimmed and joined at the ends to produce circular phage chromosome sequences. The presence of phage chromosomes in the methylamine microcosm metagenome indicates that free-living phages were propagating during the microcosm incubation with ^{13}C-methylamine. *M. mobilis* is the most likely host for these phages, due to its dominance in the labeled microcosm community. However, the phage sequences were missing from the methane, methanol and formaldehyde microcosms, indicating a specific association between phage and methylamine-utilizing *M. mobilis*. This was supported by the presence of a conspicuous gene cluster, also unique to the methylamine-utilizing *M. mobilis,* which encodes pilus assembly and secretion functions (*cpaABCEFtadBC*; Fig. S6). This pilus is a possible candidate for a specific phage receptor. In addition to these, a number of other candidate phage receptors were unique to the methylamine *M. mobilis* (a biopolymer transporter, a major facilitator). The connection between methylamine metabolism and phage association is very intriguing. One hypothetical scenario could be imagined in which a specific transporter for methylamine also serves as a specific phage receptor. However, this hypothesis will need to be tested experimentally.

We were also able to analyze other, less covered genomes by supplementing the PhyloPythia binning with protein recruitment using related genome sequences as a reference[4]. From comparisons with the *Methylococcus capsulatus* genome[9], we estimated that a large portion of the (composite) genome of *M. tundripaludum* was present in the methane microcosm dataset (not shown). We conducted metabolic reconstruction for this organism (Fig. S7) and mapped indels on the chromosome of *M. capsulatus* (not shown). Trends similar to the ones noted for gene conservation between *M. mobilis* and *M. flagellatus* were observed: the core parts of the genomes of *M. tundripaludum* and *M. capsulatus*, encoding central metabolism and house keeping genes, were conserved, while parts of the genomes encoding auxiliary functions were not. Notable omissions from the *M. tundripalidum* genome were gene clusters encoding the soluble methane monooxygenase, RuBisCO, and enzymes of the serine cycle. These genomic features agree with physiological analysis of the cultivated *M. tundripaludum* strain[21]. In a similar fashion, a large portion of a *R. eutropha* genome was recovered from the formate microcosm metagenome. It was highly similar to the published genome of Strain H-16[22], encoding all the core functions and only missing genes for a few auxiliary pathways, such as CO dehydrogenase and polysaccharide biosynthesis. It also appeared to lack the megaplasmid found in Strain H-16 (data not shown). Partial genomes were obtained for uncultivated representatives of *Burkholderiaceae, Comamonadaceae, Rhodocyclaceae*, and *Actinobacteria*, the groups that include methylotrophic representatives (Table S11).

Besides the *bona fide* methylotrophs, our functional enrichment approach suggested that phyla not traditionally classified as methylotrophs may be involved in C_1 transformations, such as *Verrucomicrobia, Nitrospirae* and *Planctomycetes.* The lower coverage of these strains may reflect either slower rates of metabolism or sub-optimal incubation conditions. Acidophilic methane-oxidizing *Verrucomicrobia* have been described recently[23-25]. However, based on 16S rRNA and functional gene comparisons, *Verrucomicrobia* uncovered in this study are only distantly related to these organisms (<90% 16S rRNA identity). We analyzed the datasets containing *Verrucomicrobia* phylogenetic markers (methane, methanol and formaldehyde microcosms) for the presence of specific functional genes potentially enabling methylotrophy in these species and identified a conspicuous gene (*mtaB*) that was present in one or more copies in each dataset, predicted to encode a methanol:corrinoid methyltransferase. This enzyme has been characterized in methylotrophic archaea[26] and suggested to be involved in methanol utilization by *Clostridia*[27]. However, MtaB sequences from the Lake Washington metagenome were most similar to a homolog from the only publicly available *Verrucomicrobia* genome, that of *Opitutaceae* bacterium TAV2 (Table S12), thus implicating this bacterium as well as the *Verrucomicrobia* detected in this study in methanol utilization. While no methylotrophic *Planctomycetes* or *Nitrospirae* have been obtained at the moment in pure cultures, these organisms are often detected in environments with high rates of C_1 metabolism[28].

This work is a proof of principle study demonstrating that the metagenomics approach can enable detailed analysis of the genomes of environmentally relevant microbes, by-passing pure culture isolation, even if the species in question comprise a minor fraction in a highly complex microbial community. A specific enrichment step, such as SIP employed here, is key to increasing the resolution of metagenomics, by focusing the sequencing effort on specific functional types. We have presented here a detailed analysis of the genome of a novel methylotroph, *M. mobilis* that comprises less than 0.4% of the total bacterial population in Lake Washington sediment. We also demonstrated the utility of SIP-enabled metagenomics in uncovering specific bacterium-phage relationships, suggesting the existence of complex population dynamics involving multiple strains of *M. mobilis* and multiple strains of novel corticoviruses. The existence of such dynamics, likely involving competition for a nutrient (methylamine or a different methylated compound), in turn highlights the potential environmental importance of C_1 compounds as components of global carbon cycling. A genome of an uncultivated *M. tundripaludum* was also analyzed in detail, expanding the current genomic knowledge of methane utilizers. In addition, we identified *Verrucomicrobia* only distantly related to recently described methanotrophic isolates, suggesting that methylotrophy may be a common attribute of this phylum. Overall, this study uncovered the existence of dynamic and diverse populations responding to C_1 substrates, pointing toward the existence of a complex, multi-tired microbial food web involved in environmental C_1 cycling in Lake Washington sediment and likely in other freshwater lake sediments.

In conclusion, the variation on the standard metagenomics approach described here, employing function-specific enrichment, allows high-resolution genomic analysis of major functional types and has the potential to be used in a wide variety of ecosystems with a wide variety of labeled substrates, as well as in combination with other types of enrichment.

Methods

Sample collection, stable isotope probing and DNA extraction. Sediment samples were collected on May 15, 2005, from a 63 m deep station in Lake Washington, Seattle, Washington (47^{0}38.075’ N, 122^{0}15.993’ W) using a box core that allowed collection of undisturbed sediment. Samples were transported to the laboratory on ice and immediately used to set up microcosms. Each microcosms contained 10 ml sediment from the oxygenated top 1 cm layer, 90 ml Lake Washington water, and one of the following ^{13}C substrates: methane (50% of air), methanol (10 mM), methylamine (10 mM), formaldehyde (1 mM), or formate (10 mM). All substrates were 99 atom % ^{13}C and were purchased from Sigma-Aldrich, with the exception of [^{13}C] methanol, which was provided by the National Stable Isotope Resource at Los Alamos National Laboratory. The samples were incubated for 3-5 (methylamine and methane), 5-7 (methanol) or 10-14 (formaldehyde and formate) days at room temperature, with shaking. It has been previously demonstrated that SIP incubations at the in situ temperature (8^{o}C) resulted in similar community structures while longer incubation times were required[29]. DNA was extracted and purified and subjected to density gradient ultracentrifugation as previously described[12, 29], with slight modifications (Supplementary Methods). ^{13}C-DNA fractions were visualized in UV (Fig. S1) and collected using 19-gauge needles.

DNA sequencing and assembly. Five shotgun libraries were constructed, one from each microcosm, in the pUC18 vector (1-3 kb inserts). The libraries were sequenced with BigDye Terminators v3.1 and resolved with ABI PRISM 3730 (ABI) sequencers. A total of 344,832 reads comprising 255.08 megabases (Mb) of Phred Q20 sequence were generated. Sequences were screened for vector contaminations and quality trimmed using LUCY[30] and assembled, both *en masse* and by sample, using the PGA assembler.

Assembly statistics are shown in Table 1. These draft quality assemblies were manually validated and used for all downstream analyses.

Compositional binning. Assembled metagenomic fragments were binned (classified) using PhyloPythia, a phylogenetic classifier that uses a multi-class Support Vector machine (SVM) for the composition-based assignment of fragments at different taxonomic ranks, essentially as previously described[16]. Generic models for the ranks of domain, phylum and class were combined with models for the dominating clades in the sample. The generic models represent all clades covered by three or more species at the corresponding ranks among the sequenced microbial isolates. At the rank of family, a sample-specific model was created with classes for the clades *Methylococcaceae, Burkholderiaceae, Rhodocyclaceae, Methylophilaceae* and *Comamonadaceae* and a class 'other'. A sample-specific model for the dominant sample populations was created with classes for the *Methylotenera* and *Methylobacter* populations and a class 'other'. The sample-specific population model was trained on 138 kb and 141 kb of contigs for the *Methylotenera* and *Methylobacter* populations, respectively, that were identified based on phylogenetic marker genes, as well as sequenced isolates for the class 'other'. The family-level model was trained using the sample-specific data and additional sequenced isolates available for the corresponding clades. For each model, five sample-specific multi-class SVMs were created using fragments of lengths of 3, 5, 10, 15 and 50 kb, respectively. All input sequences were extended by their reverse complement prior to computation of the compositional feature vectors. The parameters w and l were both set to 5 for the sample-specific models. The final classifier consisting of the sample-specific and generic clade models was applied to assign all fragments >1 kb of the samples. In case of conflicting assignments, preference was given to assignments of the sample-specific models. Data were incorporated into the Integrated Microbial Genomes with Microbiome Samples (IMG/M) system (http://img.jgi.doe.gov/m). This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under accession number XXX.

Species richness estimation.16S rRNA gene fragments were amplified from sediment DNA using the EUB27f/1496R primer set following by cloning into the pCR2.1 vector (Invitrogen), as recommended by the manufacturer. Inserts of 859 randomly selected clones were subjected to restriction fragment length polymorphism (RFLP) analysis, after digestion with AluI (Fermentas). The GeneTools imaging software (ProcessGelFiles4.m) was used to compare the restriction patterns. AluI restriction fragments resulting from pCR2.1 were used as internal locators to adjust the positions of the insert fragments. Different restriction patterns were clustered by likeness using agglomerative clustering (Matlab, Mathworks). Clones predicted to be identical by these analyses were sequenced in order to verify the efficiency of the analysis, and in each case the identity was proven. Nine groups were identified containing two identical sequences, three groups containing three identical sequences, and one group containing four identical sequences. Chao nonparametric richness estimators were implemented to estimate species richness using the computational tool EstimateS (version 8, http://purl.oclc.org/estimates), resulting in the lowest richness estimate of 5,430.

Protein recruitment. Protein recruitment was carried out essentially as previously described[4] except for protein sequences rather than DNA sequences were used. The Phylogenetic Profiler tool that is part of the IMG/M package was used. In the case of *M. tundripaludum/ M. capsulatus* pair, cut-offs of 60% to 80% were used, based on 89% 16S rRNA gene similarity between the two strains. In the case of *R. eutropha/ R. eutropha* H-16 pair (99% 16S rRNA gene similarity), a cut-off of 90% was used.

References

1. The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. The National Academies Press (2007).

2. Tyson, G.W. et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**:37-43 (2004).

3. Tringe, S.G. et al. Comparative metagenomics of microbial communities. *Science* **308**:554-557 (2005).

4. Rusch, D.B. et al. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol*. **5**:e77 (2007).

5. Hanson, R.S. & Hanson, T.E. Methanotrophic bacteria. *Microbiol Rev*. **60**:439-471 (1996).

6. Guenter, A. The contribution of reactive carbon emissions from vegetation to the carbon balance of terrestrial ecosystems. *Chemosphere* **49**:837-844 (2002).

7. Lidstrom, M.E. Aerobic methylotrophic procaryotes. In: A. Balows, H.G. Truper, M. Dworkin, W. Harder and K.-H. Schleifer (ed.) The Prokaryotes. Springer Verlag, New York, NY (2006).

8. McDonald, I.R., L. Bodrossy, Y. Chen & Murrell, J.C. Molecular ecology techniques for the study of aerobic methanotrophs. *Appl Environ Microbiol*. **74**:1305-1315 (2008).

9. Ward, N. et al. Genomic insights into methanotrophy: the complete genome sequence of *Methylococcus capsulatus* (Bath). *PLoS Biol*. **2**:e303 (2004).

10. Kane, S.R. et al. Whole-Genome Analysis of Methyl tert-Butyl Ether (MTBE)-Degrading Beta-Proteobacterium *Methylibium petroleiphilum* PM1. *J Bacteriol*. **189**: 1931-1945 (2007).

11. Chistoserdova, L. et al. The genome of *Methylobacillus flagellatus*, the molecular basis for obligate methylotrophy, and the polyphyletic origin of methylotrophy. *J. Bacteriol.* **189**:4020-4027 (2007).

12. Radajewski, S., Ineson, P., Parekh, N.R. & Murrell, J.C. Stable-isotope probing as a tool in microbial ecology. *Nature* **403**:646-649 (2000).

13. Auman, A.J., Stolyar, S., Costello, A.M. & Lidstrom, M.E. Molecular characterization of methanotrophic isolates from freshwater lake sediment. *Appl Environ Microbiol*. **66**:5259-66 (2000).

14. Kalyuzhnaya, M.G., Bowerman, S., Lara, J.C., Lidstrom, M.E. & Chistoserdova, L. *Methylotenera mobilis* gen. nov., sp. nov, an obligately methylamine-utilizing bacterium within the family *Methylophilaceae*. *Int J Syst Evol Microbiol.* **56**:2819-2823 (2006).

15. Kalyuzhnya, M.G., Lidstrom, M.E. & Chistosedova, L. Real-time detection of actively metabolizing microbes via redox sensing as applied to methylotroph populations in Lake Washington. ISME J. (2008, In Press).

16. McHardy, A.C., Garcia Martin H., Tsirigos, A. Hugenholtz P. & Rigoutsos I. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* **4**:63-72 (2007).

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

18. Ginsburg, B. et al. DMS formation by dimethylsulfoniopropionate route in freshwater. *Environ Sci Technol*. **32**:2130-2136 (1998).

19. Liu, M. et al. Genomic and genetic analysis of *Bordetella* bacteriophages encoding

reverse transcriptase-mediated tropism-switching cassettes. *J Bacteriol*. **186**:1503-1517 (2004).

20. Krupovic M. et al. Genome characterization of lipid-containing marine bacteriophage PM2 by transposon insertion mutagenesis. *J Virol*. **80**:9270-9278 (2006).

21. Wartiainen, I., Hestnes, A.G., McDonald, I.R. & Svening, M.M. *Methylobacter tundripaludum* sp. nov., a methane-oxidizing bacterium from Arctic wetland soil on the Svalbard islands, Norway (78° N). *Int J Syst Evol Microbiol*. **56**:109-113 (2006).

22. Pohlmann, A. et al. Genome sequence of the bioplastic-producing "Knallgas" bacterium *Ralstonia eutopha* H16. *Nat Biotechnol*. **24**:1257-1262 (2006).

23. Islam, T., Jensen, S., Reigstad, L.J., Larsen, & Ø, Birkeland, N.-K. Methane oxidation at 55°C and pH 2 by a thermoacidophilic bacterium belonging to the Verrucomicrobia phylum. *PNAS* **105**:300-304 (2008).

24. Dunfield, P.F. et al. Methane oxidation by an extremely acidophilic bacterium of the phylum *Verrucomicrobia*. *Nature* **450**: 879-82 (2007).

25. Pol, A. et al. Methanotrophy below pH 1 by a new *Verrucomicrobia* species. *Nature* **450**:874-878 (2007).

26. Sauer, K. and Thauer, R.K. Methanol:coenzyme M methyltransferase from *Methanosarcina barkeri*. Zinc dependence and thermodynamics of the methanol:cob(I)alamin methyltransferase reaction. *Eur J Biochem.* **249**:280-285 (1997).

27. Das, A. et al. Characterization of a corrinoid protein involved in the C1 metabolism of strict anaerobic bacterium *Moorella thermoacetica*. *Protteins: Struct Funct Bioinform.* **67**:167-176 (2007).

28. Lösekann, T. et al. Diversity and abundance of aerobic and anaerobic methane oxidizers at the Haakon Mosby mud volcano, Barents Sea. *Appl Environ Microbiol*. **73**:3348-3362 (2007).

29. Nercessian, O., Noyes, E., Kalyuzhnaya, M.G., Lidstrom, M.E. & Chistoserdova, L. Bacterial populations active in metabolism of C_1 compounds in the sediment of Lake Washington, a freshwater lake. *Appl Environ Microbiol*. **71**:6885-6899 (2005).

30. Chou, H.H. & Holmes, M.H. DNA sequence quality trimming and vector removal. *Bioinformatics* **17**:1093-1104 (2001).

Acknowledgements This research was supported by the National Science Foundation as part of the Microbial Observatories program (MCB-0604269). This work was performed, in part, under the auspices of the U.S. Department of Energy's Office of Science Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396. The sequencing for the project was provided through the US Department of Energy (DOE) Community Sequencing Program (http://www.jgi.doe.gov/CSP/index.html).

Author Contributions

M.G.K., M.E.L. and L.C. conceived the project. M.E.L. and L.C. coordinated project execution. M.G.K. collected samples, performed SIP, purified DNA for sequencing and performed microarray hybridizations. D.B. and P.M.R. oversaw library construction and sequencing. S.G.T. oversaw sequence assembly and analysis. A.C.C. and A.L. carried out assemblies. A.S. and I.V.G. conducted gene prediction and annotation. A.C.M. and I.R. carried out binning. E.S. and V.M.M. carried out data processing and loading into IMG/M. L.C. and N.I. carried out metabolic reconstruction. S.R.L. and M.G.K. performed species richness estimates. D.S. carried our microarray design. M.G.K., M.E.L. and L.C. wrote the initial draft of the paper, all other authors contributed.

Table 1. Summary sequencing and assembly and gene prediction statistics

	Methane	Methanol	Methylamine	Formaldehyde	Formate	Combined	*Methylotenera*
Assembly statistics							
Number of reads	71,808	67,200	83,712	80,640	41,472	344,832	NA
Average read length (bp)	792	797	709	712	638	741	NA
Trimmed read length (Mbp)	56.85	53.53	59.34	58.91	26.45	255.08	NA
Non-redundant sequence (bp)	52.16	50.25	37.23	57.62	17.57	211.47	11.16
Percent of reads in contigs	10.2	10.0	55.5	7.3	34.3	27.6	100
Total contigs (>2 kb)	2,797	2,871	7,558	2,583	3,618	25,877	4,078
Total singlets	59,417	56,408	29,217	69,104	18,857	215,581	0
Average sequence coverage (x)	1.6	1.6	1.9	1.7	1.9	1.7	2.1
Highest sequence coverage (x)	7.0	4.8	20.4	6.4	4.7	23.1	20.4
Average size of contigs (bp)	1,418	1,288	2,065	1,166	1,265	1,593	2,736
Largest contig (bp)	6,174	5,913	20,771	4,714	6,276	22,407	15,820
GC content (%)	58.9	59.5	53.0	57.9	65.8	58.3	46.2
Gene predictions							
Protein coding genes	81,076	77,229	54,340	89,729	28,700	321,503	12,719
Genes in COGs	43,456	40,773	33,643	46,032	17,112	174,344	10,082
Genes in Pfams	28,090	26,494	23,586	29,375	10,585	115,228	8,543
Predicted enzymes	3,089	3,047	5,005	3,065	1,417	16,780	3,264
Number of 16S rRNA genes	12	12	10	18	5	61	3
Number of tRNA genes	405	412	376	504	121	1,728	181

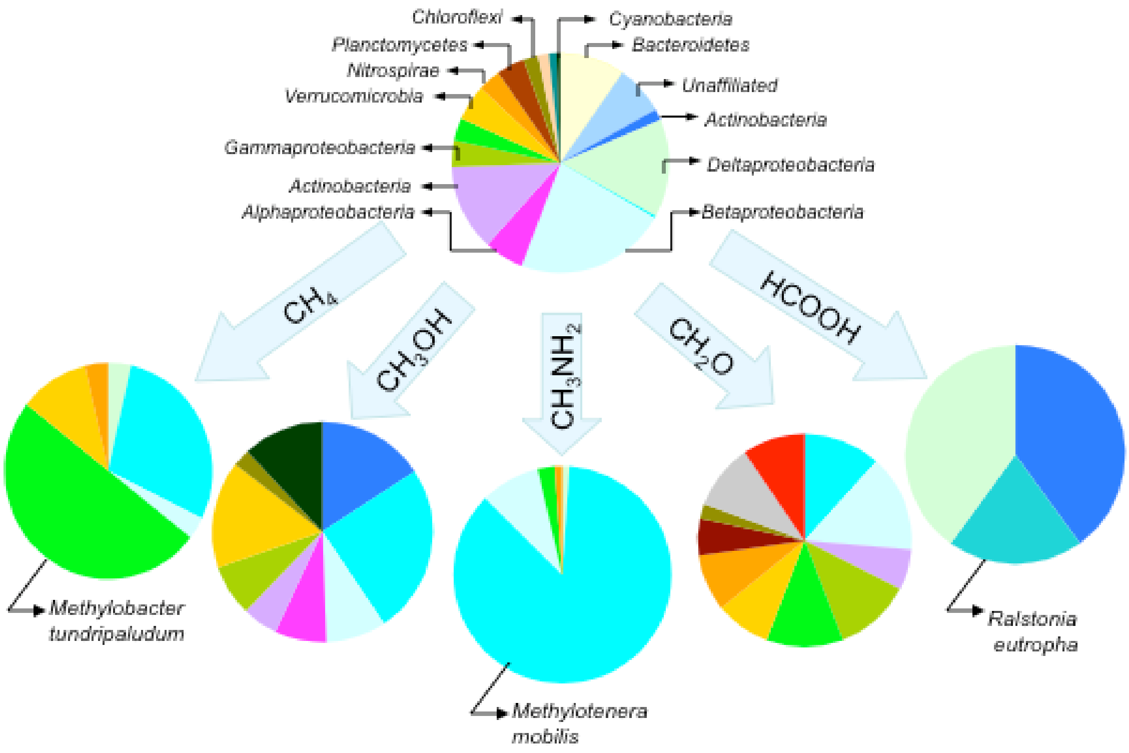
Figure legends

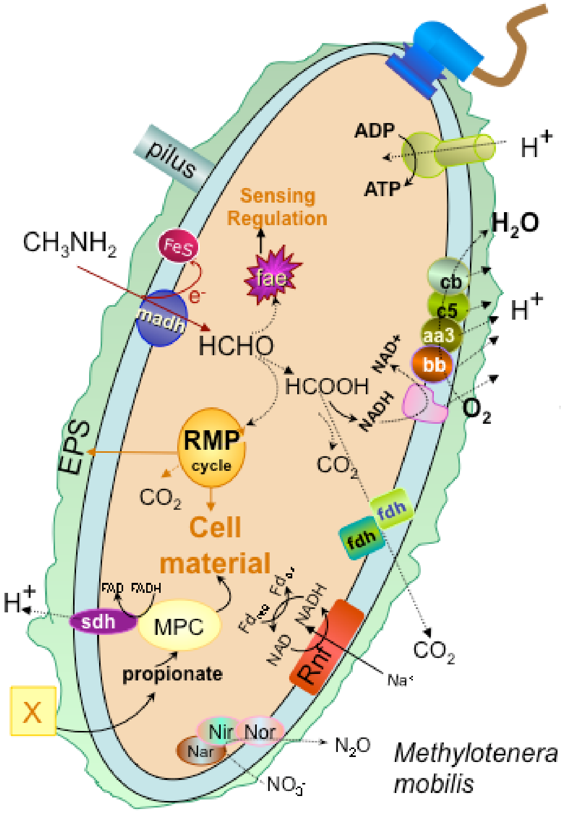
Figure 1. Taxonomic distribution of 16S rRNA gene sequences from metagenomes. The sum of coverage scores for each phylum (Table S1) were used for metagenomes generated in this work, and data from[5] were used for the non-enriched community. Similar taxonomic distributions were observed when PCR-amplified libraries generated for each microcosm were analyzed as in[5] (data not shown). ◆*Ralstonia eutropha*; ◆ *Methylotenera mobilis*; ◆ other *Betaproteobacteria*; ◆*Methylobacter tundripaludum*; ◆ other *Gammaproteobacteria*; ◆ *Alphaproteobacteria*; ◆ *Deltaproteobacteria*; ◆ *Actinobacteria*; ◆ *Acidobacteria*; ◆ Archaea; ◆ *Bacteroidetes*; ◆ *Chloroflexi*; ◆ *Cyanobacteria*; ◆ *Firmicutes*; ◆ *Gemmatimonadetes*; ◆ *Planctomycetes*; ◆ *Verrucomicrobia*; ◆ Unclassified bacteria; ◆ Chloroplasts.

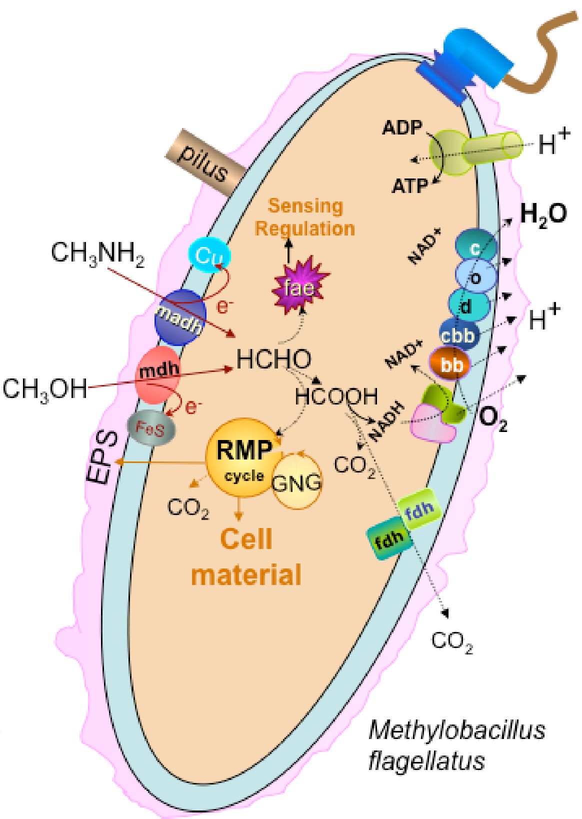
Figure 2. Metabolic features of *M. mobilis* compared to metabolic features of *M. flagellatus* as deduced from genomic comparisons. Major metabolic pathways and energy generating systems are shown. Similar shapes indicate similar functions, different colors indicate lack of homology at the protein level.

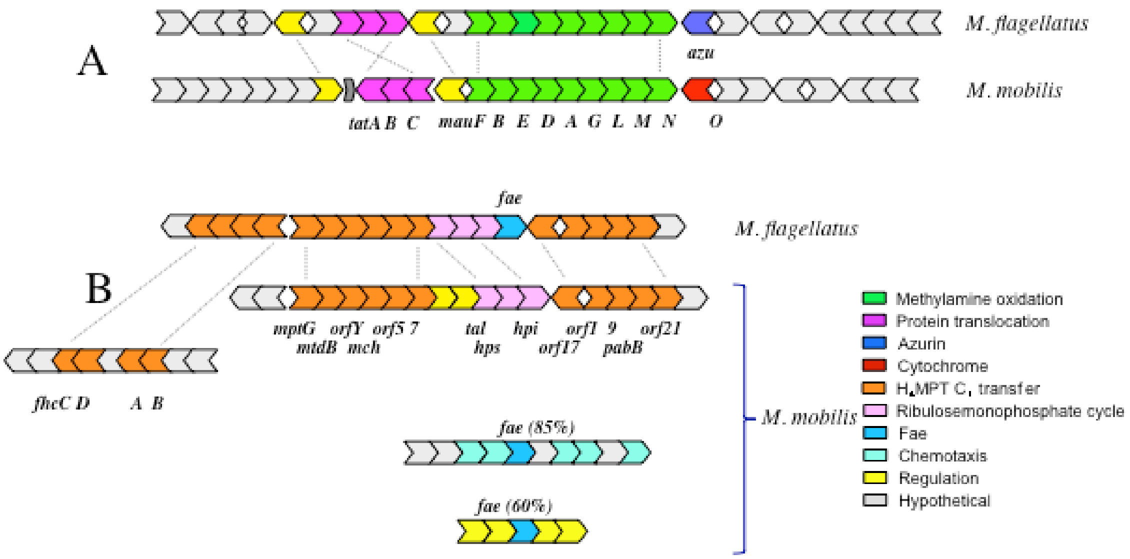
Figure 3. Comparison of gene clusters involved in methylotrophy in *M. mobilis* and *M. flagellatus*. A. In the methylamine oxidation gene cluster, the gene for azurin, an electron acceptor from methylamine dehydrogenase in *M. flagellatus* is replaced by a gene encoding cytochrome $C_{551/552}$ suggesting a functional replacement. B. Two *fae* genes in *M. mobilis* are parts of gene clusters predicted to be involved, respectively in sensing/

chemotaxis and regulation, suggesting novel/ secondary functions for Fae. The accuracy of assembly was tested by PCR amplifying portions of the clusters and re-sequencing.




pilus
ADP
H+
ATP
Sensing
Regulation
CH_3NH_2
FeS
e-
madh
fae
HCHO
cb
c5
aa3
bb
H_2O
H+
NAD+
HCOOH
NADH
O_2
EPS
RMP
cycle
CO_2
CO_2
Cell
material
fdh
fdh
H+
FAD
FADH
sdh
MPC
propionate
Fd
Fd
NADH
NAD
Rnf
Na+
CO_2
X
Nir
Nor
Nar
N_2O
NO_3^-
Methylotenera
mobilis


pilus
ADP
H+
ATP
Sensing
Regulation
H_2O
CH_3NH_2
Cu
e-
fae
NAD+
c
o
d
madh
HCHO
cbb
H+
CH_3OH
mdh
NAD+
bb
HCOOH
NADH
O_2
e-
FeS
EPS
RMP
cycle
GNG
CO_2
CO_2
Cell
material
fdh
fdh
CO_2
Methylobacillus
flagellatus


A
M. flagellatus
azu
M. mobilis
tatA B C
mauF B E D A G L M N
O
B
fae
M. flagellatus
mptG orfY orf5 7
mtdB mch
tal
hps
hpi
orf1 9
orf17
pabB
orf21
fhcC D A B
M. mobilis
fae (85%)
fae (60%)
Methylamine oxidation
Protein translocation
Azurin
Cytochrome
H$_4$MPT C$_1$ transfer
Ribulosemonophosphate cycle
Fae
Chemotaxis
Regulation
Hypothetical