Whole genome sequencing of a novel, dichloromethane-fermenting *Peptococcaceae* from an enrichment culture

Sophie I Holland [Equal first author, 1], **Richard J Edwards** [Equal first author, 2], **Haluk Ertan** [3, 4], **Yie Kuan Wong** [2], **Tonia L Russell** [5], **Nandan P Deshpande** [2], **Michael Manefield** [1, 4], **Matthew J Lee** [Corresp. 1]

[1] School of Civil and Environmental Engineering, University of New South Wales, Sydney, New South Wales, Australia

[2] School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, New South Wales, Australia

[3] Department of Molecular Biology and Genetics, Istanbul University, Istanbul, Turkey

[4] School of Chemical Engineering, University of New South Wales, Sydney, New South Wales, Australia

[5] Ramaciotti Centre for Genomics, University of New South Wales, Sydney, New South Wales, Australia

Corresponding Author: Matthew J Lee
Email address: mattlee@unsw.edu.au

Bacteria capable of dechlorinating the toxic environmental contaminant dichloromethane (DCM, CH_2Cl_2) are of great interest for potential bioremediation applications. A novel, strictly anaerobic, DCM-fermenting bacterium, "DCMF", was enriched from organochlorine-contaminated groundwater near Botany Bay, Australia. The enrichment culture was maintained in minimal, mineral salt medium amended with dichloromethane as the sole energy source. PacBio whole genome SMRT™ sequencing of DCMF allowed *de novo*, gap-free assembly despite the presence of cohabiting organisms in the culture. Illumina sequencing reads were utilised to correct minor indels. The single, circularised 6.44 Mb chromosome was annotated with the IMG pipeline and contains 5,773 predicted protein-coding genes. Based on 16S rRNA gene and predicted proteome phylogeny, the organism appears to be a novel member of the *Peptococcaceae* family. The DCMF genome is large in comparison to known DCM-fermenting bacteria and includes 96 predicted methylamine methyltransferases, which may provide clues to the basis of its DCM metabolism. Full annotation has been provided in a custom genome browser and search tool, in addition to multiple sequence alignments and phylogenetic trees for every predicted protein, available at http://www.slimsuite.unsw.edu.au/research/dcmf/.

Whole genome sequencing of a novel, dichloromethane-fermenting *Peptococcaceae* from an enrichment culture

Sophie I Holland[a]†, Richard J Edwards[b]†, Haluk Ertan[c,d], Yie Kuan Wong[b], Tonia Russell[e], Nandan Deshpande[b], Michael Manefield[a,d], Matthew Lee[a*]

†These authors contributed equally to this work.

[a]School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW, Australia

[b]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia

[c]Department of Molecular Biology and Genetics, Istanbul University, Istanbul, Turkey

[d]School of Chemical Engineering, University of New South Wales, Sydney, NSW, Australia

[e]The Ramaciotti Centre for Genomics, University of New South Wales, Sydney, NSW, Australia

Running title: Genome sequence of a novel DCM fermenter.

*Corresponding Author:

Matthew Lee

Email address: mattlee@unsw.edu.au

Abstract

Bacteria capable of dechlorinating the toxic environmental contaminant dichloromethane (DCM, CH_2Cl_2) are of great interest for potential bioremediation applications. A novel, strictly anaerobic, DCM-fermenting bacterium, "DCMF", was enriched from organochlorine-contaminated groundwater near Botany Bay, Australia. The enrichment culture was maintained in minimal, mineral salt medium amended with dichloromethane as the sole energy source. PacBio whole genome SMRT™ sequencing of DCMF allowed *de novo*, gap-free assembly despite the presence of cohabiting organisms in the culture. Illumina sequencing reads were utilised to correct minor indels. The single, circularised 6.44 Mb chromosome was annotated with the IMG pipeline and contains 5,773 predicted protein-coding genes. Based on 16S rRNA gene and predicted proteome phylogeny, the organism appears to be a novel member of the *Peptococcaceae* family. The DCMF genome is large in comparison to known DCM-fermenting bacteria and includes 96 predicted methylamine methyltransferases, which may provide clues to the basis of its DCM metabolism. Full annotation has been provided in a custom genome browser and search tool, in addition to multiple sequence alignments and phylogenetic trees for every predicted protein, available at http://www.slimsuite.unsw.edu.au/research/dcmf/.

Introduction

Dichloromethane (DCM, CH_2Cl_2) is a toxic environmental contaminant. Approximately 70% of all DCM worldwide is of anthropogenic origin, due to its extensive use in industry as a solvent and aerosol propellant (Marshall & Pottenger, 2016). It is currently present at 30% of Superfund National Priority List sites within the United States and its territories (U.S. National Library of Medicine, 2019), and global capacity for DCM continues to steadily increase (Marshall & Pottenger, 2016).

DCM in groundwater can be transformed by both aerobic and anaerobic bacteria, although the former is far better characterized (Leisinger & Braus-Stromeyer, 1995). To date, only two DCM-fermenting bacteria have been described and sequenced: *Dehalobacterium formicoaceticum* (Mägli, Wendt & Leisinger, 1996; Chen et al., 2017) and '*Candidatus* Dichloromethanomonas elyunquensis' (Kleindienst et al., 2016, 2017). Of these, only the former has been isolated (Mägli, Wendt & Leisinger, 1996). Both species are thought to metabolise DCM via incorporation of the methyl group into the Wood-Ljungdahl pathway, although the precise mechanism of dechlorination has thus far eluded description.

Here, we report the whole genome sequencing and assembly of a novel, DCM-fermenting bacterium, herein referred to as DCMF. The organism exists in an enrichment culture ("DFE", DCM-fermenting enrichment) derived from an organochlorine-contaminated sand bed aquifer adjacent to Botany Bay, an oceanic embayment 13 km south of Sydney, Australia (Lee et al., 2012).

Materials & Methods

Inoculum origin

The original inoculum was obtained from sediment drilled from 5 m beneath the surface of an organochlorine-contaminated coastal sand bed aquifer (Botany Sands aquifer), latitude -33°57'27.6"S, longitude 151°12'60.0"E. The initial, methanogenic enrichment culture using DCM as the sole energy source was reported previously (Lee et al., 2012).

Culture media

Cultures were grown in anaerobic minimal mineral salts medium that comprised (g l^{-1}): $CaCl_2.2H_2O$ (0.1), KCl (0.1), $MgCl_2.6H_2O$ (0.1), $NaHCO_3$ (2.5), NH_4Cl (1.5), NaH_2PO_4 (0.6), 1 ml of trace element solution A (1000×), 1 ml of trace element solution B (1000×), 1 ml of vitamin solution (1000×), 10 ml of 5 g l^{-1} fermented yeast extract (FYE; 100×), and resazurin 0.25 mg l^{-1}. Trace element solutions A and B were prepared as described previously (Wolin, Wolin & Wolfe, 1963), as was the vitamin solution (Adrian et al., 1998). Medium was sparged with N_2 during preparation and the pH was adjusted to 6.8 – 7.0 by a final purge with N_2/CO_2 (4:1). Aliquots were dispensed into glass serum bottles that were crimp sealed with Teflon faced rubber septa (13 mm diameter, Wheaton) before the medium was chemically reduced with sodium sulphide (0.2 mM). DCM (1 mM) was supplied as the sole electron source via a glass syringe. Methanogenic Archaea present in the early enrichment culture were inhibited with 2-bromoethanosulfonate (BES, 0.2 mM) for two generations. All cultures were incubated statically at 30˚C in the dark.

Preparation of spent media as a co-factor solution

A stock FYE solution was prepared by inoculating anoxic yeast extract (5 g l^{-1}) in defined minimal mineral salts medium (described above, excluding DCM) with the DCM-fermenting enrichment (DFE) culture. The culture was incubated for one week at 30˚C before being filter-sterilised. The filtered, spent media was re-inoculated with DFE and incubated for a further week, to ensure that growth was no longer possible on FYE (i.e. that it had been energetically exhausted). The spent media was then filter-sterilised again before use.

Analytical methods

DCM and methane were quantified on a GS-Q column (30 m × 0.32 mm; Agilent Technologies) using a Shimadzu GC-2010 gas chromatograph with flame ionisation detector (GC-FID). Headspace samples (100 µl) were withdrawn directly from culture flasks using a lockable, gas-tight syringe and injected manually. The oven was initially 150˚C, then raised by 30˚C min^{-1} to 250˚C. The inlet temperature was 250˚C, split ratio 1:10, FID temperature 250˚C. A minimum three-point calibration curve was used. DCM concentrations are reported as the nominal concentration in each serum bottle, calculated from the headspace concentration using the Henry's Law dimensionless solubility constant (H^{cc} = 0.107 at 30˚C), as per the OSWER method (US EPA, 2001).

Genomic DNA extraction

Genomic DNA was extracted as previously described (Urakawa, Martens-Habbena & Stahl, 2010). Briefly, cells were lysed with lysis buffer and bead-beating, before DNA was extracted

with phenol-chloroform-isoamyl, precipitated using isopropanol, and resuspended in molecular grade water. The nucleic acid concentration was quantified using a Qubit instrument and assay as per the manufacturer's instructions (Life Technologies).

Community analysis

Throughout the initial transfers and serial dilutions of the enrichment culture, the community was monitored via denaturing gradient gel electrophoresis (DGGE). DNA was amplified with primers GC338F and 530R (Table S1). DGGE was performed with a DCode mutation detection system (Bio-Rad) and a Cipher Electrophoresis system (CBS Scientific Company Inc) in a 1× TAE buffer at pH 7.5. PCR products were loaded onto a 10% (v/v) acrylamide gel with a 30 – 60% gradient of urea and deionised formamide before electrophoresis at 60˚C, 75V for 16.5 h. Gels were stained with SYBR Gold (Invitrogen™, Life Technologies) in 1× TAE buffer for 10 min, prior to visualisation on a Gel Doc XR (Bio-Rad). Bands of interest were excised, DNA eluted from them in molecular grade water and re-amplified using the 338F primer (Table S1). PCR products were cleaned with a Clean and Concentrate-25 kit (Zymo Research).

To confirm the absence of an archaeal population following amendment of the enrichment culture with BES, archaeal specific primers Arc340F and Arc1000R (Table S1) were used for PCR on a T100™ thermal cycler (Bio-Rad).

Quantitative PCR of the *Dehalobacter* spp. 16S rRNA gene was carried out on a CFX96 thermal cycler (Bio-Rad, Table S1). Standards ranged from 10^3 – 10^9 copies ml^{-1} and were created using serial 10-fold dilutions of a plasmid carrying the cloned gene, constructed with TOPO TA Cloning Kit (Life Technologies).

Illumina genome sequencing

DNA was prepared with the Nextera XT library prep kit (Illumina). Sequencing was carried out on an Illumina MiSeq with a v2 500-cycle kit (2 × 250 bp run) at the Ramaciotti Centre for Genomics (UNSW Sydney, Australia). Three MS110-2 libraries were used for the run. Library size ranged from 200 - 3000 bp, with an average of 955 bp. Raw reads were trimmed and filtered with SolexaQA (DynamicTrim.pl and LengthSort.pl) (Cox, Peterson & Biggs, 2010). Raw reads were submitted to the NCBI Sequence Read Archive with the identifier SRR5179547.

Pacific Biosciences SMRT sequencing

A MagAttract HMW DNA kit (Qiagen) was used to extract high-molecular weight genomic DNA, followed by purification using AMPure PB beads (Beckman Coulter). DNA concentration and purity were checked by Qubit and NanoDrop instruments, respectively. A 0.75% Pippin Pulse gel (Sage Science) was performed by the Ramaciotti Centre for Genomics (UNSW Sydney, Australia) to further verify integrity. A SMRTbell library was prepared with the PacBio 20 kb template protocol excluding shearing (Pacific BioSciences). Additional damage repair was carried out following minimum 4 kb size selection using Sage Science BluePippin. Whole genome sequencing was performed on the PacBio RS II (Pacific Biosciences), employing P6 C4 chemistry with 240 min movie lengths. DNA was initially sequenced using two Single Molecule Real Time™ (SMRT) cells. A third SMRT™ cell was added to compensate for low quality data

from the first two, due to degraded DNA yield from the sample. The SMRTbell library for this cell was prepared with the PacBio 10 kb template protocol, without size selection, and a lower input (3,624 ng) of DNA was used. In total, the three SMRT cells yielded 463,878 subreads from 169,180 ZMW, with a combined length of 1,712,588,985 bp. Reads were submitted to the NCBI Sequence Read Archive with the identifier SRR5179548.

Genome assembly and annotation

PacBio subreads were assembled using HGAP3 (Chin et al., 2013) as implemented in SMRT Portal. In-house software, SMRTSCAPE (SMRT Subread Coverage & Assembly Parameter Estimator; http://rest.slimsuite.unsw.edu.au/smrtscape) was used to predict optimal HGAP settings for several different assemblies with different predicted genome size and minimum correction depths (Table S2). The assembly with the greatest depth of coverage used for seed read error correction that still yielded a full-length (6.44 Mb) intact chromosome was selected for the draft genome. This corresponded to: min read length 4,010 bp; min seed read length 8,003 bp; min read quality 0.86; min 10× correction coverage. The genome was corrected with Quiver (Chin et al., 2013) using all subreads and circularised by identifying and trimming overlapping ends, then annotated in-house using Prokka (Seemann, 2014). Based on draft annotation, the genome was re-circularised to have its break-point in the intergenic region between the 3' of two hypothetical genes, and then subjected to a second round of Quiver correction to make sure the manually joined region was of high quality. Filtered Illumina reads were mapped onto the Quiver-corrected genome using BWA-MEM v0.7.9a (Li, 2013) and possible errors were identified with Pilon (Walker et al., 2014). Manual curation was then performed to check any discrepancies between the PacBio and Illumina data and correct small indels. Raw PacBio reads were mapped onto the completed genome with BLASR (Chaisson & Tesler, 2012). The corrected genome was re-annotated with Prokka and uploaded to the Integrated Microbial Genomes and Microbiomes (IMG/M) system of the Joint Genome Institute (JGI) for independent annotation (Chen et al., 2019). Twenty-eight fragmented pairs of genes were subject to additional manual curation and correction where a pyrrolysine or selenocysteine residue had been erroneously translated as a stop codon. The genome has subsequently been re-annotated by NCBI.

16S rRNA gene phylogeny of the novel organism

The DCMF 16S rRNA gene consensus sequence was searched against the NCBI prokaryotic 16S rRNA BLAST database as well as the 16S rRNA gene sequences of the two other known DCM-fermenting bacteria (absent from that database), *D. formicoaceticum* strain DMC (NCBI locus tags CEQ75_RS05455, CEQ75_RS05490, CEQ75_RS13675, CEQ75_RS13970, CEQ75_RS17045) and '*Ca.* Dichloromethanomonas elyunquensis' strain RM (KU341776.1). The closest phylogenetic relatives and an outgroup, *Moorella perchloratireducens* strain An10 (NR_125518.1), were aligned and a tree constructed using the neighbour-joining method with 1000 bootstrap resampling a 200PAM/k = 2 scoring matrix using 1,365 nucleotides. This was performed with MAFFT program v.7 (Kuraku et al., 2013) using the Archaeopteryx tool (Han & Zmasek, 2009), as well as manual curation.

High throughput phylogenetic analysis of predicted proteome

JGI-annotated proteins were further annotated via high-throughput homology searching, multiple sequence alignment and molecular phylogenetics using HAQESAC v1.10.2 (Edwards et al., 2007). BLAST+ v2.6.0 blastp (Camacho et al., 2009) was used to search each protein against three protein datasets: (1) all bacterial proteins in the UniProt Knowledgebase (The UniProt Consortium, 2017) (downloaded 2017-02-06); (2) the predicted DCMF proteome; (3) the nine NCBI proteomes available for closely related bacteria identified from 16S rRNA gene analysis: *D. formicoaceticum* (GCF_002224645.1), *Desulfosporosinus acididurans* (GCF_001029285.1), *Desulfosporosinus acidiphilus* (GCF_000255115.2), *Desulfosporosinus orientis* (GCF_000235605.1), *Desulfosporosinus hippei* (GCF_900100785.1), *Desulfosporosinus lacus* (GCF_900129935.1), *Desulfitobacterium metallireducens* (GCF_000231405.2), *Desulfitobacterium hafniense* (GCF_000021925.1), *Dehalobacter restrictus* (GCF_000512895.1). The top 50 blastp results for each dataset were combined and up to 60 homologues meeting the HAQESAC default filtering criteria were aligned with Clustal Omega v1.2.2 (Sievers & Higgins, 2017). Neighbour-joining phylogenetic trees (1000 bootstraps) were inferred using ClustalW v2.1 and midpoint-rooted using HAQESAC. Paralogous subfamilies arising from gene duplications were identified as nodes where the two ancestral clades each had at least two different species and shared at least one of those species. Multiple sequences from the same species within one of these paralogous subfamilies were identified as “in-paralogues” (lineage-specific duplications) or possible sequence variants. DCMF in-paralogues were kept. Possible in-paralogues or sequence variants from other species were restricted to the single closest homologue to the DCMF query. NCBI annotated proteins were subsequently subjected to the same pipeline with the addition of the JGI predicted proteome to the search database.

Putative taxonomic assignments for each JGI protein were made using an in-house tool, TaxaMap (http://rest.slimsuite.unsw.edu.au/taxamap). TaxaMap identifies the smallest clade to which the query DCMF protein can be confidently assigned by stepping ancestrally through the tree until it reaches a branch with a bootstrap support of at least 50% and at least one non-DCMF protein. If the root is reached without meeting these requirements, the full HAQESAC tree was used. Once the clade has been identified, all Uniprot species codes for that clade are extracted as putative taxonomic assignments. These are mapped onto parent species, genus, family, order, class and phylum classifications using UniProt Knowledgebase taxonomy. At each taxonomic level, the taxa list is reduced to be non-redundant and each taxon contributes equally, to reduce sampling biases. Where a species code could only be mapped to a higher taxonomic level, it was designated as an unknown taxon associated with that higher level, e.g. “Firmicutes fam.” would indicate an unknown family within the phylum Firmicutes. Where no non-DCMF homologues were found, a protein was assigned “None”. TaxaMap Assignments were made for each protein individually and then combined using two strategies: (1) Unweighted; (2) Bootstrap weighted. The unweighted assignment simply adds up the number of proteins assigned to a particular taxon. Where a protein is assigned to multiple taxa, each is given an equal proportion of that protein, e.g. if a protein mapped ambiguously to five taxa, each would receive 0.2 for that

protein. Any taxa with a combined score below 1.0 across all proteins was excluded, and scores recalculated iteratively. For the weighted score, counts were multiplied by the percentage bootstrap support for the clade, e.g. if a protein was assigned to two taxa and the bootstrap support for the clade was 80%, each taxon would receive a score of 0.4 (= 0.5 x 0.8).

Genomic analysis

CheckM (Parks et al., 2015) was used to assess the completeness and contamination in the DCMF genome. SPADE (Mori et al., 2019) was used to analyse repeat regions in the genomes, using default parameters.

The 82 full-length predicted trimethylamine (TMA) methyltransferase protein sequences were compared in a pairwise percentage distance matrix, calculated using GABLAM version 2.28.2 (Davey, Shields & Edwards, 2006). BLAST 2.5.0+ blastp (Camacho et al., 2009) results were converted into the minimum global percentage difference between each pair of proteins. This distance matrix was converted into a heatmap using the heatmap.2() function of gplots (https://CRAN.R-project.org/package=gplots) in R 3.4.0 (The R Core Team, 2013).

Results

Enrichment of DCMF in DFE

Five 1% transfers (T1 – T5) of the previously reported (Lee et al., 2012) enrichment culture DCMD were carried out. The initial three transfers produced methane in a molar ratio of 0.6 moles per mole of DCM (Figure 1A). Addition of BES to the culture medium in T4 caused methanogenesis to cease, and T5 could utilise DCM without the generation of methane in the absence of BES (Figure 1B). The absence of methanogenic populations was confirmed via archaeal specific PCR. While a clear band at ~660 bp was observed in a positive control and T3 culture, there was no archaeal PCR product from the enrichment culture after the addition and subsequent removal of BES. The non-methanogenic, DCM-fermenting enrichment culture was henceforth called DFE.

T5 was then subject to two rounds of dilution to extinction. Community diversity was monitored throughout these transfers by DGGE, which showed a trend towards purity, culminating in the presence of a single band from the lowest active dilution series culture (10^{-3}; Figure S1). Sequencing of the primary band had the highest identity match to an uncultured *Peptococcaceae*, henceforth referred to as “DCMF”.

The shift away from the *Dehalobacter* population originally shown to be linked to DCM-degradation (Lee et al., 2012), was confirmed with qPCR. The *Dehalobacter* sp. 16S rRNA gene was below the limit of detection (1.45×10^3 copies ml^{-1}) at all stages of growth in DFE cultures after the removal of methanogenic populations.

Genome assembly and annotation

Attempts were initially made to sequence the dominant, DCM-degrading organism using Illumina short read technology, which yielded 5,040,903 filtered read pairs for a total of

1,827,383,271 bp. However, the presence of the additional organisms in the DFE culture and lack of a reference genome hindered this approach. Instead, a pure PacBio long read strategy was used to assemble a full-length gap-free circular genome for DCMF (GenBank accession CP017634.1). Trimmed and filtered Illumina reads (average 242× coverage) were used for final, minor error correction. The final genome assembly had an average of 132× PacBio coverage (min >50×) and no regions of unusual read depth (Figure 2A). The genome was circularised at overlapping ends and every base was covered by long reads spanning at least 5 kb 5' and 3' (Figure 2B). In addition to these assessments, CheckM evaluated the genome as 97% complete with a contamination rate of 2%.

The DCMF genome is 6,441,270 bp long and has a G+C content of 46.44%. JGI annotation initially revealed 5,801 predicted protein-coding genes. Manual curation of the 28 pairs of genes fragmented by the presence of the amino acids pyrrolysine and selenocysteine (encoded by in-frame UAG and UGA stop codons, respectively; Table S3) brought this total down to 5,773 protein coding genes.

PacBio sequencing confirmed the presence of a number of contaminant bacteria remaining in the enrichment culture via identification of 16S rRNA genes. This included species within (or related to) the genera *Desulfovibrio*, *Ignavibacterium*, *Treponema*, and *Thermovirga* (Table S4).

16S rRNA gene phylogeny

The DCMF genome contains four full-length 16S rRNA genes (NCBI locus tags DCMF_03210, DCMF_03275, DCMF_18375, DCMF_21985; Table S4), which share 99.87% identity when aligned. Based on the consensus 16S rRNA gene sequence, the closest relative to DCMF is *D. formicoaceticum* strain DMC (94% identity). This is closely followed by '*Ca.* Dichloromethanomonas elyunquensis' strain RM, *Dehalobacter restrictus* strain PER-K23 and *Desulfosporosinus acidiphilus* strain SJ4 (all 89% identity), and *Desulfitobacterium dehalogenans* strain ATCC 51507 (88% identity) (Figure 3).

Phylogenetic analysis of the predicted proteome

Taxonomic analysis of the whole predicted DCMF proteome was inconclusive at the genus level but strongly supported assignment within the order *Clostridiales* (Figure 4). The top-ranked genus was *Dehalobacterium* (25.7% proteins, bootstrap-weighted), supporting the 16S rRNA gene phylogeny (Figure 3) with *D. formicoaceticum* as the closest known relative of DCMF. The top families were *Peptococcaceae* (39.3%) and *Clostridiaceae* (11.2%). Whole-proteome TaxaMap analysis provides a good overview but is clearly influenced by the availability of homologous sequences in the search databases and may also be disrupted by, for example, horizontal gene transfer. We therefore restricted analysis to a more robust set of eight house-keeping genes and 47 ribosomal proteins (Table S5). With the exception of one malate dehydrogenase (Ga0180325_112460) and SSU ribosomal proteins S10P (Ga0180325_114571), all proteins support *D. formicoaceticum* as the closest known relative of DCMF and placement in the *Peptococcaceae* family. All 55 genes support placement in *Clostridiales* (Table S5). Multiple sequence alignments, phylogenetic trees and TaxaMap assignments for all proteins can be found

in online supplementary material at: http://www.slimsuite.unsw.edu.au/research/dcmf/. The restricted housekeeping genes can be found at: http://www.slimsuite.unsw.edu.au/research/dcmf/dcmf-hk.php.

Genomic features of DCMF

A number of metabolic pathways were identified in the DCMF genome (Table 1). The most prominent of these is the full set of genes for the Wood-Ljungdahl pathway (Table S6). No reductive dehalogenases were identified in the genome. Additionally, numerous sets of glycine/sarcosine/betaine reductase complex genes were found (Table S7), indicating that DCMF may have a wider metabolic repertoire than close relatives.

The DCMF genome also contains an abundance of methylamine methyltransferase genes (Table S8), including 82 copies of TMA methyltransferase, *mttB*. There is a high diversity amongst the *mttB* genes, with an average amino acid sequence difference of 69.70% (Figure 5). Associated with the presence of these methyltransferases are five genes necessary to synthesise and utilise pyrrolysine (*pylTSBCD*; Table S9), a non-canonical amino acid residue present in 23 of the 96 total methylamine methyltransferases in the genome.

The presence of all genes required for *de novo* corrinoid biosynthesis (Table S10) is pertinent both to certain Wood-Ljungdahl pathway proteins and the methylamine methyltransferases, which typically require a corrinoid cofactor to function. However, the genes for methionine synthesis (*metH* and *metE*), an important precursor for corrin ring formation, were not identified in the genome. DCMF may be using an alternative route for *de novo* biosynthesis of this amino acid.

Discussion

The shift from a *Dehalobacter* species to DCMF

The novel *Peptococcaceae*, DCMF, was enriched from a previously reported methanogenic consortium, DCMD, where DCM was supplied as the sole energy source (Lee et al., 2012). That consortium was dominated by a *Dehalobacter* species whose growth was linked to DCM metabolism, producing acetate and methane. The Archaeal population was dominated by a hydrogenotrophic methanogen from the genus *Methanoculleus*. Furthermore, *Dehalobacter* sp. growth could be inhibited by addition of excess hydrogen. These two phenomena led to the conclusion that hydrogen was a DCM fermentation product along with acetate, and that a syntrophic association existed between *Dehalobacter* and *Methanoculleus*. In the present study, inhibition of methanogens with BES enabled the hitherto unknown non-hydrogenogenic DCMF to become the dominant DCM fermenter in the enrichment culture DFE.

Amongst the culture contaminants, some genera (*Desulfovibrio*, *Treponema*, *Thermovirga*) are consistent with those previously identified in DCMD, while others (*Ignavibacterium*) appear to have only risen above the quantifiable abundance threshold since the previous community analysis was carried out (Lee et al., 2012). These cohabiting bacteria have persisted despite attempts to isolate DCMF. These have been limited to serial transfers of dilution to extinction,

due to the inability of the organism to form colonies on agar plates or in semi-solid agar shakes. Nonetheless, this has lead to a highly enriched culture, with community fingerprinting results showing only a single lineage.

Optimisation for a high quality genome assembly from a mixed culture

Based on the 16S rRNA gene sequence retrieved from the DGGE community analysis, DCMF appeared to be an organism with comparatively few cultured relatives. Thus, whole genome sequencing was carried out in order to learn more about its role and function in the enrichment community. The lack of a reference genome and other organisms in the enrichment culture hindered attempts to assemble the genome from short read sequences only, making the long read capability of PacBio sequencing indispensible for this effort. Although long reads are prone to a higher proportion of sequencing errors than short reads, a series of checks were put in place to ensure that a high quality, uncontaminated genome assembly was obtained.

The use of SMRTSCAPE to predict the optimal HGAP settings allowed rapid comparison of various assembly parameters. By increasing the minimum correction coverage from 6× to 10×, the total size of the assembly (including contaminant organism DNA) decreased from ~16 Mb to ~8.8 Mb, while the size of the DCMF genome remained relatively stable around 6.4 Mb. Increasing the minimum correction coverage one step further to 11× resulted in a significant reduction of the DCMF genome to 1.9 Mb, indicating that much of the assembly was likely being lost to overzealous correction (Table S2).

The large size of the DCMF genome distinguishes it from the two other known DCM-fermenting bacteria, *D. formicocaceticum* and "*Ca.* Dichloromethanomonas elyunquensis" (Table 1). When assembling a genome *de novo* from a mixed culture, there is always the concern that stretches of other contaminating genomes will be mis-incorporated into the assembly. This likelihood was reduced by our assembly strategy of increasing stringency. The consistent sequencing coverage across the final genome (Figure 2) strongly indicates that there was no such mis-assembly. The CheckM contaminant rate of 2% further confirms that the large DCMF genome is not over-inflated due to contamination. Analysis of repeated sequence motifs with SPADE showed that they comprise just 21,395 bp (0.03%) of the total DCMF genome, which also rules this out as a source of the large genome size. Annotation predicted 5,773 protein coding genes, giving a gene density of approximately 0.9 genes per kilobase, which is consistent with normal bacterial gene density (Koonin & Wolf, 2008).

Genome annotation quality and availability of data

Despite the numerous error limiting and quality control steps taken in this study, it is almost certain that some errors will remain in both the genome sequence and genome annotation. We have therefore provided rich supplementary data to enable rapid, detailed analysis of potential genes and proteins of interest. The DCMF genome is available for browsing via a public Web Apollo (Lee et al., 2013) genome browser, accessed via the supplementary data site: http://www.slimsuite.unsw.edu.au/research/dcmf/. Results of three annotation pipelines (Prokka, JGI and NCBI) are available through the browser for direct comparison, along with mapped

PacBio reads for assessing genomic sequence quality. A search tool has also been provided, enabling Exonerate (Slater & Birney, 2005) or BLAST+ (Camacho et al., 2009) searches of cDNA, peptides or genomic DNA against the DCMF genome, with hits linking directly to the corresponding region of the Web Apollo genome browser. Furthermore, multiple sequence alignments and phylogenetic trees have been provided for every JGI- and NCBI- annotated protein, enabling rapid assessment of protein descriptions and completeness.

An abundance of methyltransferases may indicate key role in metabolism

While DCMF, *D. formicoaceticum*, and '*Ca.* Dichloromethanomonas elyunquensis' have thus far only been cultured on DCM as sole energy source, the larger genome of DCMF suggests that perhaps it is capable of other metabolisms. One standout feature is the vast abundance of predicted methyltransferases. The genome harbours 96 assorted methylamine methyltransferase genes, of which 81 are annotated as a component of a TMA methyltransferase. This hints that TMA may also be utilised as a substrate by DCMF. Additionally, the presence of numerous glycine/betaine/sarcosine reductases may allow the organism to utilise these related compounds as well. These reductase genes are also present in *D. formicoaceticum*, but absent from "*Ca.* Dichloromethanomonas elyunquensis" (Table 1).

Of the 96 methylamine methyltransferase genes, 23 contain a pyrrolysine residue, identifiable as an in-frame UAG (amber) stop codon. While the TMA methyltransferase (*mttB*) gene is widespread amongst bacteria and archaea, most organisms do not encode the pyrrolysine residue (Srinivasan, 2002; Ticak et al., 2014). Indeed, the *pylTSBCD* gene cluster to synthesise and incorporate this non-canonical amino acid is limited to only six bacterial genera, including *Desulfotomaculum, Desulfitobacterium*, and *Thermincola* (Gaston, Jiang & Krzycki, 2011) – all members of the *Peptococcaceae* family and close relatives of DCMF based on 16S rRNA phylogeny. *D. formicoaceticum* also encodes the *pyl* genes, but "*Ca.* Dichloromethanomonas elyunquensis" does not (Table 1).

Curiously, there is high diversity amongst the TMA methyltransferases in DCMF, with an average amino acid sequence diversity of 69.7% (Figure 5). This may indicate that these genes have more than one function within the cell and/or have diversified to accommodate cobalamin cofactors with various upper and lower ligands. It has previously been shown that the chloromethane dehalogenase CmuAB is functionally similar to the monomethylamine methyltransferase MtaA (Studer et al., 2001). Moreover, four corrinoid-dependent methyltransferases were highly expressed in the proteome of DCM-fermenting '*Ca.* Dichloromethanomonas elyunquensis'(Kleindienst et al., 2019), further indicating that the array of methyltransferases in DCMF, along with its complete corrinoid biosynthetic pathway, may be crucial to the metabolism of DCM.

Notably, however, '*Ca.* Dichloromethanomonas elyunquensis' also encodes reductive dehalogenase genes in its genome, while DCMF and *D. formicocaceticum* do not (Table 1). This finding, coupled with a recent dual carbon-chlorine isotopic analysis of the two previously-reported DCM-fermenters (Chen et al., 2018), suggests that there are distinct DCM

dechlorination mechanisms operating in these organisms. Based on the presence or absence of key pathways in the genome (Table 1) and phylogenetic analysis (Figure 3), DCMF appears to have more in common with *D. formicoaceticum* than '*Ca.* Dichloromethanomonas elyunquensis'.

Conclusions

DCMF is an organism that demonstrates a relatively rare metabolism and harbours a large genome. Both long and short read genome sequencing technology were used to compliment each other and assemble a singular, circular chromosome for the organism, despite the low-level presence of other bacteria in the enrichment culture. DCMF is the dominant organism in the enrichment and likely sits within the *Peptococcaceae* family, although not within any known genus. Its DCM-fermenting capabilities make it of interest to the bioremediation sector and the genome contains clues to the as-yet undiscovered DCM dechlorinating enzyme, the identification of which will be the subject of future work. Extensive supplementary data for the DCMF genome and annotation is available at http://www.slimsuite.unsw.edu.au/research/dcmf/.

Acknowledgements

We thank Dr Bat-Erdene Jugder (University of New South Wales) for his assistance with the DNA extractions for PacBio sequencing and Dr Xabier Vázquez-Campos (University of New South Wales) for assistance with data retrieval.

References

Adrian L, Manz W, Szewzyk U, Görisch H. 1998. Physiological characterization of a bacterial consortium reductively dechlorinating 1,2,3- and 1,2,4-trichlorobenzene. *Applied and Environmental Microbiology* 64:496–503.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:1. DOI: Artn 421\nDoi 10.1186/1471-2105-10-421.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* 13:238. DOI: 10.1186/1471-2105-13-238.

Chen IMA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, Huntemann M, Varghese N, White JR, Seshadri R, Smirnova T, Kirton E, Jungbluth SP, Woyke T, Eloe-Fadrosh EA, Ivanova NN, Kyrpides NC. 2019. IMG/M v.5.0: An integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Research* 47:D666–D677. DOI: 10.1093/nar/gky901.

Chen G, Murdoch RW, Mack EE, Seger ES, Löffler FE. 2017. Complete genome sequence of *Dehalobacterium formicoaceticum* strain DMC, a strictly anaerobic dichloromethane-degrading bacterium. *Genome Announcements* 5:18–19. DOI: https://doi .org/10.1128/genomeA.00897-17.

Chen G, Shouakar-Stash O, Phillips E, Justicia-Leon SD, Gilevska T, Sherwood Lollar B, Mack

448 EE, Seger ES, Löffler FE. 2018. Dual carbon-chlorine isotope analysis indicates distinct
449 anaerobic dichloromethane degradation pathways in two members of *Peptococcaceae*.
450 *Environmental Science & Technology* 52:8607–8616. DOI: 10.1021/acs.est.8b01583.

451 Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A,
452 Huddleston J, Eichler EE, Turner SW, Korlach J. 2013. Nonhybrid, finished microbial
453 genome assemblies from long-read SMRT sequencing data. *Nature Methods* 10:563–569.
454 DOI: 10.1038/nmeth.2474.

455 Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: At-a-glance quality assessment of Illumina
456 second-generation sequencing data. *BMC Bioinformatics* 11:485.

457 Davey NE, Shields DC, Edwards RJ. 2006. SLiMDisc: Short, linear motif discovery, correcting
458 for common evolutionary descent. *Nucleic Acids Research* 34:3546–3554. DOI:
459 10.1093/nar/gkl486.

460 Edwards RJ, Moran N, Devocelle M, Kiernan A, Meade G, Signac W, Foy M, Park SDE, Dunne
461 E, Kenny D, Shields DC. 2007. Bioinformatic discovery of novel bioactive peptides. *Nature*
462 *Chemical Biology* 3:108–112. DOI: 10.1038/nchembio854.

463 Gaston MA, Jiang R, Krzycki JA. 2011. Functional context, biosynthesis, and genetic encoding
464 of pyrrolysine. *Current Opinion in Microbiology* 14:342–349. DOI:
465 10.1016/j.mib.2011.04.001.

466 Han M, Zmasek C. 2009. phyloXML: XML for evolutionary biology and comparative genomics.
467 *BMC Bioinformatics* 10:356. DOI: 10.1186/1471-2105-10-356.

468 Kleindienst S, Chourey K, Chen G, Murdoch RW, Higgins SA, Iyer R, Campagna SR, Mack EE,
469 Seger ES, Hettich RL, Löffler FE. 2019. Proteogenomics reveals novel reductive
470 dehalogenases and methyltrasnferases expressed during anaerobic dichloromethane
471 metabolism. *Applied and Environmental Microbiology* 85:1–16. DOI: 10.1128/aem.02768-
472 18.

473 Kleindienst S, Higgins SA, Tsementzi D, Chen G, Konstantinidis KT, Mack EE, Löffler FE.
474 2017. '*Candidatus* Dichloromethanomonas elyunquensis' gen. nov., sp. nov., a
475 dichloromethane-degrading anaerobe of the Peptococcaceae family. *Systematic and Applied*
476 *Microbiology* 40:150–159. DOI: 10.1016/j.syapm.2016.12.001.

477 Kleindienst S, Higgins SA, Tsementzi D, Konstantinidis KT, Mack EE, Löffler FE. 2016. Draft
478 genome sequence of a strictly anaerobic dichloromethane-degrading bacterium. *Genome*
479 *Announcements* 4:e00037-16. DOI: 10.1128/genomeA.00037-16.

480 Koonin E V, Wolf YI. 2008. Genomics of bacteria and archaea: The emerging dynamic view of
481 the prokaryotic world. *Nucleic Acids Research* 36:6688–6719. DOI: 10.1093/nar/gkn668.

482 Kuraku S, Zmasek CM, Nishimura O, Katoh K. 2013. aLeaves facilitates on-demand exploration
483 of metazoan gene family trees on MAFFT sequence alignment server with enhanced
484 interactivity. *Nucleic Acids Research* 41:W22–W28. DOI: 10.1093/nar/gkt389.

485 Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH,
486 Elsik CG, Lewis SE. 2013. Web Apollo: a web-based genomic annotation editing platform.
487 *Genome Biology* 14:R93. DOI: 10.1186/gb-2013-14-8-r93.

488 Lee M, Low A, Zemb O, Koenig J, Michaelsen A, Manefield M. 2012. Complete chloroform

dechlorination by organochlorine respiration and fermentation. *Environmental Microbiology* 14:883–894. DOI: 10.1111/j.1462-2920.2011.02656.x.

Leisinger T, Braus-Stromeyer SA. 1995. Bacterial growth with chlorinated methanes. *Environmental Health Perspectives* 103:33–36.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *BMC Bioinformatics* 11:485.

Mägli A, Wendt M, Leisinger T. 1996. Isolation and characterization of *Dehalobacterium formicoaceticum* gen. nov. sp. nov., a strictly anaerobic bacterium utilizing dichloromethane as source of carbon and energy. *Archives of Microbiology* 166:101–108. DOI: 10.1007/s002030050362.

Marshall KA, Pottenger LH. 2016. Chlorocarbons and Chlorohydrocarbons. In: *Kirk-Othmer Encyclopedia of Chemical Technology*. John Wiley & Sons, Inc,. DOI: 10.1002/0471238961.1921182218050504.a01.pub3.

Mori H, Evans-Yamamoto D, Ishiguro S, Tomita M, Yachie N. 2019. Fast and global detection of periodic sequence repeats in large genomic resources. *Nucleic Acids Research* 47:e8. DOI: 10.1093/nar/gky890.

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* 25:1043–55. DOI: 10.1101/gr.186072.114.

Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069. DOI: 10.1093/bioinformatics/btu153.

Sievers F, Higgins DG. 2017. Clustal Omega for making accurate alignments of many protein sequences. *Protein Science* 27:135–145. DOI: 10.1002/pro.3290.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31. DOI: 10.1186/1471-2105-6-31.

Srinivasan G. 2002. Pyrrolysine encoded by UAG in archaea: Charging of a UAG-decoding specialized tRNA. *Science* 296:1459–1462. DOI: 10.1126/science.1069588.

Studer A, Stupperich E, Vuilleumier S, Leisinger T. 2001. Chloromethane:tetrahydrofolate methyl transfer by two proteins from *Methylobacterium chloromethanicum* strain CM4. *European Journal of Biochemistry* 268:2931–2938. DOI: 10.1046/j.1432-1327.2001.02182.x.

The R Core Team. 2013. R: A language and environment for statistical computing.

The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* 45:D158–D169. DOI: https://doi.org/10.1093/nar/gkw1099.

Ticak T, Kountz DJ, Girosky KE, Krzycki JA, Ferguson DJ. 2014. A nonpyrrolysine member of the widely distributed trimethylamine methyltransferase family is a glycine betaine methyltransferase. *Proceedings of the National Academy of Sciences of the United States of America*:E4668–E4676. DOI: 10.1073/pnas.1409642111.

U.S. National Library of Medicine. 2019.TOXMAP. *Available at https://toxmap.nlm.nih.gov/toxmap/app/*

Urakawa H, Martens-Habbena W, Stahl DA. 2010. High abundance of ammonia-oxidizing archaea in coastal waters, determined using a modified DNA extraction method. *Applied and environmental microbiology* 76:2129–2135. DOI: 10.1128/AEM.02692-09.

US EPA. 2001. *Fact Sheet: Correcting the Henry's Law Constant for Soil Temperature*.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* 9:e112963. DOI: 10.1371/journal.pone.0112963.

Wolin EA, Wolin MJ, Wolfe RS. 1963. Formation of methane by bacterial extracts. *Journal of Biological Chemistry* 238:2882–2886.

Table and Figure Legends

Table 1. Comparison of the genomes of DCM-fermenting bacteria.

Figure 1. The removal of the methanogenic population from the DCM dechlorinating culture. (A) The initial three transfers (T1 – T3) of DCMD produced methane (black circles) in a molar ratio of 0.6 moles per mole DCM. DCM is shown both as actual concentration over time (white squares) as well as the cumulative DCM consumed (black squares). (B) DCM continued to be consumed in the presence (grey squares, subculture T4) and absence (white squares, subculture T5) of 2-bromoethanosulfonate, which caused methane production to cease.

Figure 2. Average coverage depth and read length across the DCMF genome assembly. (A) PacBio read depth along the full DCMF chromosome. Horizontal lines mark median depth (132×), and gradations as 1/8 median depth. (B) Maximum PacBio read length (kb) spanning each base along the full DCMF chromosome. Horizontal lines mark median length (15.3 kb), and gradations as 1/8 median length. Colours indicate total read length (blue), longest 5' distance from base spanned by a single read (purple), and longest 3' distance from base spanned by a single read (green).

Figure 3. **16S rRNA gene phylogenetic tree of DCMF with closely related bacteria (94-87% identity).** The two other known DCM-fermenting bacteria are underlined. Numbers indicate percentage of branch support from 1000 bootstraps. The scale bar indicates an evolutionary distance of 0.01 amino acid substitutions per site. Sequence alignments and tree construction were performed with MAFFT using the Archaeopteryx tool.

Figure 4. **Bootstrap-weighted combined taxonomic assignments for the DCMF predicted proteome based on TaxaMap processing of high-throughput phylogenetic analysis.** Results are shown at five taxonomic levels: genus, family, order, class and phylum. The asterisk (*) indicates where low abundance and/or unknown Firmicutes taxa have been combined at the genus, family, order and class levels.

Figure 5. **A heatmap representing the pairwise percentage distance matrix for the 82 full-length predicted trimethylamine methyltransferase protein sequences.** Proteins with 0%

567 distance (dark blue) are identical, while those with 100% distance (white) do not share any
568 sequence homology. The distance matrix was calculated using GABLAM and converted into a
569 heatmap using the gplots package in R.

Table 1(on next page)

Comparison of the genomes of DCM-fermenting bacteria.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27718v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

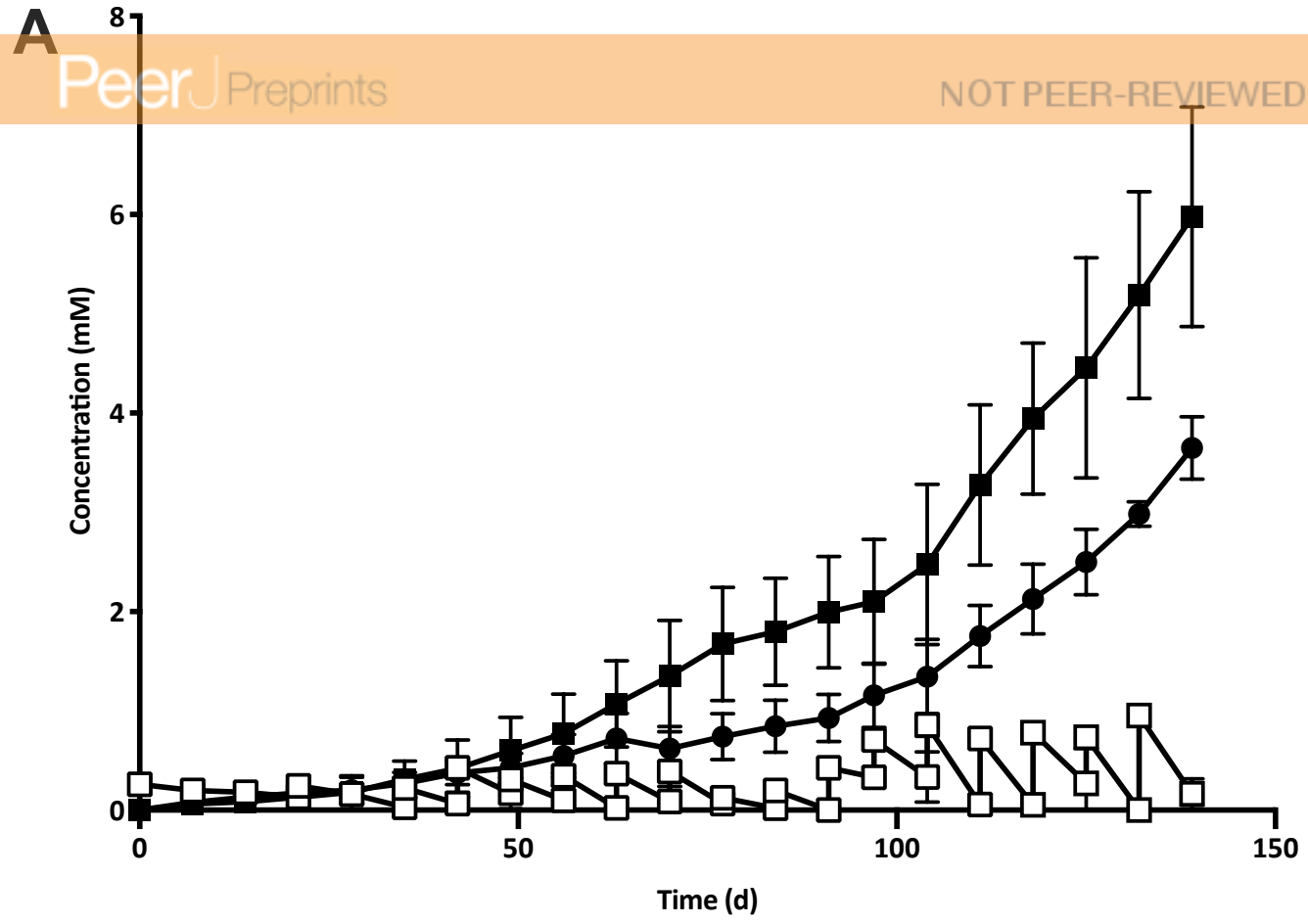
Table 1. Comparison of the genomes of DCM-fermenting bacteria.

	"DCMF"	*Dehalobacterium formicoaceticum*	"*Candidatus* Dichloromethanomonas elyunquensis"
GenBank Accession	CP017634.1	CP022121.1	LNDB00000000.1
Genome size (bp)	6,441,270	3,766,545	2,076,422
G+C content (%)	46.4	43.2	43.5
Contigs	1	1	53
Protein-coding sequences	5,773	3,935	2,323
Metabolic pathways/genes of interest			
Wood-Ljungdahl pathway	+	+	+
Reductive dehalogenases	-	-	+
Cobalamin biosynthesis	+	+	-
Glycine/betaine/sarcosine reductase complex	+	+	-
Methylamine methyltransferases	+	+	+
Pyrrolysine biosynthesis	+	+	-
Reference	This study	(Chen et al., 2017)	(Kleindienst et al., 2016)

Figure 1(on next page)

The removal of the methanogenic population from the DCM dechlorinating culture.

(A) The initial three transfers (T1 – T3) of DCMD produced methane (black circles) in a molar ratio of 0.6 moles per mole DCM. DCM is shown both as actual concentration over time (white squares) as well as the cumulative DCM consumed (black squares). (B) DCM continued to be consumed in the presence (grey squares, subculture T4) and absence (white squares, subculture T5) of 2-bromoethanosulfonate, which caused methane production to cease.



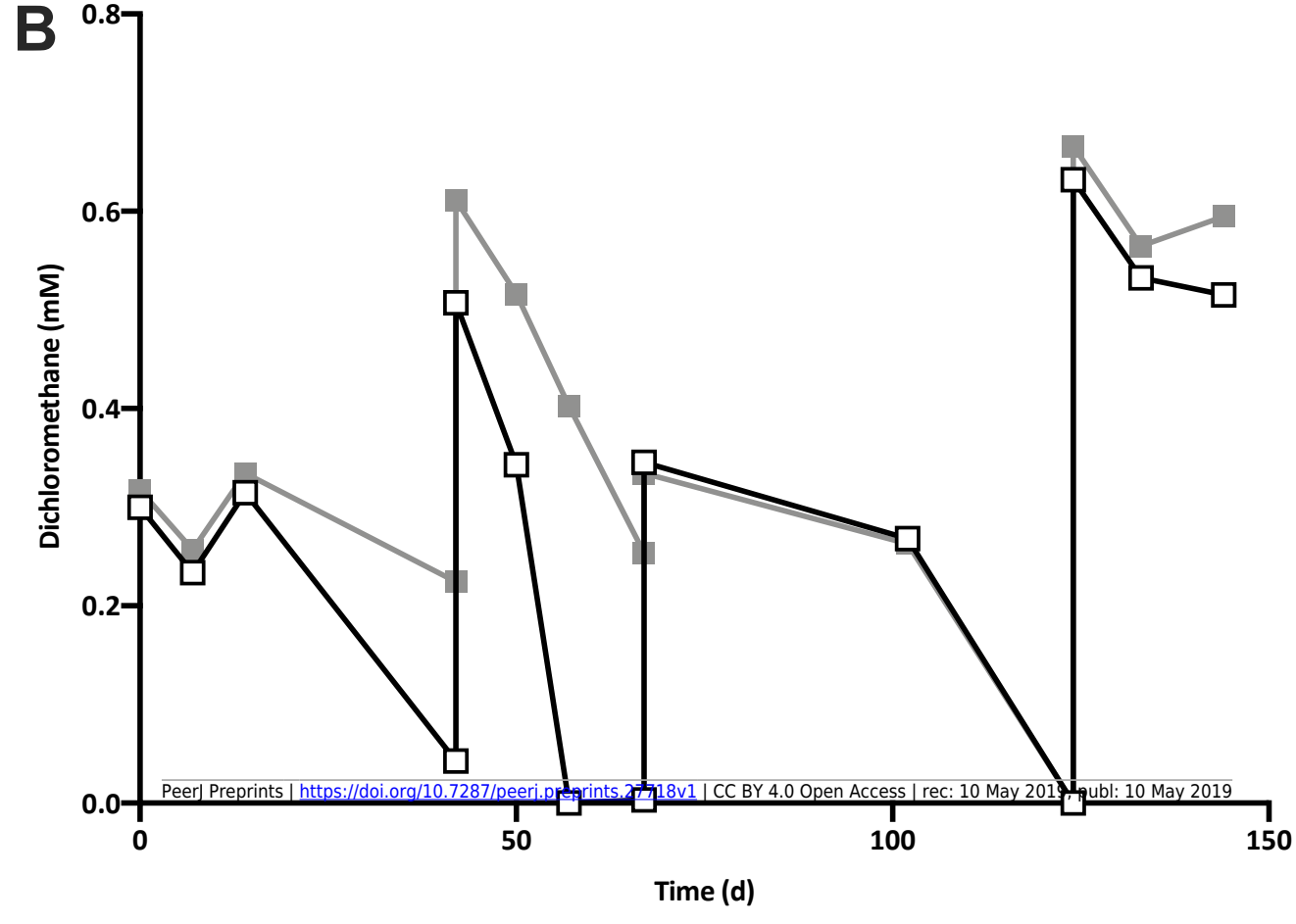


Figure 2

Average coverage depth and read length across the DCMF genome assembly.

(A) PacBio read depth along the full DCMF chromosome. Horizontal lines mark median depth (132×), and gradations as 1/8 median depth. (B) Maximum PacBio read length (kb) spanning each base along the full DCMF chromosome. Horizontal lines mark median length (15.3 kb), and gradations as 1/8 median length. Colours indicate total read length (blue), longest 5’ distance from base spanned by a single read (purple), and longest 3’ distance from base spanned by a single read (green).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27718v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

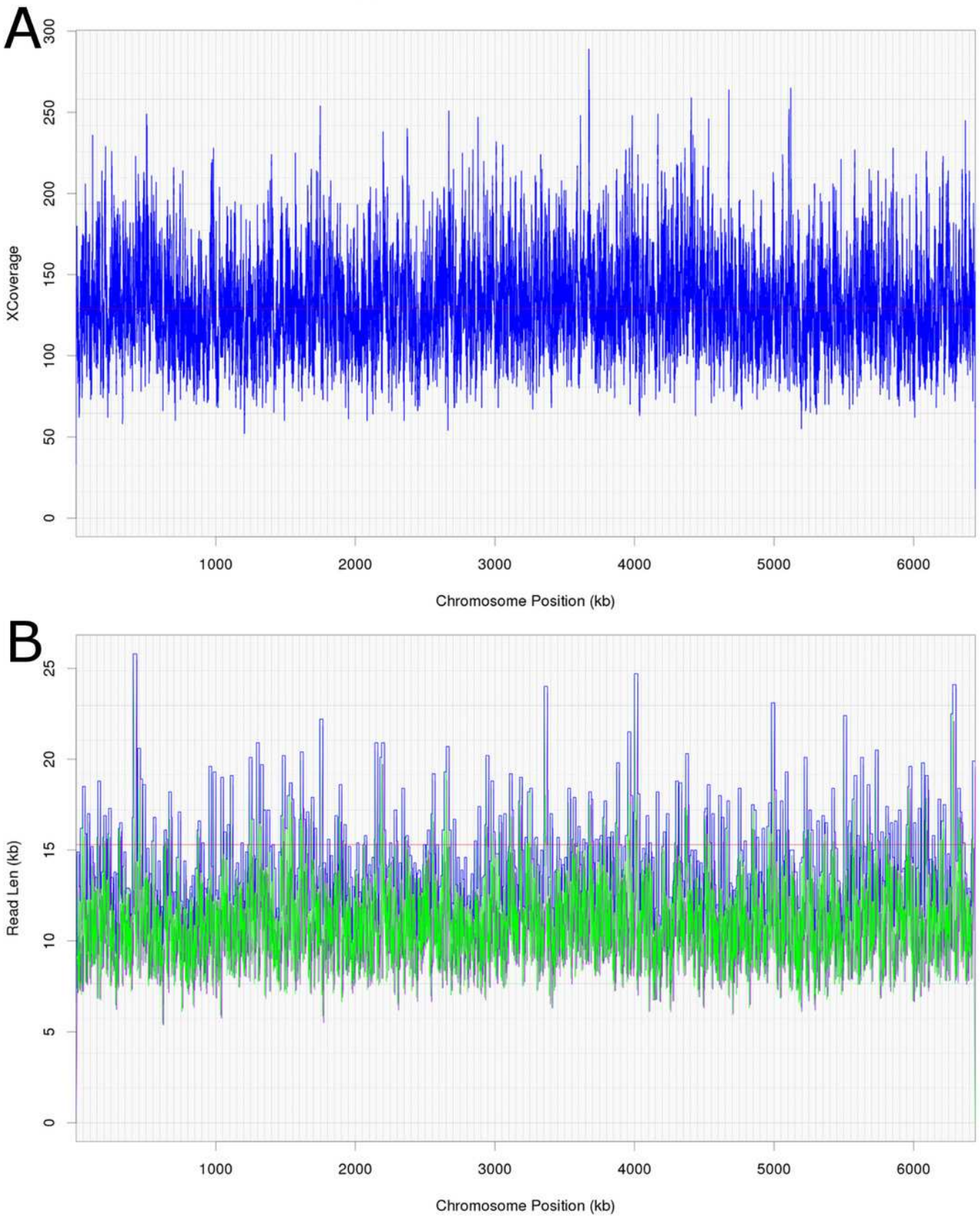

A
XCoverage
0
50
100
150
200
250
300
1000
2000
3000
4000
5000
6000
Chromosome Position (kb)
B
Read Len (kb)
0
5
10
15
20
25
1000
2000
3000
4000
5000
6000
Chromosome Position (kb)

Figure 3

16S rRNA gene phylogenetic tree of DCMF with closely related bacteria (94-87% identity).

Known DCM-fermenting bacteria are underlined . Numbers indicate percentage of branch support from 1000 bootstraps. The scale bar indicates an evolutionary distance of 0.01 amino acid substitutions per site. Sequence alignments and tree construction were performed with MAFFT using the Archaeopteryx tool.

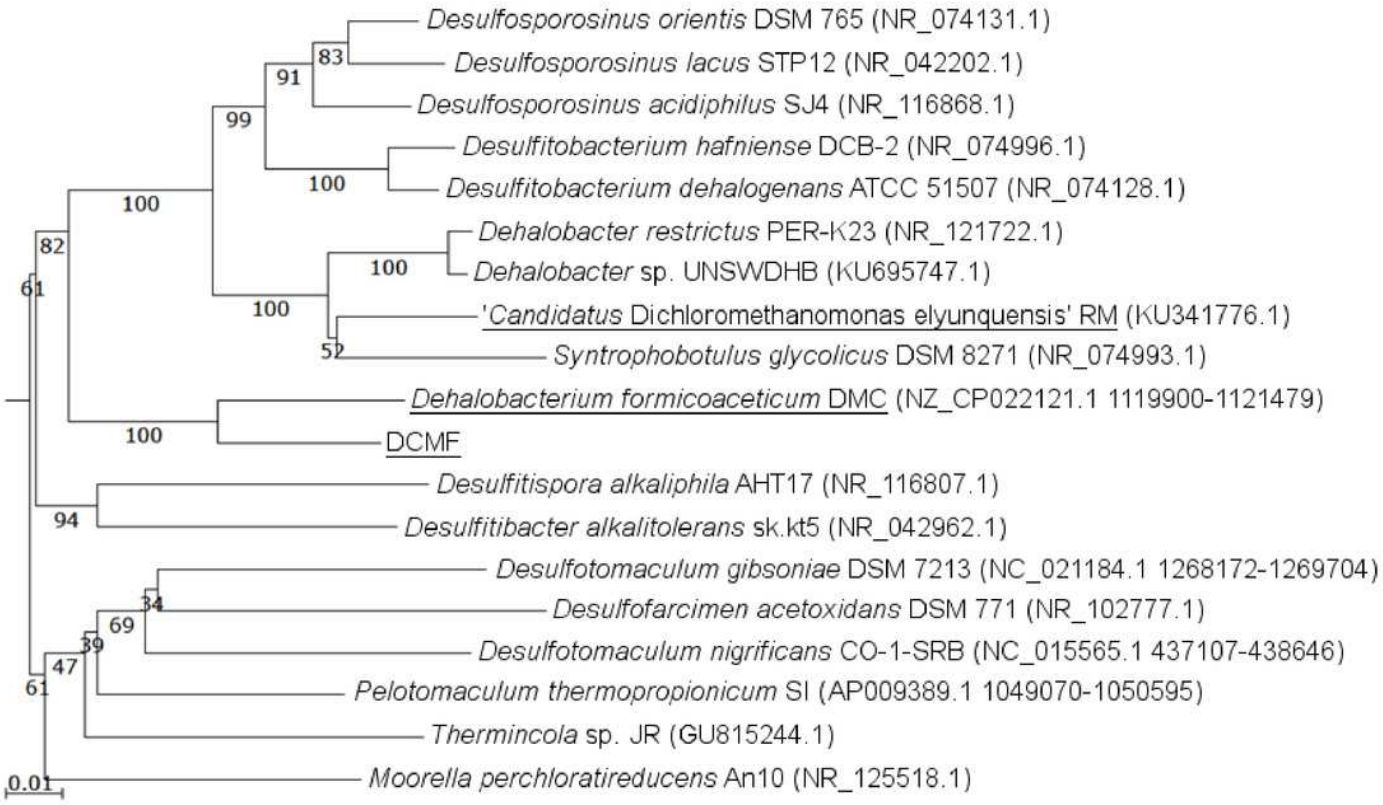


Figure 4

Bootstrap-weighted combined taxonomic assignments for the DCMF predicted proteome based on TaxaMap processing of high-throughput phylogenetic analysis.

Results are shown at five taxonomic levels: genus, family, order, class and phylum. The asterisk (*) indicates where low abundance and/or unknown Firmicutes taxa have been combined at the genus, family, order and class levels.

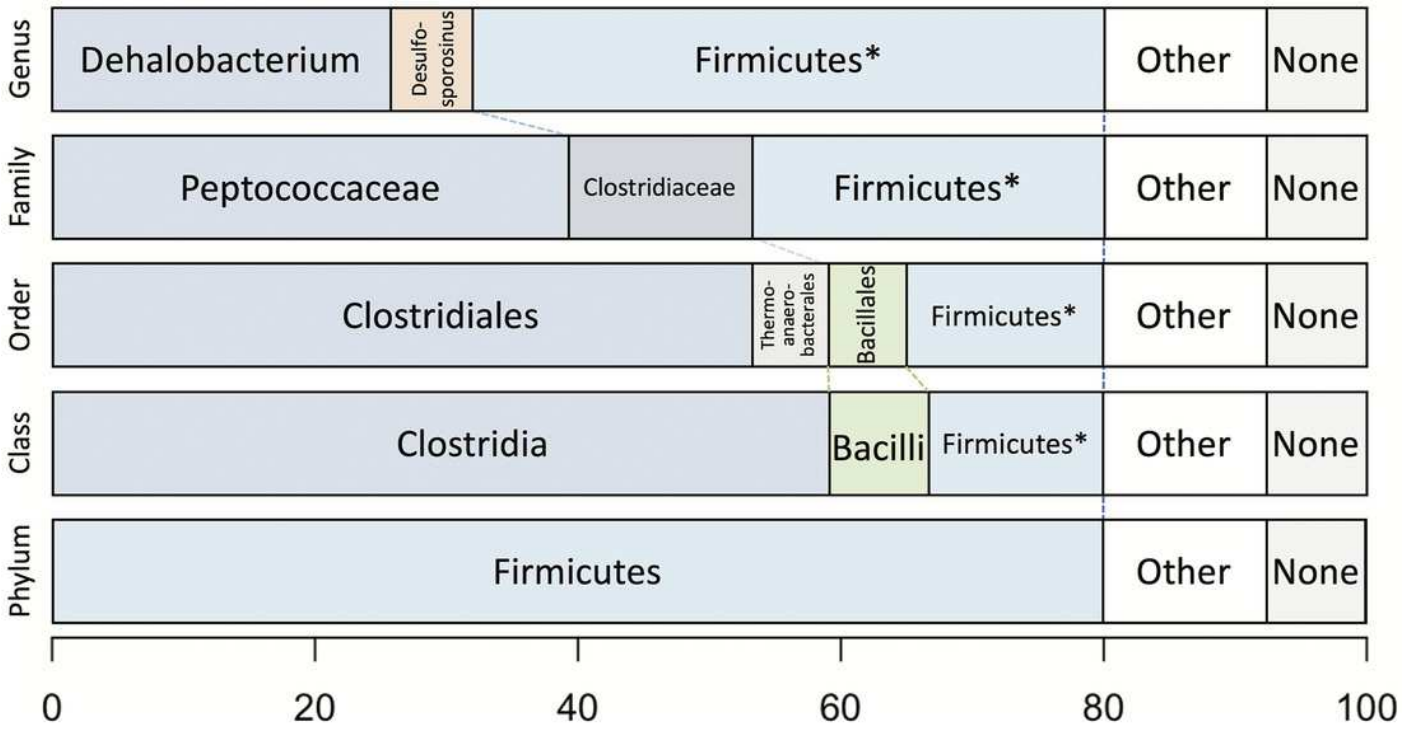
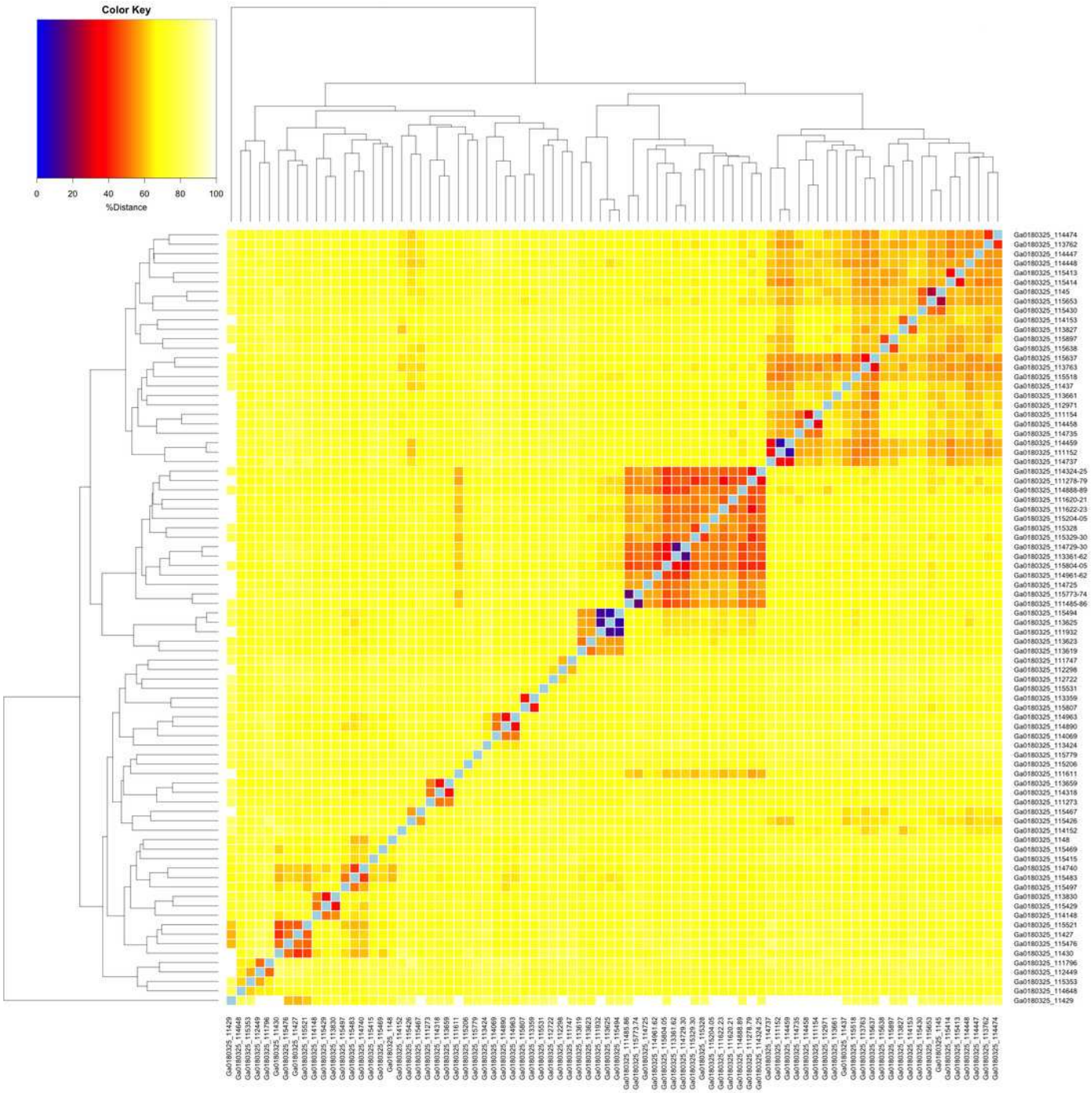


Figure 5

A heatmap representing the pairwise percentage distance matrix for the 82 full-length predicted trimethylamine methyltransferase protein sequences.

Proteins with 0% distance (dark blue) are identical, while those with 100% distance (white) do not share any sequence homology. The distance matrix was calculated using GABLAM and converted into a heatmap using the gplots package in R.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27718v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019


Color Key
0
20
40
60
80
100
%Distance