Metabolic marker gene mining provides insight in global mcrA diversity and, coupled with targeted genome reconstruction, sheds light on metabolic versatility of the *Methanomassiliicoccales*

Daan R. Speth[1] and Victoria J. Orphan[1]

[1]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA, USA

Correspondence: dspeth@caltech.edu or vorphan@gps.caltech.edu

Abstract

Over the past years, metagenomics has revolutionized our view of microbial diversity. Moreover, extracting near-complete genomes from metagenomes has led to the discovery of known metabolic traits in unsuspected lineages. Genome-resolved metagenomics relies on assembly of the sequencing reads and subsequent binning of assembled contigs, which might be hampered by strain heterogeneity or low abundance of a target organism. Here we present a complementary approach, metagenome marker gene mining, and use it to assess the global diversity of archaeal methane metabolism through the *mcrA* gene. To this end, we have screened 18,465 metagenomes for the presence of reads matching a database representative of all known mcrA proteins and reconstructed gene sequences from the matching reads. We use our mcrA dataset to assess the environmental distribution of the *Methanomassiliicoccales* and reconstruct and analyze a draft genome belonging to the 'Lake Pavin cluster', an understudied environmental clade of the *Methanomassiliicoccales*. Thus, we show that marker gene mining can enhance the discovery power of metagenomics, by identifying novel lineages and aiding selection of targets for in-depth analyses. Marker gene mining is less sensitive to strain heterogeneity and has a lower abundance threshold than genome-resolved metagenomics, as it only requires short contigs and there is no binning step. Additionally, it is computationally cheaper than genome resolved metagenomics, since only a small subset of reads needs to be assembled. It is therefore a suitable approach to extract knowledge from the many publicly available sequencing projects.

Keywords: Marker genes ; data mining ; mcrA ; methanogens ; metagenomics ; *Methanomassiliicoccales*

Introduction

Genome resolved metagenomics is allowing unprecedented, primer independent, insight in the diversity of the microbial world (Hug *et al.*, 2016). In addition to the window into microbial diversity that metagenomics sequencing offers, it also provides clues for the metabolism of the organisms observed (Sharon and Banfield, 2013). More precisely, based on the presence (or absence) of homologs of previously studied genes, an educated guess of the metabolism of an organism can be made. This has amongst others, led to the recent discovery of complete ammonium oxidation (comammox) in a single organism (van Kessel *et al.*, 2015; Daims *et al.*, 2015) and provided evidence for archaeal methane metabolism outside of the *Euryarchaeota* (Evans *et al.*, 2015; Vanwonterghem *et al.*, 2016).

Other major advances in our understanding of the diversity of archaeal methane metabolism have come from cultivation studies, including the culturing and enrichment of members of the 7th *Euryarchaeal* order of methanogens, the *Methanomassiliicoccales* (Dridi *et al.*, 2012; Borrel *et al.*, 2012; Iino *et al.*, 2013; Borrel *et al.*, 2013a), and the recent culturing of halophilic methanogens from Siberian soda lakes (Sorokin *et al.*, 2017). The latter group seems to be restricted to highly saline environments, whereas environmental sequencing indicates that the *Methanomassiliicoccales* are widely distributed, occurring in habitats ranging from animal guts to wetlands and wastewater treatment (Großkopf *et al.*, 1998; Wright *et al.*, 2004; Iino *et al.*, 2013; Söllinger *et al.*, 2016). Indeed, a recent large-scale effort to bin genomes from environmental metagenome data recovered 66 *Methanomassiliicoccales* genomes (Parks *et al.*, 2017), further supporting the environmental relevance of this order. The recent advances in our understanding of the diversity of archaeal methane metabolism raise the question whether additional novel diversity exists within previously sequenced metagenomic datasets, and whether the environmental importance and diversity of understudied clades can be further illuminated.

In the examples of metagenomics enabled discovery discussed above, homologs of marker genes known to be diagnostic for methane metabolism were discovered in metagenome assembled genomes (MAGs).

This requires the assembly of the raw sequencing reads, and subsequent binning of the assembled contigs into draft genomes (Thomas *et al.*, 2012). Assembly is computationally expensive on large datasets, and strain diversity within a sample can result in highly fragmented assemblies (Thomas *et al.*, 2012). Automated binning has improved dramatically in recent years, but this process often still requires substantial time-consuming manual curation (Delmont *et al.*, 2017). Alternatively, the diversity of organisms capable of a metabolic process can be assessed using PCR-based screening of environmental samples. However, PCR-based analyses are sensitive to primer bias, and therefore unlikely to yield highly divergent gene sequences. In addition, the amplification of a single metabolic gene makes elucidation of the taxonomic affiliation of the organism containing the gene nearly impossible.

Directly mining metagenomic reads for marker genes, and subsequently reconstructing the full-length gene sequence, combines some of the advantages of both of these strategies, while minimizing the disadvantages. Using a curated database, reads can be confidently assigned to a gene of interest, with false positive removal using a BLAST Score Ratio (BSR) (Rasko *et al.*, 2005). If a divergent variant of a gene of interest is retrieved using this approach, the genome containing this gene can be retrieved from the source metagenome dataset using targeted (manual) binning. This last step will not always be successful, because less sequencing depth is required to assemble a single short contig consisting of one gene, rather than assemble longer contigs and confidently assign them to a draft genome. However, precisely because more sequencing depth is required to assemble and bin a draft genome than a single gene, marker gene mining might yield information from datasets were genome binning is not feasible (e.g. Lüke *et al.*, 2016).

We have previously used this approach to assess the presence of nitrogen cycle genes in datasets from the Arabian Sea oxygen minimum zone (Lüke *et al.*, 2016) and to assess the environmental distribution of organisms capable of complete ammonium oxidation (van Kessel *et al.*, 2015). Here we present a more systematic use of marker gene mining to assess the diversity of the *mcrA* gene, encoding the alpha

79 subunit of the methyl-coenzyme M reductase. This enzyme is essential for (reverse) methanogenesis,
80 where it catalyzes the final reduction and release of the methyl group on coenzyme M to methane, or
81 the initial oxidation of methane (Scheller *et al.*, 2010). Moreover, the *mcrA* gene, and other genes
82 required for methanogenesis, have recently been discovered in several unexpected clades of Archaea
83 (Evans *et al.*, 2015; Vanwonterghem *et al.*, 2016; Laso-Pérez *et al.*, 2016; Sorokin *et al.*, 2017), indicating
84 the possibility that more diversity has previously been overlooked.
85 We have screened the environmental metagenomic data available in the NCBI sequencing read archive
86 (Kodama *et al.*, 2012) and MG-RAST (Meyer *et al.*, 2008) for reads matching the mcrA gene. We
87 subsequently used the obtained data to assess the diversity and environmental distribution of the
88 *Methanomassiliicoccales* order (Paul *et al.*, 2012). Finally, we reconstruct and analyze a draft genome of
89 an organism belonging to the 'Lake Pavin cluster' an understudied lineage within this group (Borrel *et*
90 *al.*, 2013b).
91

Methods

Construction of the mcrA database from PFam and NCBI-nr

To construct a mcrA protein sequence database representative of the known global diversity, we first obtained the amino acid sequences in the Pfam (version 29.0; Finn *et al.*, 2013) families MCR_alpha_N (PF02745) and MCR_alpha (PF02249), representing the N-terminal and C-terminal parts of the mcrA protein. We assessed the completeness of the Pfam dataset by downloading the NCBI-nr protein reference database in fasta format (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/) and using it as query for a DIAMOND (Buchfink *et al.*, 2014) search against the Pfam dataset. Subsequently, we calculated the BLAST score ratio (BSR) (Rasko *et al.*, 2005) between the score against the Pfam dataset and the maximum possible score (a self-hit) of the 16,260 sequences that had a DIAMOND hit against the Pfam dataset (Supplemental Figure 1). This allowed us to identify real mcrA sequences not covered in the Pfam, while eliminating false positive hits. Using this method, we identified that mcrA sequences from the clades *Bathyarchaeota*, *Methanofastidiosales* (WSA2), ANME-1, *Methermicoccus* and *Methanoperedens* (ANME-2d) were not represented in the Pfam at the time of database construction (June 2016) (Supplemental Figure 1). All 203 full-length mcrA sequences in the NCBI-nr were added to the mcrA dataset, which was subsequently clustered at 90% identity using UCLUST (Edgar, 2010), resulting in a mcrA reference database containing 69 non-redundant full-length sequences representing the full diversity of mcrA sequences included in the NCBI-nr (June 2016).

mcrA read data acquisition from SRA and MG-RAST

To obtain a list of metagenome datasets of potential interest, metadata was downloaded for all runs in the sequencing read archive (SRA) (https://www.ncbi.nlm.nih.gov/sra) corresponding to the query "metagenomic AND WGS NOT human NOT gut NOT oral" (gut and oral were excluded because of the

high number of datasets from these environments) on June 16th 2016. Additionally, metadata for all runs in MG-RAST (https://metagenomics.anl.gov/) was downloaded, and all datasets labeled 'WGS' (whole genome shotgun) were selected. This resulted in a list of 10,613 SRA run accession numbers and 7,852 MG-RAST identifiers, representing over 60 Terabases of sequencing data (Figure 1, Supplemental files 1 and 2). As storing this amount of sequence data was not feasible, the accession lists were used as input for the 'sra_trawler.sh' and 'mgrast_trawler.sh' shell scripts, respectively. Briefly, these scripts download an accession number on the list, then use the downloaded dataset as query in a DIAMOND (Buchfink *et al.*, 2014) search against a database of interest (in this case the mcrA protein database described above). Then, hits are written to a new file in fasta format using the script 'blast_based_read_lookup.pl', the dataset is discarded, and the process is repeated for the next dataset. Scripts are available at (https://github.com/dspeth/bioinfo_scripts/metagenome_screening). Using this approach, it took approximately 4 months to process the 18,465 selected datasets using 16 cores on a server. To speed up the process, the sra-trawler.sh script uses the sra-toolkit (https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/) to split paired-end files and only analyzes the forward reads of paired-end datasets. 2,083,349 reads from 6,105 datasets matched the mcrA database and were combined and used as the query in a DIAMOND search against the NCBI-nr database to calculate the BSR between the hit against the mcrA database and the NCBI-nr (Figure 1). This results in a BSR between the score against the database of interest and the score against an outgroup, rather than a BSR between the score against the database of interest and a self-hit described above for mcrA database construction. Applying BSR in this manner allows for detection of real mcrA sequences with low identity to the database while still removing false positives (Lüke *et al.*, 2016). The resulting 2,081,632 post-BSR hits originated from 3,684 datasets (Figure 1). Dataset SRR398144, an mcrA amplicon sequencing effort (Denonfoux *et al.*, 2013), accounted for nearly 10% of these hits (197,371).

As the aim of our work was to reconstruct near full-length mcrA sequences, SRR398144 was excluded and the remaining steps were done with 1,884,261 reads.

Assessing global mcrA protein diversity

The true-positive reads from 1,080 datasets containing over 20 hits, enough for approximately 2-fold gene coverage in a 150bp dataset, were assembled separately using IDBA-UD (Peng *et al.*, 2012), resulting in 1,511 sequences >1,000 bp. Prodigal (Hyatt *et al.*, 2010) was used for open reading frame prediction on the assembled contigs and mcrA protein sequences over 300 amino acids were selected. The resulting 1,460 sequences were sorted by length and added to the mcrA database. Additionally, the eight recently published *Syntrophoarchaeum* mcrA sequences (Laso-Pérez *et al.*, 2016) were added to the mcrA database, and the new set was clustered using UCLUST at 90% identity (Edgar, 2010). The resulting 150 sequences were aligned using MUSCLE (Edgar, 2004) and a maximum likelihood phylogeny was calculated using RAxML (Stamatakis, 2014), which was visualized using iTOL (Letunic and Bork, 2016; Figure 2). To estimate previous detection of the mcrA sequences, all 15,888 hits post-BSR selection in the NCBI-nr that were too short to be included in the database (PCR products) were used as DIAMOND query against the 150 non-redundant mcrA sequences (Supplemental figure 2A). Additionally, all 1,884,261 metagenome hits were used as DIAMOND query against the 149 non-redundant mcrA sequences (Supplemental figure 2B). Relative counts of all 13,007 NCBI-nr hits and 1,487,226 metagenomic reads over 90% identity (as these would be clustered with the sequence) were visualized using iTOL. Reconstructed gene sequences are included as supplemental file 3 and 4, containing the nucleotide and amino acid sequences respectively.

Environmental distribution of the *Methanomassiliicoccales*

The assembled mcrA sequences belonging to the *Methanomassiliicoccales* clade were obtained using BSR between the best hit against the five *Methanomassiliicoccales* mcrA sequences present in our mcrA database, before addition of the newly assembled sequences (Figure 2), and the best hit against our entire mcrA database. Sequences with a BSR over 0.75 were assigned to the *Methanomassiliicoccales* (Supplemental figure 2). The resulting 116 sequences, combined with the five reference sequences discussed above and a *Methanofastidiosales* sequence (KYC45731.1, as outgroup), were aligned using MUSCLE (Edgar, 2004) and a phylogeny was calculated using RaxML (Stamatakis, 2014) that was visualized using iTOL (Letunic and Bork, 2016). The source environment of each sequence was assigned manually, using the NCBI-SRA or MG-RAST sample record.

Genome reconstruction of a Lake Pavin cluster *Methanomassiliicoccales*

Reads from dataset SRR636597 (Tan *et al.*, 2015) were assembled using Megahit (Li *et al.*, 2015) resulting in 125,408 contigs longer than 1,000bp. Reads from datasets SRR636597, SRR636559, and SRR636569, which contained highly similar mcrA sequences (Figure 3), were mapped to the assembled contigs using BBmap (http://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/). Kmer frequency of the contigs was calculated using the script calc.kmerfreq.pl by Mads Albertsen (https://github.com/MadsAlbertsen/multi-metagenome/tree/master/R.data.generation). The contig fasta file was converted to tab-delimited form and length and GC content were calculated using fasta_to_gc_length_tab.pl (https://github.com/dspeth/bioinfo_scripts/metagenome_screening). The contig containing the mcrA sequence of interest was identified using DIAMOND (Buchfink *et al.*, 2014) and used to guide the manual binning of the cluster representing the draft genome containing this gene, using R, as previously described (Speth *et al.*, 2016). The resulting 1.6Mbp draft genome was further refined by 10 cycles of mapping with BBmap and assembly with SPAdes (Bankevich *et al.*, 2012) using the iterative_denovo_spades.sh shell script

(https://github.com/dspeth/bioinfo_scripts/metagenome_screening). A final manual refining was performed using anvi'o, using the coverage of the contigs in datasets SRR636597, SRR636559, and SRR636569 to remove contaminating contigs from the bin (Eren *et al.* 2015). The final 1.55Mb draft genome on 58 contigs was quality checked using checkM (Parks *et al.*, 2015), annotated using Prokka (Seemann, 2014), analyzed using Artemis (Carver *et al.*, 2012), compared to the other available *Methanomassiliicoccales* genomes using ProteinOrtho (Lechner *et al.*, 2011), and the annotation was manually curated. As NCBI does not accept third party annotation without experimental validation (see: https://www.ncbi.nlm.nih.gov/genbank/tpa/), the assembled annotated draft genome is included as supplemental file 5.

Results & discussion

McrA gene diversity across all sampled datasets

To leverage the available metagenomic sequencing data for a diversity analysis of functional marker genes, we established a workflow based on automated sequential downloading and processing of the public data in the Sequencing Read Archive (SRA) and the Metagenomics RAST (MG-RAST) repositories (Figure 1). We applied this workflow to the *mcrA* gene, a marker for the production and anaerobic oxidation methane, because of the environmental relevance of these processes (Knittel and Boetius, 2009). Additionally, the *mcrA* gene has recently been discovered in several clades of *Archaea*, indicating that methane metabolism is more widespread in the domain than previously thought (Mondav *et al.*, 2014; Evans *et al.*, 2015; Vanwonterghem *et al.*, 2016; Sorokin *et al.*, 2017). Notably, *Syntrophoarchaeum*, an archaeon in a syntrophic anaerobic butane-oxidizing enrichment culture, contained multiple copies of highly divergent *mcrA* genes thought to be involved in the activation of higher alkanes, but not methane (Laso-Pérez *et al.*, 2016). Of these recently discovered groups, only the *Bathyarchaeal* mcrA sequences were included in our reference database, as they were present in the NCBI-nr database at the time of database construction (June 2016). The mcrA sequences of the *Verstraetearchaeota*, *Syntrophoarchaea*, *Methanonatronarchaeia*, and *Methanoflorens* were not present in the NCBI-nr as of June 2016. Our analysis does retrieve the mcrA sequences associated with *Verstraetearcheota*, *Methanonatronarchaeia*, and *Methanoflorens* (Figure 2), but was not sensitive enough to retrieve sequences related to the highly divergent *Syntrophoarchaea*, simultaneously illustrating both the power to detect novel diversity, and the limit of screening of metagenomic reads based on sequence identity. The HMM based search strategy implemented in GraftM (Boyd *et al.*, 2018), offers a complementary strategy for mining unassembled reads, further increasing the potential for discovery of novel diversity.

Besides the independent recovery of the *Verstraetearcheota*, *Methanonatronarchaeia*, and *Methanoflorens* from public databases, two sequences deeply branching within the *Euryarchaeota* were retrieved. Both sequences were assembled from dataset mgm4537093.3, originating from a marine sediment sample at the oil seeps of the coast of Santa Barbara, California. One of these sequences is basal to the *Methanomassiliicoccales*, and distantly related to *Methanonatronarchaeia*, and the other sequence is basal to the *Methanosarcinales*/*Methanocellales*/*Methanomicrobiales* cluster (Figure 2). However, the organisms containing these sequences were not present in sufficient abundance in their respective samples to extract a draft genome from the metagenome, leaving the taxonomic association of these divergent sequences unclear. Furthermore, several sequences associated with anaerobic methanotrophic (ANME) archaea were retrieved (Figure 2), including a sequence from a South African gold mine (Lau *et al.*, 2014) related to ANME-1, and the first full length ANME-3 mcrA sequence from a dataset obtained from the deep-sea Haakon Mosby mud volcano, the site where this group was originally discovered (Niemann *et al.*, 2006; Lösekann *et al.*, 2007). Our analysis also substantially expands the known diversity of mcrA sequences of methanogenic clades within the *Euryarchaeota* (Figure 2). Aligning the BSR-filtered reads from our analysis to the mcrA database amended with the newly assembled sequences shows higher average sequence identity of the aligned sequences after addition of the new sequences. Before our analysis, 555,598 reads (29.5% of BSR-filtered reads) had lower than 90% identity to a sequence in our reference database, whereas after addition of the newly retrieved sequences that number dropped to 313,904 reads below 90% identity (16.6% of BSR-filtered reads; Figure 2; Supplemental Figure 3). For the PCR amplicons present in the NCBI-nr (Supplemental Figure 1) these numbers were comparable, with 4628 (29.1%) sequences below 90% identity to any sequence in the database before including the newly retrieved sequences, and 2841 (17.9%) below 90% identity afterwards. Although this is an improvement, the high number of reads still unassigned (>90% sequence identity to any database sequence) does indicate there are yet more divergent mcrA variants

to be discovered (Supplemental Figure 3). This highlights the potential for ongoing exploration of metagenomic sequencing data as it becomes available, using our sequence-identity based analysis, and an HMM based approach (Boyd *et al.*, 2018). The highest number of novel sequences retrieved in our survey of the SRA and MG-RAST was associated with the *Methanomassiliicoccales*, bringing the known diversity within this recently discovered order on par with that of more intensively studied groups.

Environmental distribution of the *Methanomassiliicoccales*

The *Methanomassiliicoccales* are the most recently described order of methanogens within the *Euryarchaeota* (Tajima *et al.*, 2001; Paul *et al.*, 2012; Dridi *et al.*, 2012) and, compared to other methanogenic lineages, the diversity and environmental distribution of this clade is least comprehensively studied (Figure 2). It has recently been proposed that the *Methanomassiliicoccales* should be divided into an environmental clade and a gastrointestinal tract (GIT) clade (Söllinger *et al.*, 2016). In addition, a 'Lake Pavin' clade, named after the site where it was detected, was previously proposed based on analysis of environmental 16S ribosomal RNA gene sequencing (Borrel *et al.*, 2013b). Another recent study suggested the existence of *Methanomassiliicoccales* in marine sediments, based on the presence of butanetriol dibiphytanyl glycerol tetraether (BDGT) lipids in *Methanomassiliicoccus luminyensis*, and the detection of these lipids in marine sediments (Becker *et al.*, 2016), but the specificity of this biomarker is unclear.

To further investigate the diversity and environmental distribution of the *Methanomassiliicoccales*, we retrieved all 116 mcrA sequences over 300 amino acids belonging to this clade from our analysis and assessed their environmental origin. This analysis confirmed the presence of three major clades ('GIT', 'Environmental' and 'Lake Pavin'), but unlike the study by Söllinger *et al.* we do not observe clear clustering by environment (Figure 3). Even though our original dataset selection was biased against gut samples by the SRA query used, several sequences of fecal origin are represented in both the

environmental and GIT clades (Figure 3). Conversely, both the 'Environmental' and 'GIT' clades contain many sequences originating from the same environment; a single study that characterizes the microbial diversity in open fermentation pits for liquor production (Guo *et al.*, 2014). None of the assembled sequences originated from marine sediments, implying that the *Methanomassiliicoccales* are not abundant in these systems, and that other archaeal clades (possibly) associated with the *Thermoplasmatales* are responsible for the detected BDGT lipids (Becker *et al.*, 2016).

Four sequences, of which three were nearly identical, belonging to the 'Lake Pavin' clade were retrieved in our analysis (Figure 3). As this group was previously only detected by environmental PCR, and genomic data is lacking, we focused on the three near-identical sequences for more in-depth analysis. Using the mcrA sequences as a guide, we extracted a draft genome of a representative of this clade from dataset SRR636597, originating from an oil mining tailing pond (Tan *et al.*, 2015). This draft genome is referred to as MALP (MAssiliicoccales Lake Pavin) for the remainder of the manuscript.

Genomic analysis of MALP, a representative of 'Lake Pavin' clade *Methanomassiliicoccales*

The MALP draft genome assembled and binned from dataset SRR636597 consists of 1.55 megabases on 58 contigs, and is over 92% complete with 0% contamination, as assessed using checkM (Parks *et al.*, 2015). A recent assembly and binning effort of 1550 SRA datasets recovered a highly similar MAG, likely representing the same microbial population, designated UBA248 (Parks *et al.*, 2017). At an estimated 89% completeness, 1.6% contamination, and on 138 contigs UBA248 is somewhat more fragmented and slightly lower quality. Furthermore, Parks *et al.* did not perform any analysis on this specific MAG. In addition, Parks *et al.* also assembled a related MAG, designated UBA472, from a different site. Contig alignment using Mauve (Rissman *et al.*, 2009) and comparison of gene content using anvi'o (Eren *et al.*, 2015) supports that MALP and UBA248 represent the same population, whereas UBA472 likely represents a closely related organism (Supplemental Figure 4).

The size of the MALP genome falls within the range of previously obtained *Methanomassiliicoccales* genomes (1.4-2.6 Mbp; Borrel *et al.*, 2014; Lang *et al.*, 2015). As in the other *Methanomassiliicoccales*, the ribosomal rRNA genes in MALP are not organized in an operon, and it encodes two copies of the 5S rRNA gene. Our mcrA analysis and BLAST searches using the MALP 16S rRNA gene indicate that organisms closely related to MALP (>98% 16S rRNA gene identity) are previously found in (contaminated) sediments and wastewater treatment systems, but not in fecal samples.

MALP encodes all genes required for hydrogen dependent reduction of methanol to methane as proposed in the other *Methanomassiliicoccales* (Figure 4A). These include the methanol:CoM methyltransferases (*mtaABC*) for formation of methyl-CoM (Sauer and Thauer, 1999), methyl-CoM reductase (*mcrABG*) to release methane and form the CoM-CoB heterodisulfide (Ermler *et al.*, 1997), soluble heterodisulfide reductase (*hdrABC*) and [NiFe]-hydrogenase (*mvhAGD*) to reduce ferredoxin and heterodisulfide coupled to hydrogen oxidation (Thauer *et al.*, 2008; Wagner *et al.*, 2017), and the Fpo-like complex (*fpoABCDHIJKLMN*) proposed to oxidize ferredoxin and establish a proton gradient (Welte and Deppenmeier, 2011), potentially coupled to heterodisulfide reduction using *hdrD* (figure 4) (Lang *et al.*, 2015). Like "*Candidatus* Methanomethylophilus alvus", MALP does not encode an energy conserving hydrogenase to couple ferredoxin oxidation to hydrogen formation and the buildup of a proton gradient (Borrel *et al.*, 2014).

From both physiological studies on the *Methanomassiliicoccus luminyensis* culture and the "*Candidatus* Methanomethylophilus alvus" enrichment, as well as comparative genomics, it has become clear that previously enriched *Methanomassiliicoccales* are also capable of growth on methylamines (Borrel *et al.*, 2014; Lang *et al.*, 2015). Metabolizing methylamines requires pyrrolysine-containing methyl transferases (*mtmBC*/*mtbBC*/*mttBC*), which are present in the genomes of the previously sequenced *Methanomassiliicoccales*. In contrast, the reconstructed MALP genome does not encode the pyrrolysine-containing methyltransferases, the operon for pyrrolysine biosynthesis, the pyrrolysine tRNA

316 synthetase, or the pyrrolysine tRNA. Considering that MALP is a representative of the most basal cluster
317 of *Methanomassiliicoccales* (Figure 3), the absence of pyrrolysine usage from the genome suggests the
318 ability to generate methane from methylated amines was acquired recently within the
319 *Methanomassiliicoccales* order. In agreement with this, a previous comparative genomics study found
320 that the number of pyrrolysine containing genes in other sequenced *Methanomassiliicoccales* varied
321 between 3 (*Methanomassiliicoccus luminyensis*) and 19 ('*Candidatus* Methanomethylophilus alvus')
322 (Borrel *et al.*, 2014).
323 Another unexpected feature of the MALP genome was the presence of the two catalytic subunits of N5-
324 methyltetrahydromethanopterin:CoM methyltransferase (*mtrAH*) (Wagner *et al.*, 2016). In CO_2-reducing
325 methanogens methyltetrahydromethanopterin:CoM methyltransferase is an eight-subunit membrane-
326 associated complex that catalyzes the second to last step of the methanogenic pathway, the transfer of
327 a methyl group from tetrahydromethanopterin (THMPT) to coenzyme M, coupled to translocation of a
328 sodium ion. However, only the catalytic subunits (*mtrAH*), and none of the membrane associated
329 subunits (*mtrB-G*), are present in the MALP genome. This is surprising, as there is no known role for
330 THMPT in *Methanomassiliicoccales*, including MALP. The *mtrA* subunit, which donates the methyl group
331 to Coenzyme M is conserved, but contains an unusual C-terminal extension. On the other hand, the
332 *mtrH* subunit that is responsible for the transfer of the methyl group from THMPT is divergent from the
333 *mtrH* of CO_2-reducing methanogens. Considering the likely absence of THMPT from MALP, and the
334 divergent *mtrH* subunit, we propose this *mtrAH* may be a N5-methyltetrahydrofolate:CoM
335 methyltransferase instead (figure 4B).
336 In addition to the *mtrAH* genes, MALP encodes acetyl-CoA synthetase for the formation of acetyl-CoA
337 from acetate (Jetten *et al.*, 1989), and acetyl-CoA synthase/CO dehydrogenase for the
338 disproportionation of acetyl-CoA into methyl-tetrahydromethanopterin and CO, and ferredoxin
339 dependent oxidation of CO to CO_2 (figure 4B) (Ferry, 1992). This gene set, combined with the presence

of the Fpo-like complex, in theory provides MALP with a complete pathway for aceticlastic methanogenesis, similar to the pathway observed in *Methanosaeta thermophila* (figure 4B) (Welte and Deppenmeier, 2014). Even though genes required for aceticlastic methanogenesis were detected, there are a number of differences between the pathway in the obligate aceticlastic methanogen *Methanosaeta thermophila*, and the hypothetical pathway in MALP that make it doubtful that MALP has the capability to produce methane from acetate.

Notably, MALP lacks the membrane complexes thought to conserve energy for ATP production in *Methanosaeta thermophila*; the membrane subunits of the *mtr* complex (*mtrBCDEFG*) and the membrane-bound heterodisulfide reductase (*HdrDE*). More specifically, the 2 ATP equivalents expended in the conversion of acetate to acetyl-CoA using acetyl-CoA synthetase require a minimum translocation of 7 protons/sodium atoms to regenerate the ATP (at 3 charge translocations/ATP) and build up a potential of 1 net proton/sodium atom per molecule of acetate. *M. thermophila* might achieve this by translocating 2 Na^+ atoms using the *mtr* complex, 2 protons using the membrane-bound heterodisulfide reductase (*HdrDE*), and 3 protons using the Fpo-like complex (Welte and Deppenmeier, 2014). Compared to *M. thermophila*, MALP lacks the sodium translocating subunits of *mtr* (*mtrBCDEFG*) and the integral membrane subunit of the membrane-bound heterodisulfide reductase (*HdrE*). However, MALP does encode an energy conserving pyrophosphatase, and the *hdrD* subunit of membrane bound heterodisulfide reductase has been proposed to interact with the Fpo-like complex, potentially raising the number of protons translocated by the Fpo complex to four (Lang *et al.*, 2015). In addition, MALP encodes an "energy-conserving hydrogenase related" (*ehr*) complex, first observed in *Geobacter sulfurreducens* (Coppi, 2005). The function of this complex is unknown, but it harbors several proton-pumping subunits (Marreiros *et al.*, 2013) and could be involved in energy conservation in MALP. Although these complexes could account for sufficient charge translocation, there is not enough biochemical evidence to confidently predict MALP has the ability to produce methane from acetate.

In addition to a pathway for aceticlastic methanogenesis, the presence of the *mtrAH* genes, combined with the pathway for N5-methyltetrahydrofolate oxidation to formate, could enable MALP to grow using an unconventional type of methanol disproportionation shown in equation 1 and figure 4C.

(1) $3\ CH_3OH \rightarrow HCOO^- + H^+ + 2\ CH_4 + H_2O$ ($\Delta G^{0'} = -203.4$ kJ/mol)

(2) $4\ CH_3OH \rightarrow CO_2 + 3\ CH_4 + 2\ H_2O$ ($\Delta G^{0'} = -319.4$ kJ/mol)

Although thermodynamically feasible (based on ΔG_f^0 from Thauer *et al.*, 1977), there are various caveats to this proposed metabolism. First, it is less energetically favorable than methanol disproportionation to methane and CO_2 (equation 2), suggesting that the metabolism would not be competitive in the environment. However, when expressed per mol substrate (methanol) the energy difference drops to 67.8 kJ/mol methanol versus 79.85 kJ/mol methanol for disproportionation to formate/methane and CO_2/methane respectively. This relatively small difference could be overcome by efficient formate removal by other organisms in the environment.

Another caveat is that the endergonic methyl transfer from Coenzyme M to THMPT in methanol disproportionating methanogens is thought to be driven by dissipation of a sodium gradient. As *mtrAH* in MALP is likely not membrane associated, this mechanism seems unlikely. However, methyl transfer from CoM to tetrahydrofolate (THF) is likely less endergonic than transfer to THMPT (Chistoserdova *et al.*, 1998; Maden, 2000) and might proceed without being driven by a sodium gradient. The remainder of the C1-THF pathway is reversible, albeit less favorable than the C1-THMPT pathway (Maden, 2000). The entire C1-THF pathway is also present in *Methanomassiliicoccus lumiyensis* and *Methanomassiliicoccus intestinalis*, while the other *Methanomassiliicoccales* lack only the gene for conversion between N5-methyltetrahydrofolate and 5-10 methylenetetrahydrofolate (*metF*). This pathway is also proposed to be used in the oxidative direction, to supply intermediates for purine biosynthesis (Lang *et al.*, 2015). None of the *Methanomassiliicoccales*, including MALP, encode formate dehydrogenase (Lang *et al.*, 2015). Oxidation of the methyl group would thus stop at formate,

388 generating 4 electrons and resulting in the stoichiometry shown in equation (1). A final caveat with this
389 proposed pathway is the conversion between NAD(P)H generated in the oxidation of N5-
390 methyltetrahydrofolate to formate, and the ferredoxin that is oxidized at the Fpo-like complex. MALP
391 does not encode a homolog of Ferredoxin:NADP reductase, thus at present the candidate for this
392 reaction is unknown.
393 In summary, the MALP genome, belonging to the 'Lake Pavin' clade of the *Methanomassiliicoccales*,
394 indicates that the MALP organism is a hydrogenotrophic methyl-reducing methanogen, capable of
395 growth on methanol. Unlike the other members of the *Methanomassiliicoccales*, MALP does not encode
396 the genes required for growth on other methylated compounds, such as methylamines or
397 methylsulfides. However, MALP does encode an unusual *mtrAH* complex, which might allow for
398 aceticlastic methanogenesis, as well as methanol disproportionation. However, as outlined above, the
399 latter two predicted metabolic capabilities are highly uncertain without further physiological and
400 biochemical data. Therefore, obtaining a cultured representative of this clade would greatly aid in
401 testing these hypotheses.
402

Concluding remarks

Metagenomic marker gene mining is a complementary approach to genome resolved metagenomics, and can be used to assess phylogenetic diversity and environmental distribution of a microbial process. Due to the size (and continued rapid growth) of public sequence databases our implementation is slow. However, it this approach is broadly applicable as it requires minimal computational power due to the small database size, and minimal storage (in contrast to locally storing a version of the public databases) because datasets are processed sequentially and then deleted. Using this marker gene mining approach, we recovered novel mcrA gene diversity, and identified promising targets for more in-depth analysis leading to better understanding of the habitat distribution and metabolic versatility of the environmental *Methanomassiliicoccales*. Marker gene mining can be used to query the large amount of data from the many publicly available sequencing projects for specific questions, and potentially lead to discoveries outside the scope of the original studies. The rapidly increased throughput, and reduced cost, of next generation sequencing ensures that much more data will become available in years to come, and complementary strategies to analyze this sequencing data effectively will be increasingly important going forward.

Acknowledgements

We thank Woody Fischer and Connor Skennerton for helpful discussion and Grayson Chadwick for critically reading the manuscript. This manuscript is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under award number DE-SC0016469 to VJO. In addition, DRS was supported by NWO Rubicon 019.153LW.039.

Figure Legends

Figure 1. Overview of the marker gene mining workflow

A) Construction of the mcrA reference database using the two mcrA Pfam families and the NCBI non-redundant protein database. B) Screening metagenomes in the sequencing read archive and MG-RAST for the presence of mcrA and reconstructing mcrA gene sequences.

Figure 2. Phylogeny and environmental detection of recovered mcrA sequences

A) Maximum likelihood phylogeny of the translated mcrA sequences. Background shading is used to delineate major clades. Bootstrap values over 70% are indicated by black circles. The two copies of *mcrA* in *Methanococcales* and *Methanobacteriales* are indicated with *mcrA* & *mrtA*. B) Protein accession number or dataset accession number of the sequences in the phylogeny. Sequences obtained from the NCBI-nr are highlighted in orange, sequences assembled in this study from SRA and MG-RAST datasets are highlighted in blue. C) Number of sequences present in the NCBI-nr with over 90% amino acid identity to the sequences in the phylogeny. This includes mostly gene fragments amplified using PCR D) Number of sequencing reads, after BLAST Score Ratio (BSR) filtering, from metagenomes in the SRA and MG-RAST with over 90% amino acid identity to the sequences in the phylogeny.

Figure 3. Phylogeny and environmental distribution of the mcrA sequences within the *Methanomassiliicoccales* order

Maximum likelihood phylogeny of the translated *mcrA* sequences belonging to the *Methanomassiliicoccales* order retrieved in this study, and the five *Methanomassiliicoccales mcrA* sequences present in the NCBI-nr, after dereplication at 90% identity, at the time of database construction. Bootstrap values over 70% are indicated by black circles. Leaf labels are accession numbers

451 of the protein sequence (in the case of the five reference sequences) or source dataset. Coloring of leaf
452 labels indicates source environment. Shading indicates the three clusters discussed in the text: (A) Lake
453 Pavin cluster, (B) Environmental cluster, (C) Gastrointestinal tract (GIT) cluster.
454
455 **Figure 4.** Proposed energy metabolism of the *Methanomassiliicoccales* Lake Pavin (MALP) cluster
456 genome
457 Proposed energy metabolism in the MALP genome. A) Hydrogen dependent reduction of methanol to
458 methane. B) Acetate disproportionation to methane and carbon dioxide C) methanol disproportionation
459 to methane and formate. Substrates of energy metabolism are indicated in red, products in blue.
460 Enzymes are indicated by numbered circles. 1) methanol:coM methyltransferase 2) Methyl-coenzyme M
461 reductase 3) [NiFe]-hydrogenase/heterodisulfide reductase 4) Fpo-like complex/heterodisulfide
462 reductase 5) ATP synthase 6) Acetyl-CoA synthetase 7) Acetyl-CoA synthase 8) N5-
463 tetrahydrofolate:Coenzyme M methyltransferase 9) CO-dehydrogenase 10) Carbonic anhydrase 11)
464 energy conserving pyrophosphatase 12) N5-methyltetrahydrofolate oxidation pathway. CoM: Coenzyme
465 M, CoB: Coenzyme B, Fd: ferredoxin, THF: tetrahydrofolate, THMPT: tetrahydromethanopterin.
466
467

References

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, *et al.* (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19: 455–477.

Becker KW, Elling FJ, Yoshinaga MY, Söllinger A, Urich T, Hinrichs K-U. (2016). Unusual Butane- and Pentanetriol-Based Tetraether Lipids in *Methanomassiliicoccus luminyensis*, a Representative of the Seventh Order of Methanogens *Appl Environ Microbiol* 82: 4505–4516.

Borrel G, Harris HMB, Parisot N, Gaci N, Tottey W, Mihajlovski A, *et al.* (2013a). Genome Sequence of '*Candidatus* Methanomassiliicoccus intestinalis' Issoire-Mx1, a Third *Thermoplasmatales*-Related Methanogenic Archaeon from Human Feces. *Genome Announc* 1 doi:10.1128/genomeA.00453-13.

Borrel G, Harris HMB, Tottey W, Mihajlovski A, Parisot N, Peyretaillade E, *et al.* (2012). Genome Sequence of '*Candidatus* Methanomethylophilus alvus' Mx1201, a Methanogenic Archaeon from the Human Gut Belonging to a Seventh Order of Methanogens. *Journal of bacteriology* 194: 6944–6945.

Borrel G, O'Toole PW, Harris HMB, Peyret P, Brugère J-F, Gribaldo S. (2013b). Phylogenomic data support a seventh order of Methylotrophic methanogens and provide insights into the evolution of Methanogenesis. *Genome Biol Evol* 5: 1769–1780.

Borrel G, Parisot N, Harris HMB, Peyretaillade E, Gaci N, Tottey W, *et al.* (2014). Comparative genomics highlights the unique biology of *Methanomassiliicoccales*, a *Thermoplasmatales*-related seventh order of methanogenic archaea that encodes pyrrolysine *BMC Genomics* 15: 679

Boyd, J. A., Woodcroft, B. J., & Tyson, G. W. (2018). GraftM: a tool for scalable, phylogenetically informed classification of genes within metagenomes. *Nucleic acids res* doi:https://doi.org/10.1093/nar/gky174

Buchfink B, Xie C, Huson DH. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12: 59–60.

Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* 28: 464–469.

Chistoserdova L, Vorholt JA, Thauer RK, Lidstrom ME. (1998) C1 transfer enzymes and coenzymes linking methylotrophic bacteria and methanogenic Archaea. *Science* 281: 99-102

Coppi MV. (2005). The hydrogenases of *Geobacter sulfurreducens*: a comparative genomic perspective. *Microbiology* 151: 1239–1254.

Daims H, Lebedeva EV, Pjevac P, Han P, Herbold C, Albertsen M, *et al.* (2015). Complete nitrification by *Nitrospira* bacteria. *Nature* 528: 504–509.

Delmont TO, Quince C, Shaiber A, Esen OC, Lee STM, Lücker S, *et al.* (2017). Nitrogen-fixing populations of *Planctomycetes* and *Proteobacteria* are abundant in the surface ocean. *bioRxiv*. doi:https://doi.org/10.1101/129791

Denonfoux J, Parisot N, Dugat-Bony E, Biderre-Petit C, Boucher D, Morgavi DP, *et al.* (2013). Gene capture coupled to high-throughput sequencing as a strategy for targeted metagenome exploration. *DNA Res* 20: 185–196.

Dridi B, Fardeau ML, Ollivier B, Raoult D, Drancourt M. (2012). *Methanomassiliicoccus luminyensis* gen. nov., sp. nov., a methanogenic archaeon isolated from human faeces. *Int J Syst Evol Microbiol* 62: 1902–1907.

Edgar RC. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32: 1792–1797

Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460–2461.

532 Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, *et al.* (2015) Anvi'o: an advanced
533 analysis and visualization platform for 'omics data. *PeerJ* 3: e1319.
534 Ermler U, Grabarse W, Shima S, Goubeaud M, Thauer RK. (1997). Crystal structure of methyl-coenzyme
535 M reductase: the key enzyme of biological methane formation. *Science* 278: 1457–1462.
536
537 Evans PN, Parks DH, Chadwick GL, Robbins SJ, Orphan VJ, Golding SD, *et al.* (2015). Methane metabolism
538 in the archaeal phylum *Bathyarchaeota* revealed by genome-centric metagenomics. *Science* 350: 434–
539 438.
540
541 Ferry JG. (1992). Methane from acetate. *Journal of bacteriology* 174: 5489–5495.
542
543 Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, *et al.* (2013). Pfam: the protein
544 families database. *Nucleic Acids Res* 42: D222–D230.
545
546 Großkopf R, Stubner S, Liesack W. (1998). Novel Euryarchaeotal lineages detected on rice roots and in
547 the anoxic bulk soil of flooded rice microcosms. *Appl Env Microbiol* 64: 4983-4989.
548
549 Guo M-Y, Huo D-Q, Ghai R, Rodriguez-Valera F, Shen C-H, Zhang N, *et al.* (2014). Metagenomics of
550 Ancient Fermentation Pits Used for the Production of Chinese Strong-Aroma Liquor. *Genome*
551 *announcements* 2.5 (2014): e01045-14
552
553 Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, *et al.* (2016). A new view of the tree
554 of life *Nature microbiology* 1: 16048.
555
556 Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. (2010). Prodigal: prokaryotic gene
557 recognition and translation initiation site identification. *BMC Bioinformatics* 11: 119.
558
559 Iino T, Tamaki H, Tamazawa S, Ueno Y, Ohkuma M, Suzuki KI, *et al.* (2013) '*Candidatus* Methanogranum
560 caenicola': a novel methanogen from the anaerobic digested sludge, and proposal of
561 *Methanomassiliicoccaceae* fam. nov. and *Methanomassiliicoccales* ord. nov., for a methanogenic lineage
562 of the class *Thermoplasmata*. *Microbes and environments* 28: 244-50.
563

564 Jetten MS, Stams AJ, Zehnder AJ. (1989). Isolation and characterization of acetyl-coenzyme A synthetase
565 from Methanothrix soehngenii. *Journal of bacteriology* 171: 5430–5435.
566
567 Knittel K, Boetius A. (2009). Anaerobic Oxidation of Methane: Progress with an Unknown Process. *Annu*
568 *Rev Microbiol* 63: 311–334.
569
570 Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database Collaboration. (2012).
571 The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res* 40: D54–6.
572
573 Lang K, Schuldes J, Klingl A, Poehlein A, Daniel R, Brune A. (2015). New mode of energy metabolism in
574 the seventh order of methanogens as revealed by comparative genome analysis of '*Candidatus*
575 methanoplasma termitum'. *Appl Environ Microbiol* 81: 1338–1352.
576
577 Laso-Pérez R, Wegener G, Knittel K, Widdel F, Harding KJ, Krukenberg V, *et al.* (2016). Thermophilic
578 archaea activate butane via alkyl-coenzyme M formation. *Nature* 539: 396–401.
579
580 Lau MCY, Cameron C, Magnabosco C, Brown CT, Schilkey F, Grim S, *et al.* (2014). Phylogeny and
581 phylogeography of functional genes shared among seven terrestrial subsurface metagenomes reveal N-
582 cycling and microbial evolutionary relationships. *Frontiers in Microbiology* 5: 531.
583
584 Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ. (2011). Proteinortho: detection of (co-
585)orthologs in large-scale analysis. *BMC Bioinformatics* 12: 124.
586
587 Letunic I, Bork P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of
588 phylogenetic and other trees. *Nucleic Acids Res* 44: W242–5.
589
590 Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. (2015). MEGAHIT: An ultra-fast single-node solution for large
591 and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. 31: 1674-1676.
592
593 Lösekann T, Knittel K, Nadalig T, Fuchs B, Niemann H, Boetius A, *et al.* (2007). Diversity and abundance of
594 aerobic and anaerobic methane oxidizers at the Haakon Mosby Mud Volcano, Barents Sea. *Appl Environ*
595 *Microbiol* 73: 3348–3362.

Lüke C, Speth DR, Kox MAR, Villanueva L, Jetten MSM. (2016). Metagenomic analysis of nitrogen and methane cycling in the Arabian Sea oxygen minimum zone. *PeerJ* 4: e1924.

Maden BE. (2000). Tetrahydrofolate and tetrahydromethanopterin compared: functionally distinct carriers in C1 metabolism. *Biochem J* 350 Pt 3: 609–629.

Marreiros BC, Batista AP, Duarte AMS, Pereira MM. (2013). A missing link between complex I and group 4 membrane-bound [NiFe]-hydrogenases. *Biochim Biophys Acta* 1827: 198–209.

Meyer F, Paarmann D, D'souza M, Olson R, Glass EM, Kubal M, *et al.* (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9: 386.

Mondav R, Woodcroft BJ, Kim E-H, McCalley CK, Hodgkins SB, Crill PM, *et al.* (2014). Discovery of a novel methanogen prevalent in thawing permafrost. *Nat Comms* 5: 3212.

Niemann H, Lösekann T, de Beer D, Elvert M, Nadalig T, Knittel K, *et al.* (2006). Novel microbial communities of the Haakon Mosby mud volcano and their role as a methane sink. *Nature* 443: 854–858.

Parks D, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, *et al.* (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* 2: 1533–1542.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* 25: 1043–1055.

Paul K, Nonoh JO, Mikulski L, Brune A. (2012). '*Methanoplasmatales*,' *Thermoplasmatales*-related archaea in termite guts and other environments, are the seventh order of methanogens. *Appl Env Microbiol* 78: 8245-53.

628 Peng Y, Leung HCM, Yiu SM, Chin FYL. (2012). IDBA-UD: a de novo assembler for single-cell and
629 metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28: 1420–1428.
630
631 Rasko DA, Myers GS, Ravel J. (2005). Visualization of comparative genomic analyses by BLAST score
632 ratio. *BMC Bioinformatics* 6: 2.
633
634 Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD, Perna NT. (2009). Reordering contigs of draft
635 genomes using the Mauve Aligner. *Bioinformatics* 25: 2071–2073.
636
637 Sauer K, Thauer RK. (1999). Methanol:coenzyme-M methyltransferase from *Methanosarcina barkeri* --
638 substitution of the corrinoid harbouring subunit *MtaC* by free cob(I)alamin. *European Journal of*
639 *Biochemistry* 261: 674–681.
640
641 Scheller S, Goenrich M, Boecher R, Thauer RK, Jaun B. (2010). The key nickel enzyme of methanogenesis
642 catalyses the anaerobic oxidation of methane. *Nature* 465: 606–608.
643
644 Seemann T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30: 2068–2069.
645
646 Sharon I, Banfield JF. (2013). Genomes from Metagenomics. *Science* 342: 1057–1058.
647
648 Sorokin DY, Makarova KS, Abbas B, Ferrer M, Golyshin PN, Galinski EA, *et al.* (2017). Discovery of
649 extremely halophilic, methyl-reducing euryarchaea provides insights into the evolutionary origin of
650 methanogenesis. *Nat Microbiol* 2: 17081.
651
652 Söllinger A, Schwab C, Weinmaier T, Loy A, Tveit AT, Schleper C, *et al.* (2016). Phylogenetic and genomic
653 analysis of *Methanomassiliicoccales* in wetlands and animal intestinal tracts reveals clade-specific
654 habitat preferences. *FEMS Microbiology Ecology* 92: fiv149.
655
656 Speth DR, In 't Zandt MH, Guerrero-Cruz S, Dutilh BE, Jetten MSM. (2016). Genome-based microbial
657 ecology of anammox granules in a full-scale wastewater treatment system. *Nat Comms* 7: 11172.
658

659 Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
660 phylogenies. *Bioinformatics* 30: 1312–1313.
661 Tajima K, Nagamine T, Matsui H, Nakamura M, Aminov RI. (2001). Phylogenetic analysis of archaeal 16S
662 rRNA libraries from the rumen suggests the existence of a novel group of archaea not associated with
663 known methanogens. *FEMS Microbiology Letters* 200: 67–72.
664
665 Tan B, Fowler SJ, Abu Laban N, Dong X, Sensen CW, Foght J, *et al.* (2015). Comparative analysis of
666 metagenomes from three methanogenic hydrocarbon-degrading enrichment cultures with 41
667 environmental samples. *The ISME Journal* 9: 2028–2045.
668
669 Thauer RK, Jungermann K, Decker K. (1977). Energy conservation in chemotrophic anaerobic bacteria.
670 *Bacteriol Rev* 41: 809–180.
671
672 Thauer RK, Kaster A-K, Seedorf H, Buckel W, Hedderich R. (2008). Methanogenic archaea: ecologically
673 relevant differences in energy conservation. *Nat Rev Micro* 6: 579–591.
674
675 Thomas T, Gilbert J, Meyer F. (2012). Metagenomics - a guide from sampling to data analysis. *Microb*
676 *Inform Exp* 2: 3.
677
678 van Kessel MAHJ, Speth DR, Albertsen M, Nielsen PH, Op den Camp HJM, Kartal B, *et al.* (2015).
679 Complete nitrification by a single microorganism. *Nature* 528: 555–559.
680
681 Vanwonterghem I, Evans PN, Parks DH, Jensen PD, Woodcroft BJ, Hugenholtz P, *et al.* (2016).
682 Methylotrophic methanogenesis discovered in the archaeal phylum *Verstraetearchaeota*. *Nat Microbiol*
683 1: 16170.
684
685 Wagner T, Ermler U, Shima S. (2016). *MtrA* of the sodium ion pumping methyltransferase binds
686 cobalamin in a unique mode. *Sci Rep* 6: 28226.
687
688 Wagner T, Koch J, Ermler U, Shima S. (2017). Methanogenic heterodisulfide reductase (*HdrABC-*
689 *MvhAGD*) uses two noncubane [4Fe-4S] clusters for reduction. *Science* 357: 699–703.
690

691 Welte C, Deppenmeier U. (2014). Bioenergetics and anaerobic respiratory chains of aceticlastic
692 methanogens. *Biochim Biophys Acta* 1837: 1130–1147.
693 Welte C, Deppenmeier U. (2011). Membrane-bound electron transport in *Methanosaeta thermophila.*
694 *Journal of bacteriology* 193: 2868–2870.
695
696 Wright A-DG, Williams AJ, Winder B, Christophersen CT, Rodgers SL, Smith KD. (2004). Molecular
697 diversity of rumen methanogens from sheep in Western Australia. *Appl Env Microbiol*. 70:1263-70.
698