Draft Genome Sequences of Freshwater *Methylophilaceae*

# Authors

Joshua J. Hamilton1, Sarah L. R. Stevens1, JGI Personnel2, Katherine D. McMahon1,3\*

# Institutional Affiliation

1 Department of Bacteriology, University of Wisconsin-Madison, Madison, WI

2 Department of Energy Joint Genome Institute, Walnut Creek, CA

3 Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI

# Corresponding author

Corresponding author: Katherine D. McMahon, trina.mcmahon@wisc.edu

# Abstract (Heading 1)

Authors should provide a concise, non-redundant and meaningful abstract that describes the nature of the article. It should summarize the rationale, the objectives and the findings of the report and provide key details (e.g., relevant INSDC identifiers, culture collection identifiers, other project metadata that is accessible in standardized form).

# Keywords: (Heading 1)

Authors should include five to seven descriptive keywords. These may include the article type, the name(s) of the organism(s) sequenced, the next higher taxonomic rank, the sampling site and other significant details about the nature of the study. (Format Keywords)

# Abbreviations: (optional) (Heading 1)

Authors should include any non-standard abbreviations that are used throughout the article. Do not include well-known abbreviations (e.g., NCBI, EMBL, DNA, RNA) and do not use non-standard abbreviations for organism names. Species and subspecies names must be fully spelled out on first use as binomials (genus name and species epithet) or trinomials (genus name, species epithet subsp. subspecific epithet). Following first usage, the genus name may be abbreviated by using the first letter of the genus name, followed by a period and the epithets.

# Rationale (Heading 1)

The *Betaproteobacteria* are often the numerically dominant group in freshwater lakes [1]. Freshwater *Betaproteobacteria* have been classified into seven broad monophyletic lineages based on 16s rRNA gene sequences [2], one of which exclusively contains freshwater members of the family *Methylophilaceae* [2]. Freshwater 16s rRNA gene sequences within the *Methylophilaceae* fall into two clusters [2], known as LD28 and PRD01a011 [3]. Members of the LD28 group are "typical" freshwater bacteria, showing both persistence and abundance across lakes in Europe [3–8] and Asia [9]. The LD28 group is also closely related to the marine OM43 group of the *Methylophilaceae* [10,11]. Isolates of both LD28 and PRD01a001B have recently been obtained [12], revealing these microbes to be psychrophilic methylotrophs with small cell sizes and streamlined genomes (<2 Mb).

This publication describes one single-cell genome (SAG) and seven metagenome-assembled genomes (MAGs) for freshwater members of the *Methylophilaceae* from two lakes in Wisconsin, United States. These genomes are part of two larger sequencing projects. Single-cell genomes were collected in 2009, and selected to represent common and abundant freshwater bacteria. Metagenomes were collected at numerous time-points from 2005 to 2013 as part of a long-term monitoring project under the auspices of the National Science Foundation's Long-Term Ecological Research program. All genomes were sequenced and assembled through the United States Department of Energy's Joint Genome Institute (US-DoE-JGI). Metagenome-assembled genomes (MAGs) were also binned by the JGI.

# Organism Information (Heading 1)

## Classification (Heading 2)

A phylogenetic tree was constructed showing the relationship of Lake Mendota (ME) and Trout Bog (TB) genomes to other sequenced members of *Methylophilaceae*. The tree was built based on a concatenated alignment of 37 universal marker genes [13]. A a maximum-likelihood tree was constructed using RAxML. Allowing the software to determine the best protein substitution model (‘-m PROTGAMMAUTO’ flag). Bootstrap values are shown next to each node, as a percentage of 100 rapid bootstraps. Multiple marker genes are required because these genomes are incomplete (meaning any given marker gene is unlikely to be found in all genomes). For these 37 marker genes, phylogenetic trees built from individual markers have been shown to be congruent [14].

Phylogenetic analysis identified members of both the LD28 and PRD01a011 lineages, with the remainder of the samples belonging to the genus *Methylotenera* (Table 1). In contrast a previous study[12], our analysis indicates members of the genus *Methylopumilus* belong to separate phylogenetic groups, revealing an opportunity to further refine the phylogeny of the family *Methylophilaceae*.

# Genome sequencing information (Heading 1)

## Genome project history (Heading 2)

This section of the manuscript should provide a detailed summary of the sequencing, assembly binning (for metagenome-assembled genomes) and annotation methodology. The section should include an introductory paragraph that provides the readers with specific information about the sequencing project, when the project began and was completed, whether the sequence is complete or remains as a draft genome, and the quality of the draft, which public databases contain the project data and other relevant information. These data should be summarized in Table 2. If the genome collection contains reports from multiple projects, the authors need to provide the preceding information for each project.

## Growth conditions and genomic DNA preparation (Heading 2)

In the case of cultivated organisms, please provide the source of the organism (e.g., culture collection and accession number) and the conditions that were used to grow the strains(s) for DNA extraction (media, temperature, aeration, volume of culture, length of incubation). For all genomes, provide the method used to harvest and lyse the cells, and to extract and purify the DNA and to assess its purity. For single-cell genomes, authors should provide on the MDA protocol used.

## Genome sequencing, assembly, and binning (Heading 2)

Provide a succinct and detailed description of the methods used to sequence, assemble, and bin the genome(s). Identify the sequencing center where the work was performed, the sequencing technology(ies) used, library construction, number of reads and read length. Cite any relevant references regarding methods used. Also, provide a succinct and detailed description of the assembly and binning, including the software used for preliminary assembly, finishing and error checking and correction of mis-assemblies. Provide a brief description of the size of the final assembly, the number of contigs, and coverage.

## Genome annotation (Heading 2)

Provide a brief and succinct description of the methods used to identify and annotate genes, and any software used in the annotation pipeline.

# Genome Properties (Heading 1)

Provide a summary description of the size of the genome(s) (in base pairs), the number of chromosomes and plasmids. For single-cell genomes and metagenome-assembled genomes, describe how genome completeness was estimated. Include the number of predicted genes (RNA genes, protein coding genes, pseudogenes) by number and percent of total. This section should be linked to a chromosome map, map(s) of any plasmids and two or three tables providing a more detailed summary of the genome properties.

# Insights from the genome sequences (optional) (Heading l)

In many cases, authors may wish to provide a brief, yet more detailed description of major findings arising from the genome sequence. This can be a comparison of the genomes themselves, such a discussion of shared gene content or pairwise ANI values. This could also be a comparison of major differences found between the genome sequences that are the subject of the study and others (e.g., major differences is specific metabolic pathways, significant differences in gene content, etc.). This section is intended to permit the authors to make preliminary observations rather than to serve as detailed comparative study. In a short genome report, this section should be limited to two to three paragraphs. Authors wanting to provide greater detail and to incorporate additional genomes into their study, or to incorporate additional tables and figures are invited to submit their articles as extended genome reports.

## Extended insights (Heading 2)

Authors are encouraged to provide more detailed descriptions about insights gained from the genome sequence. This may include comparisons of the genome to that of closely relate species, detailed discussions about specific metabolic pathways that are noteworthy or unique, or other features that may be of interest to the readers. Authors may include additional tables and figures in this section (see below).

# Conclusions (Heading 1)

Provide a brief summary of the findings arising from the genome sequence. This should place the current genome into context with genomes sharing the same unifying feature.

# Taxonomic and nomenclatural proposals (optional) (Heading 1).

Authors are free to make taxonomic proposals and revisions of existing taxa, providing that the proposals are made in accordance with the rules of the relevant code of nomenclature. Taxonomic proposals must include the following sections appearing after the Conclusions section: A formal description for each taxon - Each taxonomic proposal must have its own subsection heading, and must appear in the proper order. Proposals for new genera must precede proposals for new species or subspecies. New species must precede new subspecies. A proposed name and etymology – For each new taxon proposed authors must propose a new name, in accordance to the appropriate rules of nomenclature. The proposed name should be followed by the etymology of the name, in grammatically correct Latin. *Authors are responsible for ensuring that proposed names meet these requirements.*

A protologue – for each new taxon and name that is proposed, authors must provide a description (also referred to as a diagnosis in botany) that provides readers with a summarized statement of differential features that can be used to distinguish the proposed taxon from other, closely related taxa. The protologue should include information about the morphology, physiology, habitat and genetics, along with any marker genes or features that can be used for identification purposes. The protologue must conclude with a statement that positively establishes the type strain (prokaryotes) or specimen (botany and zoology). If a new species or subspecies of bacteria or archaea is proposed, authors must provide the accession numbers from at least two internationally recognized culture collections (in different countries) from which viable samples of the type strain are available without restriction. Proposals that fail to provide this information cannot be considered for valid publication. If one or more new genera are proposed, the genus name(s) and description(s) must precede those of newly proposed member species. Genus descriptions must indicate the type species of the genus and differential/diagnostic features. In many cases, these may be the same as that for member species. Proposals for novel higher taxa (family and above) should appear after the species or subspecies proposals. Emendations of existing taxa should be made in separate sections, indicating the changes in membership and phenotypic and genotypic characteristics on which the taxa were originally formed. Assertions of synonymy should be presented in a separate section, with a full description of which taxa are being combined and an assertion of which name has priority. Taxonomic proposals of eukaryotic and virus taxa will follow the same general outline described above, but the identification and deposition of type material differs.

# Authors' contributions

In order to give appropriate credit to each author of a paper, the individual contributions of authors to the manuscript should be specified in this section.

According to [ICMJE guidelines](http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html), An 'author' is generally considered to be someone who has made substantive intellectual contributions to a published study. To qualify as an author one should 1) have made substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; 2) have been involved in drafting the manuscript or revising it critically for important intellectual content; 3) have given final approval of the version to be published; and 4) agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Each author should have participated sufficiently in the work to take public responsibility for appropriate portions of the content. Acquisition of funding, collection of data, or general supervision of the research group, alone, does not justify authorship.

We suggest the following kind of format (please use initials to refer to each author's contribution): AB carried out the molecular genetic studies, participated in the sequence alignment and drafted the manuscript. JY carried out the immunoassays. MT participated in the sequence alignment. ES participated in the design of the study and performed the statistical analysis. FG conceived of the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

All contributors who do not meet the criteria for authorship should be listed in an acknowledgements section. Examples of those who might be acknowledged include a person who provided purely technical help, writing assistance, or a department chair who provided only general support.

# Acknowledgements (heading 1 style)

Authors are encouraged to include an acknowledgement section recognizing the contributions to their work made by non-authors (skilled technicians, contributors of materials or those providing specialized advice or commentary on various aspects of their work. Authors should also acknowledge the source of funding for their work provided by various agencies, foundations or benefactors as well as the relevant contract or agreement numbers.

# References (heading 1 style)

Authors are expected to fully reference the relevant work of others in their reports. In addition to the citation of specific references describing the source organism and the sequencing and annotation methods, authors must also cite the appropriate references for the standards that are applied throughout the work. This includes the proper citation of the relevant literature for organism and gene nomenclature, ontologies, and other core methodologies so that readers may be able to follow backward and forward pointing links through the literature.

All references, including URLs, must be numbered consecutively, in square brackets, in the order in which they are cited in the text, followed by any in tables or legends. Each reference must have an individual reference number. Please avoid excessive referencing. If automatic numbering systems are used, the reference numbers must be finalized and the bibliography must be fully formatted before submission.

## Article within a journal

Smith JJ. The world of science. Am J Sci. 1999;36:234-5.

**Note that the journal style lists the first six authors before inserting *et al*. Note also that PubMed identifiers and DOIs will be added to the bibliography during production**.

## In-press article

Smith JJ. The world of science. Am J Sci. in press

**Article within a journal (no page numbers)**  
Rohrmann S, Overvad K, Bueno-de-Mesquita HB, Jakobsen MU, Egeberg R, Tjønneland A, et al. Meat consumption and mortality - results from the European Prospective Investigation into Cancer and Nutrition. BMC Medicine. 2013;11:63.

**Article within a journal by DOI**  
Slifka MK, Whitton JL. Clinical implications of dysregulated cytokine production. Dig J Mol Med. 2000; doi:10.1007/s801090000086.

**Article within a journal supplement**  
Frumin AM, Nussbaum J, Esposito M. Functional asplenia: demonstration of splenic activity by bone marrow scan. Blood 1979;59 Suppl 1:26-32.

**Book chapter, or an article within a book**  
Wyllie AH, Kerr JFR, Currie AR. Cell death: the significance of apoptosis. In: Bourne GH, Danielli JF, Jeon KW, editors. International review of cytology. London: Academic; 1980. p. 251-306.

**OnlineFirst chapter in a series (without a volume designation but with a DOI)**  
Saito Y, Hyuga H. Rate equation approaches to amplification of enantiomeric excess and chiral symmetry breaking. Top Curr Chem. 2007. doi:10.1007/128\_2006\_108.

**Complete book, authored**  
Blenkinsopp A, Paxton P. Symptoms in the pharmacy: a guide to the management of common illness. 3rd ed. Oxford: Blackwell Science; 1998.

**Online document**   
Doe J. Title of subordinate document. In: The dictionary of substances and their effects. Royal Society of Chemistry. 1999. http://www.rsc.org/dose/title of subordinate document. Accessed 15 Jan 1999.

**Online database**  
Healthwise Knowledgebase. US Pharmacopeia, Rockville. 1998. http://www.healthwise.org. Accessed 21 Sept 1998.

**Supplementary material/private homepage**  
Doe J. Title of supplementary material. 2000. http://www.privatehomepage.com. Accessed 22 Feb 2000.

**University site**  
Doe, J: Title of preprint. http://www.uni-heidelberg.de/mydata.html (1999). Accessed 25 Dec 1999.

**FTP site**  
Doe, J: Trivial HTTP, RFC2169. ftp://ftp.isi.edu/in-notes/rfc2169.txt (1999). Accessed 12 Nov 1999.

**Organization site**  
ISSN International Centre: The ISSN register. http://www.issn.org (2006). Accessed 20 Feb 2007.

**Dataset with persistent identifier**  
Zheng L-Y, Guo X-S, He B, Sun L-J, Peng Y, Dong S-S, et al. Genome data from sweet and grain sorghum (Sorghum bicolor). GigaScience Database. 2011. http://dx.doi.org/10.5524/100012.

**Citable Micropublication**  
Name Abstract for Escherichia coli. NamesforLife, LLC. Retrieved June 11, 2014. http://doi.org/10.1601/nm.3093.

# Table 1 (required, fixed format)

The taxonomic placement and names must be referenced to the alignment as described in Organism Information. For isolate genomes, summarized phenotypic features are based on either published reports from the literature. For SAGs and MAGs, summarized phenotypic features are based on the authors’ observations. Highlighted phenotypic features are required only for isolate genomes.

**Table 1.** Classification and general features of *Genusspecies* strain designationT [cite MIGS reference]

|  |  |  |  |
| --- | --- | --- | --- |
| **MIGS ID** | **Property** | **Term** | **Evidence codea** |
|  | Classification | Domain *Use only validly published or* | TAS [] |
|  |  | Phylum *available names. Cite the* | TAS [] |
|  |  | Class *appropriate taxonomic authority* | TAS [] |
|  |  | Order *and references establishing the name* | TAS [] |
|  |  | Family -. | TAS [] |
|  |  | Genus | TAS [] |
|  |  | Species | TAS [] |
|  |  | (Type) strain: *StrainT (Accession #s)* |  |
|  | Gram stain | *Positive/negative/vaiable* | TAS [] |
|  | Cell shape | *Rod/coccus/filments/chains, etc* | TAS [] |
|  | Motility | *Motile/non-motile* | TAS [] |
|  | Sporulation | *Spore type/position or not reported* | NAS |
|  | Temperature range | *°C* | TAS [] |
|  | Optimum temperature | *°C* | TAS [] |
|  | pH range; Optimum | *3.5–6.5; 5* | TAS [] |
|  | Carbon source | *Specify known carbon sources sustaining growth* | TAS [] |
| MIGS-6 | Habitat |  | TAS [] |
| MIGS-6.3 | Salinity | *as% NaCl (w/v)* | TAS [] |
| MIGS-22 | Oxygen requirement | *Aerobic/anaerobic/microaerophilic/aerotolerant* | TAS [] |
| MIGS-15 | Biotic relationship | *free-living/symbiont/commensal* | TAS [] |
| MIGS-14 | Pathogenicity | *Pathogenic/non-pathogen* | NAS |
| MIGS-4 | Geographic location | *Country/region* | TAS [] |
| MIGS-5 | Sample collection | *Date* | TAS [] |
| MIGS-4.1 | Latitude | *DMS* | TAS [] |
| MIGS-4.2 | Longitude | *DMS* | TAS [] |
| MIGS-4.4 | Altitude | *M* | TAS [] |

a Evidence codes - IDA: Inferred from Direct Assay; TAS: Traceable Author Statement (i.e., a direct report exists in the literature); NAS: Non-traceable Author Statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). These evidence codes are from the Gene Ontology project [cite this reference]

# Table 2 (required, fixed format)

Project Information. Authors must provide the requested data, in conformance with the MIGS standard. Supply the information for column 3. Do not modify the order of columns or rows.

**Table 2.** Project information.

|  |  |  |
| --- | --- | --- |
| **MIGS ID** | **Property** | **Term** |
|  | Genome Type | One of: isolate, single cell, population |
| MIGS 31 | Finishing quality |  |
| MIGS-28 | Libraries used |  |
| MIGS 29 | Sequencing platforms |  |
| MIGS 31.2 | Fold coverage |  |
| MIGS 30 | Assemblers |  |
|  | Binning method |  |
| MIGS 32 | Gene calling method |  |
|  | Locus Tag |  |
|  | Genbank ID |  |
|  | GenBank Date of Release |  |
|  | GOLD ID |  |
|  | BIOPROJECT |  |
| MIGS 13 | Source Material Identifier |  |
|  | Project relevance |  |

# Table 3. (Optional, fixed format)

Summary of genome. This table should only be used if the report describes a chromosome and one or more plasmids.

**Table 3.** Summary of genome: one chromosome and X plasmids

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Label | Size (Mb) | Topology | INSDC identifier | RefSeq ID |
| Chromosome |  |  |  |  |
| Plasmid 1 |  |  |  |  |
| Plasmid 2 |  |  |  |  |
| Plasmid 3 |  |  |  |  |
| Plasmid 4 |  |  |  |  |
| Plasmid 5 |  |  |  |  |

# Table, 4 or 3 if optional table is not used (required, fixed format)

Genome statistics, listed in base pairs and percent of total. Provide values for columns 2 and 3. Do not modify ordering of rows. Genome size, DNA coding, and DNA G+C should be reported as the actual size, not as estimates for a complete genome.

**Table 4**. Genome statistics.

|  |  |  |
| --- | --- | --- |
| Attribute | Value | % of Total |
| Genome size (bp) |  |  |
| DNA coding (bp) |  |  |
| DNA G+C (bp) |  |  |
| DNA scaffolds |  |  |
| Total genes |  |  |
| Estimate completeness (%) |  |  |
| Estimated genome size (bp) |  |  |
| Protein coding genes |  |  |
| RNA genes |  |  |
| Pseudo genes |  |  |
| Genes in internal clusters |  |  |
| Genes with function prediction |  |  |
| Genes assigned to COGs |  |  |
| Genes with Pfam domains |  |  |
| Genes with signal peptides |  |  |
| Genes with transmembrane helices |  |  |
| CRISPR repeats |  |  |

# Table 5 or 4 if optional table is not used (required, fixed format)

Number of genes associated with general COG functional categories. Provide values for columns 2 and 3. Do not modify ordering of rows.

**Table 5**. Number of genes associated with general COG functional categories.

|  |  |  |  |
| --- | --- | --- | --- |
| Code | Value | %age | Description |
| J |  |  | Translation, ribosomal structure and biogenesis |
| A |  |  | RNA processing and modification |
| K |  |  | Transcription |
| L |  |  | Replication, recombination and repair |
| B |  |  | Chromatin structure and dynamics |
| D |  |  | Cell cycle control, Cell division, chromosome partitioning |
| V |  |  | Defense mechanisms |
| T |  |  | Signal transduction mechanisms |
| M |  |  | Cell wall/membrane biogenesis |
| N |  |  | Cell motility |
| U |  |  | Intracellular trafficking and secretion |
| O |  |  | Posttranslational modification, protein turnover, chaperones |
| C |  |  | Energy production and conversion |
| G |  |  | Carbohydrate transport and metabolism |
| E |  |  | Amino acid transport and metabolism |
| F |  |  | Nucleotide transport and metabolism |
| H |  |  | Coenzyme transport and metabolism |
| I |  |  | Lipid transport and metabolism |
| P |  |  | Inorganic ion transport and metabolism |
| Q |  |  | Secondary metabolites biosynthesis, transport and catabolism |
| R |  |  | General function prediction only |
| S |  |  | Function unknown |
| - |  |  | Not in COGs |

The total is based on the total number of protein coding genes in the genome.

# Instructions for additional tables.

Any additional tables must be formatted in the same manner as tables one through five. This format is detailed below in an example table:

**Table Number.** Table title

|  |  |  |  |
| --- | --- | --- | --- |
| **Table header** | - | - | - |
| Row 1 | - | - | - |
| Row 2 | - | - | - |
| Row 3 | - | - | - |

Table footer

Table title must not exceed one row and this row must begin with the table number. The top row of the table must have a top and bottom border and the bottom row of the table must have a bottom border. 12pt Times New Roman must be used for both table title and header. Table rows and footer must be 10 pt Times New Roman. Table footer is used to explain different elements. Authors should use superscript a, b, c, etc. to refer to the element which is being described (as in table one). All additional tables should be able to fit in a 10 × 9 space. Font should be no smaller than 10 pt. additional tables normally only appear in an extended genome report.

# Figure legends

The legends should be included in the main manuscript text file at the end of the document, rather than being a part of the figure file. For each figure, the following information should be provided: Figure number (in sequence, using Arabic numerals - i.e. Figure 1, 2, 3 etc.); short title of figure (maximum 15 words); detailed legend, up to 300 words.

**Please note that it is the responsibility of the author(s) to obtain permission from the copyright holder to reproduce figures or tables that have previously been published elsewhere.**

**Required Figure(s): a** phylogenetic tree indicating current placement and a photomicrograph or electron photomicrograph of the source organism. Authors should include a brief explanation as to the source of the data, algorithms used to create the tree and any relevant references. Each terminal node should indicate the current name of the species from the 16 rRNA gene or other marker gene originated, and the Genbank identifier. Type strains should be identified as such using a superscripted “T”. Authors should also include information about which species/strains used in a tree have a sequenced genome. Trees should not exceed a single page (7 x 9 in, including figure legend). Text should be no smaller than 8 pt when drawn to scale and should not overlap.

**Optional Figure: A genome map should be provided, for complete genome sequences, keyed to the COGS groups.**

# Figure 1

Macintosh HD:Users:joshamilton:Documents:Research:2015a-GenomeAnnouncements:GenAnnc Repository:phylogeny:betIVProvisional.pdf

Phylogenetic tree showing the relationship of Lake Mendota (ME) and Trout Bog (TB) genomes to members of *Methylophilaceae*. To build the tree, all finished genomes belonging to the family *Methylophilaceae* were downloaded from IMG. Additionally, draft genomes from the genera *Methylopumilus* (LD28) and *Methylotenera* (another freshwater group) were downloaded. Polynucleobacter and Limnohabitans were selected as outgroups due to their ubiquity in freshwater lakes.