Genomic insights into metabolism and phylogeography of the mesophilic Thermotogae genus *Mesotoga*

Camilla L. Nesbø[1,2], Rhianna Charchuk[1], Stephen M. J. Pollo[1], Karen Budwill[3], Ilya V. Kublanov[4], Thomas H.A. Haverkamp[5,6] and Julia Foght[1]

1 Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada

2 BioZone, Dept. of Chemical Engineering and Applied Chemistry, Wallberg Building, University of Toronto, Toronto, ON, Canada.

3 InnoTech Alberta, 250 Karl Clark Road, Edmonton, Alberta, Canada T6N 1E4

4 Winogradsky Institute of Microbiology, Russian Academy of Sciences, Moscow, Russia

5 Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Blindern, Oslo, Norway.

6 Norwegian Veterinary Institute, Oslo, Norway.

*Corresponding Authors: nesbo@ualberta.ca

Running title: Genome analysis of *Mesotoga*.

Key words: Thermotogae, subsurface, recombination, oil reservoir, phylogeny, sulfur metabolism, hydrogenase, anaerobe.

Subject Category: Evolutionary Genetics

Abstract

The genus *Mesotoga*, the only described mesophilic Thermotogae lineage, is commonly detected in low-temperature anaerobic hydrocarbon-rich environments. Besides mesophily, *Mesotoga* displays lineage-specific phenotypes, such as no or little H_2 production, dependence on sulfur-compound reduction and ability to oxidize acetate, which may influence its ecological role. We used comparative genomics of 18 *Mesotoga* strains and a transcriptome of *M. prima* to investigate how life at moderate temperatures affects phylogeography, and to interrogate the genomic features of its lineage-specific metabolism. Phylogenetic analysis revealed three distinct *Mesotoga* lineages having different geographic distributions patterns and high levels of intra-lineage recombination but little gene-flow between lineages. All 16S rRNA genes from the genomes had > 99% identity whereas average nucleotide identity among genomes was 90.6–98.6% within groups and 80–86% between groups. Including data from metagenomes, phylogeographic patterns reveled that geographical separation is more important for *Mesotoga* than their thermophilic relatives, and we suggest its distribution is constrained by their strictly anaerobic lifestyle. We propose that H_2 oxidation and thiosulfate reduction are accomplished by a sulfide dehydrogenase and a hydrogenase-complex and that a pyruvate:ferredoxin oxidoreductase acquired from *Clostridia* is responsible for oxidizing acetate.

Introduction

The genus *Mesotoga* is the only characterized mesophilic lineage within the otherwise thermophilic bacterial phylum Thermotogae (Pollo *et al.*, 2015). *Mesotoga* spp. have been isolated from and detected in polluted marine sediments, low temperature oil reservoirs, and waste water treatment facilities (Nesbø *et al.*, 2010; 2012; Hania *et al.*, 2011; 2013), and are common in anaerobic methanogenic environments (Nesbø *et al.*, 2010) where they may be involved in syntrophic acetate degradation (Nobu *et al.*, 2015). The first described member of this genus, *Mesotoga prima* MesG1Ag4.2 (hereafter, *M. prima*), was isolated from a PCB-degrading enrichment culture inoculated with sediments from Baltimore Harbor, Maryland (USA) (Nesbø *et al.*, 2006; 2012). Sequencing the genomes of *M. prima* and the very closely related *M. prima* PhosAc3 (hereafter, PhosAc3) isolated in Tunisia (Hania *et al.*, 2015) revealed larger genomes than thermophilic Thermotogae, with more genes involved in regulatory functions and interactions with the environment (Zhaxybayeva *et al.*, 2012).

Genome size in Thermotogae inversely correlates with optimum growth temperature (Pollo *et al.*, 2015; Zhaxybayeva *et al.*, 2012). However, it is unclear how growth temperature affects other aspects of genome evolution including levels of homologous recombination. Hyperthermophilic *Thermotoga* display extremely high levels of homologous recombination, which could be a side effect of the need for DNA repair at high temperatures (Nesbø *et al.*, 2015). Nesbø *et al.* (2015) also found that *Thermotoga* spp. from similar environments have exchanged more genes than geographically close isolates from different environments. This has been observed for metagenomes in general, where geographical distance has little effect on gene sharing,

and gene composition is more strongly affected by ecology than proximity (Fondi *et al.*, 2016). Moreover, the phylogeography and patterns of gene flow among *Thermotoga* genomes suggested that oil reservoirs were colonized from subsurface populations rather than being buried with the sediments that mature into oil reservoirs (a corollary of the paleosterilization hypothesis; Wilhelms *et al.*, 2001). Comparative genomic analyses of mesophilic Thermotogae may shed light on the role of growth temperature on recombination and phylogeography.

In addition to lower optimal growth temperature (37°C - 40°C), *Mesotoga*'s core energy metabolism also differs from that of other characterized Thermotogae. For instance, while growth of most Thermotogae is stimulated by adding sulfur-compounds to the medium (Ravot *et al.*, 1995; Boileau *et al.*, 2016), reduction of sulfur compounds appears to be essential for growth of *Mesotoga* in pure culture and they produce little or no H_2 (Hania *et al.*, 2011; 2013; Fadhlaoui *et al.*, 2017).

Here we compare 18 *Mesotoga* genomes obtained from isolates and single cells originating from six geographically different sites, including three low temperature continental oil reservoirs, in order to elucidate genomic markers of metabolic differences and to investigate how growth temperature influences phylogeography and prevalence of recombination. We also compare our genomes to *Mesotoga* sequences available in publicly available metagenomes. We infer that geographic separation has had a significant influence on the phylogeography of *Mesotoga,* possibly due to selective pressures of dispersal of strict anaerobes through aerobic environments. Finally, we present a model that accounts for *Mesotoga*'s distinct sulfur-dependent metabolism involving a hydrogenase complex.

Materials and Methods

Sources of genome sequences

Nine *Mesotoga* strains (BR, HF and BH designations) were isolated from oil reservoirs and anaerobic sediments in Canada and the USA (Table 1). In addition, seven single cells were physically selected from oil field fluids or oil sands enrichment cultures from Canada or a contaminated aquifer in the USA (PW, NAPDC and TOLDC designations, respectively) and amplified by PCR to produce single cell amplified genomes (SAGs). Detailed descriptions of isolation procedures, DNA extraction, genome assembly and annotation are provided in Supplementary Information.

To augment the strain genomes, 15 publicly available metagenomes containing large numbers of *Mesotoga* spp. sequences were identified using blastn searches of IMG (JGI; accessed February 2017) and SRA (NCBI; accessed December 2016) using *rpoB* from *M. prima* as a probe and expected (exp.) set to $< e\text{-}^{50}$. For additional details on search parameters and information on assembly of draft genomes from metagenomic sequences or contigs see Supplementary Information.

Genome content and genome alignments

Shared genes were identified in IMG Version 4 (Markowitz *et al.*, 2014) using translated proteins and 70% identity cut-off and exp. $< e^{-10}$, whereas 30% sequence identity cut-off and exp. $< e^{-5}$ were used to identify strain-specific genes and for comparing *Mesotoga* genomes to other Thermotogae genomes.

Pan-genome calculations were performed in Panseq (Laing *et al.*, 2010) using a fragment size of 500 bp and 70% identity cutoff, and exp. < e^{-10}. The data matrices of shared core single nucleotide polymorphisms (SNPs) and 500-bp fragments were converted into uncorrected distances and visualized in SplitsTree 4 (Huson and Bryant, 2006) using NeighborNet clustering.

Whole genome alignments were carried out in MAUVE version 2.3.1 (Darling *et al.*, 2010) using automatically calculated seed weights and minimum Locally Collinear Blocks (LCB) scores. LCB positions with gaps were removed and the edited LCB were concatenated in Geneious v.10 (www.geneious.com). Average nucleotide identities (ANI) were calculated at http://enve-omics.ce.gatech.edu/ani/ (Goris *et al.*, 2007).

Recombination detection

The relative rate of recombination to mutation within lineages, as well as the average recombination tract length, were assessed using the LDhat package (McVean *et al.*, 2002; Jolley, 2004) as described by Nesbø *et al.* (2015) on concatenated alignments (including LCB > 10,000 bp) of genomes from the W and the A lineage separately. Recombinant fragments between lineages were detected using LikeWind Version 1.0 (Archibald 2002) on the concatenated MAUVE alignment (above), using a sliding window of 1000 bp with 100-bp increments.

RNAseq analysis

RNA isolation from a culture of *M. prima* (grown at 45ºC for 73 h in 0.5% yeast extract, 0.01 M thiosulfate and 0.5% xylose) and subsequent sequencing as one of five barcoded

133 libraries were performed as described by Pollo et al. (2017). RNAseq analysis was carried
134 out in CLC Genomics Workbench version 7.0.4 as described by Pollo *et al.*(2017).
135
136 *H_2 and H_2S measurements*
137 Standard gas chromatographic analysis of culture headspace gas was performed using an
138 Agilent CP4900 Micro Gas Chromatograph to detect H_2 production by the cultures, as
139 described in Supplementary Information. Dissolved sulfide concentrations were measured
140 using a VACUettes® Visual High Range Kit (Chemetrics), following the manufacturer's
141 recommendations.
142
143 ## Results and Discussion
144 **Phylogenetic and genomic analyses of isolates and single cells reveal three distinct**
145 ***Mesotoga* lineages**
146 We generated draft genomes for eight newly isolated *Mesotoga* strains from two Albertan
147 oil reservoirs (H and B), and one *Mesotoga* strain from a PCB-degrading enrichment
148 culture from Baltimore Harbor, Maryland (Table 1). Seven partial SAGs were obtained
149 from cells sorted from produced water from an Albertan oil reservoir (PW), a naphtha-
150 degrading enrichment culture inoculated with sediments from an Albertan oil sands
151 tailings pond (NAPDC), and a toluene-degrading enrichment culture inoculated with
152 sediments from a contaminated aquifer in Colorado (TOLDC). We also included in our
153 analyses the draft genome of PhosAc3, previously isolated in Tunisia (Hania *et al.*, 2015)
154 and the closed genome of *M. prima* (Nesbo *et al.*, 2012) from Baltimore Harbor.

155 The 16S rRNA genes of all 17 genomes had ≥99% identity to *M. prima*;
156 phylogenetic trees revealed three distinct lineages (Figure 1a). Genome networks based
157 on core SNPs also had topologies consistent with 16S rRNA gene phylogeny, with three
158 distinct lineages being evident (Figure 1b). Two lineages have a widespread geographical
159 distribution: the World lineage (W; found in all regions represented) and the lineage
160 found in locations in USA (US; found in Baltimore Harbor and Colorado). The Alberta
161 (A) lineage was observed in the Albertan samples only. Interestingly, *M. prima* has one
162 16S rRNA from the W lineage and one from the US lineage, suggesting one copy has
163 been acquired laterally.
164 Very little reticulate evolution was observed among the three *Mesotoga* lineages
165 (Figure 1b) and the A lineage in particular showed very little connection with the other
166 two groups, suggesting that the three lineages have evolved independently for a relatively
167 long time. The same pattern was observed for the pan-genome, with most lateral
168 connections occurring within groups (Figure 1c).
169 A high level of recombination was detected, with the majority (> 200) of
170 recombination events involving genomes from the same lineage (Figure S1). For the W
171 and A lineages, respectively, the average recombination tract length was estimated to be
172 36,000 – 56,000 bp and 17,000–23,000 bp; the population mutation rate (θ) was
173 estimated to be 0.022 and 0.013, and the population recombination rate (γ) to be 1.8 (1.5–
174 2.2) and 1.5 (1.3–1.7). The resulting high γ/θ ratios of ~82–115 indicate high levels of
175 recombination and are similar to estimates for *Thermotoga* spp. (Nesbø *et al.*, 2015).
176 Phylogenetic analysis identified 52 regions with recombination likely occurring
177 between lineages: 39 regions showed evidence of recombination between *Mesotoga* sp.

178 BH458 and the W lineage, eight regions suggested recombination between *Mesotoga* sp.
179 BH458 and the A lineage, and only five regions showed possible recombination between
180 A- and W-lineage genomes (Figure 2). The regions with recombination involving the A
181 lineage were relatively short (range 230–530 bp) and the sequences more divergent,
182 whereas several of the fragments involving the W lineage and *Mesotoga* sp. BH458 were
183 > 5 kb (average 3000 bp, range 260–20,000), suggesting that recombination events
184 between the W lineage and *Mesotoga* sp. BH458 are more recent than those involving the
185 A lineage. Very high levels of recombination were observed for a few genes. Among
186 these is Theba_0319 in *M. prima*, the fourth most highly expressed gene (Supplementary
187 Table S1) that encodes the OmpB protein (Petrus *et al.*, 2012), a major component of the
188 toga of Thermotogae. This high level of recombination is consistent with strong
189 diversifying selective pressure on surface proteins, commonly observed in pathogenic
190 bacteria (e.g., Wachter and Hill, 2016).
191 The observation of several recent recombination events between the W and US
192 lineages, which currently co-exist in at least one location (i.e., Baltimore Harbor),
193 demonstrate that recombination between lineages is possible. We therefore suggest that
194 the three *Mesotoga* lineages have evolved independently due to geographical, not genetic,
195 isolation. Although it may seem counterintuitive that mesophilic *Mesotoga* would be
196 more affected by geographical separation than hyperthermophilic *Thermotoga*, this may
197 be a consequence of their strictly anaerobic metabolism. Chakraborty *et al.* (2018)
198 showed that bacteria are dispersed out of deep hot subsurface oil reservoirs and into the
199 ocean through hydrocarbon seeps, and this might serve as a major route of migration
200 between these environments. Hyperthermophilic *Thermotoga* cells will be inactive in

cold aerobic ocean water, but mesophilic *Mesotoga* cells will be active at these temperatures and therefore may quickly succumb to oxygen exposure, limiting viable dispersal. In support of this observation, many *Mesotoga*-specific genes appear to be involved in O_2 or H_2O_2 detoxification (see below).

Mesotoga lineages have conserved core genomes and diverse pan-genomes

Comparative analysis of *Mesotoga* genomes revealed greater gene-content diversity than was observed in hyperthermophilic Thermotogae (Nesbø *et al.*, 2015). The pan-genome of the *Mesotoga* isolates and SAGs was estimated to be 7,452,537 bp with an accessory genome of 5,664,475 bp; each genome contained a considerable amount of lineage-specific DNA (Figure S1; see Supplementary Information for additional details of the pan-genome and within-sample site diversity). In pairwise comparisons, the genomes shared on average 76% of their genes (Supplementary Table S2). Genomes from each lineage defined in Figure 1 shared, on average, 81% - 89% of their genes whereas in pairwise comparisons *Thermotoga* spp. share > 90% of their genes (Nesbø *et al.*, 2015). Comparing genomes from different lineages, the US lineage had an intermediate position, sharing more genes with the A and W lineages: on average, genomes from A and W share 69% of genes, W and US share 73%, and A and US share 75% of their genes.

Comparing nucleotide divergence within the core genomes revealed 'species-level' divergence among the three lineages detected (ANI < 86%), whereas ANI within the A and W lineages was very high at 98.6% and 97.9%, respectively (Supplementary Table S3). In comparison, ANI among *Thermotoga* genomes was 95.3% (Nesbø *et al.*, 2015). Thus *Mesotoga* spp., particularly those from the W lineage, appears to have more

conserved core genomes and more diverse pan-genomes than their hyperthermophilic relatives.

Nesbø *et al.* (2015) suggested that high levels of recombination may be responsible for homogenizing the *Thermotoga* spp. genomes. However, since we observed similar high levels of recombination within the *Mesotoga* W and A lineages, additional forces must be responsible for the differences in pan-genome sizes. Perhaps more cryptic niches are available in low-temperature than in high-temperature subsurface environments (McInerney *et al.*, 2017), or *Mesotoga* may have larger effective population sizes *in situ* than the hyperthermophiles (Andreani *et al.*, 2017). *Mesotoga* spp. might also interact with viruses more frequently, which may drive population diversity (Koskella and Brockhurst, 2014).

Phylogeographic patterns of the three *Mesotoga* lineages: isolation by distance, range expansion and burial with isolation

To challenge our interim conclusions about genome structure and geographic distribution, we expanded the *Mesotoga* sequence dataset by searching IMG/M (in JGI) and SRA (in NCBI) databases for metagenomes containing *Mesotoga* spp. sequences. Fifteen metagenomes containing sequences closely related to the *Mesotoga* genomes investigated here were identified, arising from two environments already described (tailings pond and oil reservoir in Alberta), as well as oil reservoirs, contaminated sediments, wastewaters and hotspring sediments across the continental USA, and wastewaters in China (Table 2 and Supplementary Information).

246 ***Recent range expansion of the W-lineage:*** Blastn comparisons confirmed wide
247 distribution of the *Mesotoga* W lineage in wastewater treatment systems. A network
248 including *Mesotoga* contigs from three metagenomes dominated by W lineage sequences
249 (Long Beach, Boston and Hong Kong) revealed no geographical structuring, which,
250 together with high identity in core sequences (Figure S3a), suggests a recent range
251 expansion of this lineage (Choudoir *et al.*, 2017). Interestingly, W lineage *Mesotoga* have
252 been detected only at sites heavily influenced by human activities (e.g., drilling,
253 contamination), suggesting an anthropogenic role in their dispersal. One of the genes
254 specific to the W lineage (Theba_0620, Supplementary Information) is involved in
255 synthesis of poly-gamma-glutamate, which has been implicated in cell survival under
256 harsh conditions and may have contributed to this lineage's wide distribution.
257 ***Isolation by distance can explain the distribution of US-genomes:*** The
258 metagenome data greatly expanded the observed distribution of the US lineage. As
259 expected, metagenome IMG 15764 from Albertan oil reservoir E (the source of *Mesotoga*
260 sp. SC_PW1-3) contained sequences with high identity to the A lineage. However, it also
261 contained many sequences with high identity to the US lineage (Table 2), and sequence
262 binning yielded two *Mesotoga* metagenome-assembled genomes (MAGs): one most
263 similar to NA-genomes (Figure S3b) and one with a mix of sequences from the A lineage
264 and US lineage (not shown). The intermediate position of US between the A- and W-
265 lineages noted above may therefore be due to this lineage co-existing with both the W
266 and A lineages.
267 The network of NA-*Mesotoga* including the metagenomes in Table 2 revealed
268 three groups (Figure S3b) where metagenomes from New York and Blank Spring

(California) form a cluster that does not contain any genomes sequenced by us. The remaining metagenomes clustered according to geography as well as by environment type: the MAG assembled from oil reservoir E (Alberta), two MAGs from an Alaskan oil reservoir, and the *Mesotoga* sequences from Alameda (California) clustered with SC_TOLDC from Colorado (western North America), while the *Mesotoga* sequences from New Jersey clustered with *Mesotoga* sp. BH458 from Baltimore Harbor (eastern North America). We therefore suggest that the divergence patterns seen for this lineage can be explained at least partly by an isolation-by-distance model.

Evolution of the A-lineage in isolation in North-American oil reservoirs: The metagenome sequences revealed that the A lineage is not restricted to Alberta, nor is it specific to oil reservoirs (Table 2), with substantial numbers of A-lineage sequences detected in wastewater metagenomes. For this lineage, MAGs were available from the same oil reservoir in Alaska as where we observed the NA-lineage (Hu *et al.*, 2016), an anaerobic wastewater digester in Oakland (California), and one, assembled by us, from a PCB-fed culture inoculated with sediments from Liangjiang River, China (Wang and He, 2013). Network analysis revealed that the genome from the Alaskan oil reservoir is most similar to those from the Albertan oil reservoir B, whereas the genomes from China and California show high similarity (> 99%) to each other and to *Mesotoga* sp. SC_NapDC from a northern Albertan oil sands tailings pond (Figure S3c). In fact, these genomes show the second highest level of pairwise identity among the A lineage genomes (Figure S3d), suggesting a recent dispersal possibly due to human activities in the last decades.

The A lineage is more isolated from the other lineages (Figure 1 and 2), which might suggest that this clade evolved in isolation since the formation of oil reservoir

292 sediments in Alberta 55–120 Ma (Head, 2014; Schaefer, 2005; Selby, 2005)). The high
293 similarity of the MAGs from the Alaskan oil field to the Albertan genomes and MAGs
294 from the A and NA lineages (Figure S3) could be due to these oil reservoir sediments
295 being laid down around the same time (~100 Ma Hu *et al.*, 2016). However, the position
296 of these MAG in the genome networks could also be explained by these oil reservoirs
297 being colonized by the same subsurface population, as suggested for *Thermotoga* spp.
298 (Nesbø *et al.*, 2015). Additional indigenous oil reservoir genomes are needed to resolve
299 this question.
300
301 **Distinct metabolism in mesophilic Thermotogae**
302 Phylogeographic analyses provided intriguing evidence that *Mesotoga*'s lower growth
303 temperature has influenced its distribution patterns. However, the *Mesotoga* genus also
304 displays additional lineage-specific phenotypes. We therefore examined the newly
305 available genomes for metabolic insights, which may be linked to *Mesotoga*'s lower
306 growth temperatures and may influence the role(s) *Mesotoga* play in their environments.
307 ***Mesotoga*-specific genes:** Comparing the *Mesotoga* isolate genomes to other
308 Thermotogae genomes in IMG (Markowitz *et al.*, 2014) revealed 200 *M. prima* genes
309 found in all *Mesotoga* genomes (including the more distantly related *M. infera* not
310 included in the phylogenomic analyses), but in no other Thermotogae genomes. The
311 majority of these genes were hypothetical proteins (N=119, Supplementary Table S1).
312 When *Mesotoga*-specific genes with a predicted function were classified according to
313 Clusters of Orthologous Groups (COG) categories, the largest category was 'Amino Acid
314 metabolism and transport' with 11 genes, most of which were dipeptidases (COG4690,

315 N=6). Interestingly, amino acid metabolism genes were up-regulated during growth at
316 low temperatures in *Mesotoga*'s closest relative *Kosmotoga olearia* (Pollo *et al.*, 2017),
317 suggesting a role in low temperature adaptations.

318 ***Reducing equivalents and thiosulfate reduction:*** *Mesotoga*'s core metabolism
319 differs from that of other characterized Thermotogae. While growth of most Thermotogae
320 is stimulated by adding thiosulfate, sulfur, or other reduced sulfur compounds to the
321 medium (Ravot *et al.*, 1995; Boileau *et al.*, 2016), reduction of sulfur compounds appears
322 to be essential for growth of *Mesotoga* in pure culture (Hania *et al.*, 2011; 2013;
323 Fadhlaoui *et al.*, 2017). The first description of *M. prima* (Nesbø *et al.*, 2012) reported
324 that growth was only slightly stimulated by the presence of thiosulfate and sulfur.
325 However, here we observed growth of this isolate only with sulfur or thiosulfate
326 (Supplementary Table S4 and Table S5), confirming that this is a general trait of
327 *Mesotoga* spp. Additionally, while other Thermotogae produce H_2 (and H_2S if grown with
328 partially reduced sulfur compounds) *Mesotoga* spp. produce large amounts of H_2S and no
329 or little H_2 (Supplementary Table S5). Fadhlaoui *et al.*, (2017) suggested that *Mesotoga*'s
330 inability to ferment sugars without sulfur compounds is due to the lack of a bifurcating
331 hydrogenase observed, for instance, in *Thermotoga* spp. However, *K. olearia* also lacks
332 this enzyme yet produces large amounts of H_2 while fermenting pyruvate by using a Fe-
333 hydrogenase (Kole_0172) (Pollo *et al.*, 2017).

334 To reconcile these observations with genomic data, transcriptome analysis was
335 performed using a culture of *M. prima* grown with 0.5% yeast extract, xylose and
336 thiosulfate. RNAseq analysis revealed high expression of Theba_0443 (RPKM of 3650;
337 Supplementary Tables S1 and S6) encoding a Fe-hydrogenase homologous to the one

used by *K. olearia* (Kole_0172). Hydrogenases are indeed essential in Thermotogae for recycling of ferredoxins (Schut *et al.*, 2013); therefore, finding the same hydrogenase to be highly expressed in *M. prima* and *K. olearia*, and conserved in all *Mesotoga* genomes investigated here, suggests that *Mesotoga* possesses a mechanism relying on oxidized sulfur compounds, efficiently converting all intracellularly produced H_2 to H_2S. Notably, there was no change in the culture headspace gas H_2:N_2 ratio after incubating *Mesotoga* spp. in a 1:9 H_2:N_2 atmosphere for > 5 months (Supplementary Table S5), suggesting that *Mesotoga* neither produces nor takes up externally supplied H_2.

No homologs of characterized thiosulfate reductases were identified, although the *Mesotoga* genomes carry homologs (Theba_0076; Theba_0077 in *M. prima*) of an archaeal intracellular ferredoxin:NADP oxidoreductase (SudAB; Hagen *et al.*, 2000) capable of acting as a sulfide dehydrogenase in the presence of elemental sulfur or polysulfide (Figure 3). Both genes were transcribed at moderate levels in *M. prima* grown with thiosulfate (RPKM 341 and 243, respectively), while the *K. olearia* homologs (Kole_1827, Kole_1828) were highly expressed under similar conditions (RPKM > 1000, (Pollo *et al.*, 2017). SudAB complexes, however, are not known to be involved in thiosulfate reduction. This is probably due to an unfavorable E°= 82 mV for the reaction when NADH acts as electron donor: $E^{\circ`}$ [$S_2O_3^{2-}$/ $HS^-$ + $SO_3^{2-}$] = -402 mV and $E^{\circ`}$ [NAD+/ NADH] = -320 mV. The $E^{\circ`}$ of [Fd_{Ox} / Fd_{Red}] is similarly high at -390 mV. Comparable endergonic reactions are catalyzed by the *Salmonella enterica* thiosulfate reductase Phs by utilizing proton-motive force (Stoffels *et al.*, 2012). However, the cytoplasmic SudAB complex cannot couple proton-motive force and reduction of an external electron acceptor. Thus, neither NADH nor Fd_{Red} can function as electron donors

for thiosulfate reduction by *M. prima*. Instead molecular H_2 with $E^{\circ`}$ $[2H^+ / H_2] = -410$ mV appears to be a thermodynamically preferable electron donor for thiosulfate reduction. The only hydrogenase present in the *M. prima* genome is the highly expressed FeFe-hydrogenase (Theba_0443), which usually is involved in Fd-dependent H_2 production (Vignais and Billoud, 2007). A cluster of five highly transcribed genes (Theba_0461 – 0465, RPKM 1203-3697, Supplementary Tables S1 and S6) encode proteins homologous to all subunits of the NADP-reducing hydrogenase Hnd of *Desulfovibrio fructosovorans* (Nouailler *et al.*, 2006) except the catalytic subunit (HndD). These proteins may work together with Theba_0443 to form a FeFe-hydrogenase complex (Figure 3). We hypothesize that this complex is involved in intracellular synthesis of molecular hydrogen for thiosulfate reduction by SudAB coupled to NADH oxidation (Fd_{Red}, Figure 3). The Hnd genes have homologs in other Thermotogae, however, similar genomic context is observed only in genomes of other *Mesotoga* and *Kosmotoga* spp. (Supplementary Table S7).

Mesotoga cells require enzymes that re-oxidize Fd_{red} formed during sugar oxidation. This might be carried out by either the NADP:ferredoxin oxidoreductase complex (Mbx; Theba_1796-1808 in *M. prima*. Schut et al., 2013) or the Rnf ion-motive electron transport complex (Theba_1343-1348; Müller *et al.*, 2008). Conserved motifs (Mulkidjanian *et al.*, 2008) suggested a Na^+-translocating F-type ATP synthase operating in *M. prima.* As a consequence, both Mbx and Rnf complexes are predicted to export Na^+ generating sodium- motive force instead of protein-motive force. Genes encoding Mbx and Rnf show low and moderate expression (RPKM 37-88 and 236-478, respectively) during growth on thiosulfate, and the expression values suggests that Rnf is the main

384 complex involved. The model shown in Figure 3 accounts for the observed dependence
385 of *M. prima* on sulfur or thiosulfate for growth, the lack of H_2 production, and involves
386 genes already implicated in hydrogen and sulfur metabolism. Importantly, however,
387 currently there are no known enzymes that couple H_2 oxidation and thiosulfate/sulfur
388 reduction. It is therefore possible that *M. prima* SudAB uses NADH as the electron donor
389 and it is much more effective than the hydrogenase which results in almost no H_2 as
390 growth product (Figure 3 panel C).

391 Alternatively, thiosulfate reduction coupled to H_2 oxidation (i.e., the postulated
392 role of SudAB; Figure 3a) may be performed solely by one of the highly-transcribed
393 hypothetical *Mesotoga* protein with no match to genes in *Kosmotoga* and other
394 Thermotogae, or in combination with SudAB (Figure 3 panel b). Several candidate genes
395 listed in Supplementary Table S6 encode proteins with unknown functions. Functional
396 studies of these genes, as well as the gene products shown in Figure 3, are needed to
397 assess their role, if any, in thiosulfate reduction. Additional genes that may be involved in
398 thiosulfate uptake and electron transfer are also discussed in Supplementary Information.

399 Interestingly, PhosAc3 and *M. infera* were found to reduce only elemental sulfur
400 (Hania *et al.*, 2015; 2013) whereas the strains isolated by us also reduce thiosulfate.
401 These differences may reflect selection during isolation; all the isolates obtained in the
402 current study were from enrichment cultures containing thiosulfate, whereas PhosAc3
403 and *M. infera* were enriched with sulfur. This suggests that the sulfur-compound-
404 preference may be a variable and flexible phenotype in *Mesotoga* populations.

405 ***Acetate and xylose utilization:*** Growth on acetate was reported for *Mesotoga*
406 PhosAc3 (Hania *et al.*, 2015), and we observed weak stimulation of growth of its close

407 relative *M. prima* on acetate (day 5-10 in Supplementary Figure S4 and Table S4). Nobu
408 *et al.* (2015) suggested that Ca. "*Mesotoga acetoxidans*", a MAG closely related to *M.*
409 *infera,* oxidizes acetate by using a novel pathway even though the genes comprising the
410 pathway are conserved in all Thermotogae genomes. Yet, this phenotype is uncommon
411 among Thermotogae and has been reported only for *Pseudothermotoga lettingae* (Balk,
412 2002). Instead, many Thermotogae are inhibited by acetate, including one of *Mesotoga*'s
413 closest relatives, *K. olearia* (Dipippo *et al.*, 2009). Our search for *Mesotoga*-specific
414 genes that may be responsible for their observed growth on acetate revealed candidate
415 gene encoding a bacterial homodimeric pyruvate:ferredoxin oxidoreductase (PFOR;
416 Theba_1954). Its close homologs were found only in *Kosmotoga pacifica* (L'Haridon *et*
417 *al.*, 2013) and *Mesoaciditoga lauensis* (Reysenbach *et al.*, 2013). Unfortunately, the
418 description of these two species did not investigate growth on acetate. The *pfor* gene is
419 distantly related to the archaeal multi-subunit-type used by other Thermotogae (Ragsdale,
420 2003) and almost all its close homologs fall within the *Clostridia* (Supplementary Figure
421 S5). Genes having 97-99% identity to *pfor* from *M. infera*, and 83-85% identity to the *M.*
422 *prima* homolog, were found in both the metagenome and metatranscriptome published by
423 Nobu *et al.*; 2015) (locus tag JGI12104J13512_10052834 and
424 JGI11944J13513_10066464) but were not included in their model. We propose that
425 PFOR may work with the acetate kinase (Theba_0428 in *M. prima*) and
426 phosphotransacetylase (Theba_0782 in *M. prima*) found in all Thermotogae to enable
427 *Mesotoga* to grow on acetate. At high extracellular acetate concentrations we suggest that
428 PFOR shifts the balance favoring the production of pyruvate from acetyl-CoA (i.e. serves
429 as an acetate switch Wolfe, 2005).

M. prima grows optimally on xylose, a sugar fermented by many Thermotogae (Bhandari and Gupta, 2014). The D-xylose utilization pathway is similar to that observed in Firmicutes (Gu *et al.*, 2010)(Figure 3). Several possible xylulose kinases genes were found co-localized with genes encoding xylosidases, sugar transporters and kinases, suggesting their synergetic activities in xylan hydrolysis, xylose import and utilization.

***Mesotoga-specific genes related to O_2 exposure**:* Several *Mesotoga*-specific genes are predicted to be involved in oxygen-radical defense (Supplementary Table S6). One of the most highly conserved genes across all the *Mesotoga* genomes (Theba_1553; average pairwise identity 96.3%) shows similarity to peroxiredoxin and alkyl hydroperoxide reductase domain-encoding genes. Moreover, a catalase gene (Theba_0075) is found in all isolate genomes except those from oil reservoir H.

The higher abundance of oxygen-radical defense may be linked to the lower growth temperatures of *Mesotoga* versus thermophilic Thermotogae. O_2 solubility in water is greater and free radicals are stabilized at low temperatures, and organisms living at low temperatures are therefore exposed to higher concentrations of reactive oxygen species (Piette *et al.*, 2010). It should be noted that the transcriptome of *M. prima* grown anaerobically revealed that some of the genes possibly involved in O_2 or H_2O_2 defense (e.g., catalase) were highly expressed (top 5% of expressed genes; Supplementary Table S1 and S6), suggesting that these genes may have additional or alternative functions under anaerobic conditions. Further investigation is needed to clarify the transcriptional responses of these genes and identify the targets of their enzymes.

Concluding remarks.

Our genomic analyses show that the lower growth temperature of *Mesotoga* spp. compared to the hyperthermophilic *Thermotoga* likely has significantly influenced *Mesotoga* phylogeography, with geographic separation having a greater influence than genetic separation, possibly due to the damaging effects of oxygen exposure during dispersal. Whether this is a general feature of strictly anaerobic organisms remains to be resolved. There is also some indication of possible eco-type differentiation among the *Mesotoga* lineages, with the US lineage being common in communities degrading aromatic pollutants (PCB, tolulene) and the A lineage in hydrocarbon-impacted sites. However, for both of these lineages, inspection of metagenomes revealed they are not restricted to these environments.

The ecological role of *Mesotoga in situ* may differ from their thermophilic relatives. For instance, hydrogen-producing *Thermotoga* spp. have been shown to grow in syntrophy with hydrogenotrophic methanogens (e.g., Johnson *et al.*, 2005) but this is likely not the case for *Mesotoga* that produce only trace amounts or no detectable H_2. Supporting this proposal, we were unable to establish co-cultures of *M. prima* and a hydrogenotrophic methanogen (not shown). Instead Fadhlaoui *et al.* (2017) showed that *Mesotoga* spp. prefer to grow in syntrophy with hydrogenotrophic sulfate-reducing bacteria. This, together with the ability to both produce and consume acetate, suggests that *Mesotoga* will assume different environmental roles than their thermophilic relatives, for instance by supporting the growth of sulfate reducers rather than methanogens. An interesting question is whether they also grow syntrophically with other common hydrogenotrophic organisms in their niches, such as organohalide-respiring Dehalococcoides (e.g. Fagervold *et al.*, 2007). Finally, the large amounts of H_2S

476 produced by *Mesotoga* could have detrimental effects on oil reservoirs, production
477 facilities and pipelines where *Mesotoga* is commonly found. Monitoring the presence of
478 *Mesotoga* spp. in addition to the more commonly targeted sulfate reducers in these
479 industrial environments (Lee *et al.*, 1995) may be informative and valuable.

480

481 **Acknowledgements**

482 This work was supported by a Norwegian Research Council award (project no.
483 180444/V40) to C.L.N. and by a Genome Canada grant (Hydrocarbon Metagenomics
484 Project) to J.F. The work of IVK was supported by the Federal Agency for Scientific
485 Organiations of Russia. We thank Dr. Alexander Lebedinsky for constructive criticism
486 and helpful suggestions.

487

488 **References**

489 Andreani NA, Hesse E, Vos M. (2017). Prokaryote genome fluidity is dependent on
490 effective population size. *The ISME Journal* **11**: 1719–1721.

491 Balk M, Weijma J, Stams AJM. (2002). *Thermotoga lettingae* sp. nov., a novel
492 thermophilic, methanol-degrading bacterium isolated from a thermophilic anaerobic
493 reactor. *Int J Syst Evol Micr* **52**: 1361–1368.

494 Bhandari V, Gupta RS. (2014). The Phylum Thermotogae. In: *The Prokaryotes*. Springer,
495 Berlin, Heidelberg: Berlin, Heidelberg, pp 989–1015.

496 Boileau C, Auria R, Davidson S, Casalot L, Christen P, Liebgott P-P, *et al.* (2016).
497 Hydrogen production by the hyperthermophilic bacterium *Thermotoga maritima* part I:
498 effects of sulfured nutriments, with thiosulfate as model, on hydrogen production and
499 growth. *Biotechno Biofuels* **9**: 269.

500 Chakraborty A, Ellefson E, Li C, Gittins D, Brooks JM, Bernard BB, *et al.* (2018).
501 Thermophilic endospores associated with migrated thermogenic hydrocarbons in deep
502 Gulf of Mexico marine sediments. *The ISME Journal* 1–12.

503 Choudoir MJ, Panke-Buisse K, Andam CP, Buckley DH. (2017). Genome Surfing As
504 Driver of Microbial Genomic Diversity. *Trends Microbiol* 1–13.

505 Darling AE, Mau B, Perna NT. (2010). ProgressiveMauve: Multiple Genome Alignment

with Gene Gain, Loss and Rearrangement. *PLoS ONE* **5**: e11147.

Dipippo JL, Nesbø CL, Dahle H, Doolittle WF, Birkland N-K, Noll KM. (2009). *Kosmotoga olearia* gen. nov., sp. nov., a thermophilic, anaerobic heterotroph isolated from an oil production fluid. *Int J Syst Evol Micr* **59**: 2991–3000.

Eckford RE, Fedorak PM. (2002). Planktonic nitrate-reducing bacteria and sulfate-reducing bacteria in some western Canadian oil field waters. *J Ind Microbiol Biotechnol* **29**: 83–92.

Fadhlaoui K, Hania WB, Armougom F, Bartoli M, Fardeau M-L, Erauso G, *et al.* (2017). Obligate sugar oxidation in *Mesotoga* spp., phylum Thermotogae, in the presence of either elemental sulfur or hydrogenotrophic sulfate-reducers as electron acceptor. *Environ Microbiol* **20**: 281–292.

Fagervold SK, May HD, Sowers KR. (2007). Microbial reductive dechlorination of Aroclor 1260 in Baltimore Harbor sediment microcosms is catalyzed by three phylotypes within the Phylum Chloroflexi. *Appl Environ Microbiol* **73**: 3009–3018.

Fondi M, Karkman A, Tamminen MV, Bosi E, Virta M, Fani R, *et al.* (2016). 'Every Gene Is Everywhere but the Environment Selects': Global Geolocalization of Gene Sharing in Environmental Samples through Network Analysis. *Genome Biol Evol* **8**: 1388–1400.

Fowler SJ, Dong X, Sensen CW, Suflita JM, Gieg LM. (2012). Methanogenic toluene metabolism: community structure and intermediates. *Environ Microbiol* **14**: 754–764.

Gieg LM, Kolhatkar RV, McInerney MJ, Tanner RS, Harris SH, Sublette KL, *et al.* (1999). Intrinsic bioremediation of petroleum hydrocarbons in a gas condensate-contaminated aquifer. *Environ Sci Technol* **33**: 2550–2560.

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Micr* **57**: 81–91.

Gu Y, Ding Y, Ren C, Sun Z, Rodionov DA, Zhang W, *et al.* (2010). Reconstruction of xylose utilization pathway and regulons in Firmicutes. *BMC Genomics* **11**: 255.

Hagen WR, Silva PJ, Amorim MA, Hagedoorn PL, Wassink H, Haaker H, *et al.* (2000). Novel structure and redox chemistry of the prosthetic groups of the iron-sulfur flavoprotein sulfide dehydrogenase from *Pyrococcus furiosus*; evidence for a [2Fe-2S] cluster with Asp Cys). *JBIC* **5**: 527-534

Hania WB, Fadhlaoui K, Brochier-Armanet C, Persillon C, Postec A, Hamdi M, *et al.* (2015). Draft genome sequence of *Mesotoga strain* PhosAC3, a mesophilic member of the bacterial order Thermotogales, isolated from a digestor treating phosphogypsum in Tunisia. *Stand Genomic Sci* **10**:12.

542 Hania WB, Ghodbane R, Postec A, Brochier-Armanet C, Hamdi M, Fardeau M-L, *et al.*
543 (2011). Cultivation of the first mesophilic representative ('mesotoga') within the order
544 Thermotogales. *Systematic and Applied Microbiology* **34**: 581–585.

545 Hania WB, Postec A, Aüllo T, Ranchou-Peyruse A, Erauso G, Brochier-Armanet C, *et al.*
546 (2013). *Mesotoga infera* sp. nov., a mesophilic member of the order Thermotogales,
547 isolated from an underground gas storage aquifer. *Int J Syst Evol Micr* **63**: 3003–3008.

548 Head IM, Gray ND, Larter SR. (2014). Life in the slow lane; biogeochemistry of
549 biodegraded petroleum containing reservoirs and implications for energy recovery and
550 carbon management. *Front Microbiol* **5**: Article 566.

551 Holoman TR, Elberson MA, Cutter LA, May HD, Sowers KR. (1998). Characterization
552 of a defined 2,3,5, 6-tetrachlorobiphenyl-ortho-dechlorinating microbial community by
553 comparative sequence analysis of genes coding for 16S rRNA. *Appl Environ Microbiol*
554 **64**: 3359–3367.

555 Hu P, Tom L, Singh A, Thomas BC, Baker BJ, Piceno YM, *et al.* (2016). Genome-
556 resolved metagenomic analysis reveals roles for Candidate Phyla and other microbial
557 community members in biogeochemical transformations in oil reservoirs. *mBio* **7**:
558 e01669–15–12.

559 Hulecki JC, Foght JM, Gray MR, Fedorak PM. (2009). Sulfide persistence in oil field
560 waters amended with nitrate and acetate. *J Ind Microbiol Biotechnol* **36**: 1499–1511.

561 Huson DH, Bryant D. (2006). Application of phylogenetic networks in evolutionary
562 studies. *Mol Biol Evol* **23**: 254–267.

563 Johnson MR, Conners SB, Montero CI, Chou CJ, Shockley KR, Kelly RM. (2005). The
564 *Thermotoga maritima* phenotype is impacted by syntrophic interaction with
565 *Methanococcus jannaschii* in hyperthermophilic coculture. *Appl Environ Microbiol* **72**:
566 811–818.

567 Jolley KA, Wilson DJ, Kriz P, McVean G, Martin MCJ (2004). The influence of
568 mutation, recombination, population history, and selection on patterns of genetic diversity
569 in *Neisseria meningitidis*. *Mol Biol Evol* **22**: 562–569.

570 Koskella B, Brockhurst MA. (2014). Bacteria-phage coevolution as a driver of ecological
571 and evolutionary processes in microbial communities. *FEMS Microbiol Rev* **38**: 916–931.

572 Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, *et al.* (2009).
573 Circos: An information aesthetic for comparative genomics. *Genome Research* **19**: 1639–
574 1645.

575 Laing C, Buchanan C, Taboada EN, Zhang Y, Kropinski A, Villegas A, *et al.* (2010).
576 Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core
577 and accessory genomic regions. *BMC Bioinformatics* **11**: 461.

578 Lee W, Lewandowski Z, Nielsen PH, Hamilton WA. (1995). Role of sulfate-reducing
579 bacteria in corrosion of mild steel: A review. *Biofouling* **8**: 165–194.

580 L'Haridon S, Jiang L, Alain K, Chalopin M, Rouxel O, Beauverger M, *et al.* (2013).
581 *Kosmotoga pacifica* sp. nov., a thermophilic chemoorganoheterotrophic bacterium
582 isolated from an East Pacific hydrothermal sediment. *Extremophiles* **18**: 81–88.

583 Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Pillay M, *et al.* (2014).
584 IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic*
585 *Acids Res* **42**: D560–7.

586 Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. (2010). RDP3: a
587 flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**:
588 2462–2463.

589 McInerney JO, McNally A, O'Connell MJ. (2017). Why prokaryotes have pangenomes.
590 *Nat Microbiol* **2**: 1–5.

591 McVean G, Awadalla P, Fearnhead P. (2002). A Coalescent-Based method for detecting
592 and estimating recombination from gene sequences. *Genetics* **160**: 1231–1241.

593 Mulkidjanian AY, Galperin MY, Makarova KS, Wolf YI, Koonin EV. (2008).
594 Evolutionary primacy of sodium bioenergetics. *Biology Direct* **3**: 13.

595 Müller V, Imkamp F, Biegel E, Schmidt S, Dilling S. (2008). Discovery of a
596 Ferredoxin:NAD+-Oxidoreductase (Rnf) in *Acetobacterium woodii*. *Ann NY Acad Sci*
597 **1125**: 137–146.

598 Nesbø CL, Bradnan DM, Adebusuyi A, Dlutek M, Petrus AK, Foght J, *et al.* (2012).
599 *Mesotoga prima* gen. nov., sp. nov., the first described mesophilic species of the
600 Thermotogales. *Extremophiles* **16**: 387–393.

601 Nesbø CL, Dlutek M, Zhaxybayeva O, Doolittle WF. (2006). Evidence for existence of
602 'mesotogas,' members of the order Thermotogales adapted to low-temperature
603 environments. *Appl Environ Microbiol* **72**: 5061–5068.

604 Nesbø CL, Kumaraswamy R, Dlutek M, Doolittle WF, Foght JM. (2010). Searching for
605 mesophilic Thermotogales bacteria: 'mesotogas' in the wild. *Appl Environ Microbiol* **76**:
606 4896–4900.

607 Nesbø CL, S Swithers K, Dahle H, Haverkamp THA, Birkeland N-K, Sokolova T, *et al.*
608 (2015). Evidence for extensive gene flow and *Thermotoga* subpopulations in subsurface
609 and marine environments. *The ISME Journal* **9**: 1532–1542.

610 Nobu MK, Narihiro T, Rinke C, Kamagata Y, Tringe SG, Woyke T, *et al.* (2015).
611 Microbial dark matter ecogenomics reveals complex synergistic networks in a
612 methanogenic bioreactor. *The ISME Journal*. **9**: 1710-1722.

613 Nouailler M, Morelli X, Bornet O, Chetrit B, Dermoun Z, Guerlesquin F. (2006).
614 Solution structure of HndAc: a thioredoxin-like domain involved in the NADP-reducing
615 hydrogenase complex. *Protein Sci* **15**: 1369–1378.

616 Petrus AK, Swithers KS, Ranjit C, Wu S, Brewer HM, Gogarten JP, *et al.* (2012). Genes
617 for the major structural components of Thermotogales species' togas revealed by
618 proteomic and evolutionary analyses of OmpA and OmpB homologs. *PLoS ONE* **7**:
619 e40236.

620 Piette F, D'Amico S, Struvay C, Mazzucchelli G, Renaut J, Tutino ML, *et al.* (2010).
621 Proteomics of life at low temperatures: trigger factor is the primary chaperone in the
622 Antarctic bacterium *Pseudoalteromonas haloplanktis* TAC125. *Mol Microbiol* **76**: 120–
623 132.

624 Pollo SMJ, Adebusuyi AA, Straub TJ, Foght JM, Zhaxybayeva O, Nesbø CL. (2017).
625 Genomic insights into temperature-dependent transcriptional responses of Kosmotoga
626 olearia, a deep-biosphere bacterium that can grow from 20 to 79 °C. *Extremophiles* **21**:
627 963–979.

628 Pollo SMJ, Zhaxybayeva O, Nesbø CL. (2015). Insights into thermoadaptation and the
629 evolution of mesophily from the bacterial phylum *Thermotogae*. *Can J Microbiol* **61**:
630 655–670.

631 Ragsdale SW. (2003). Pyruvate Ferredoxin Oxidoreductase and Its Radical Intermediate
632 †. *Chem Rev* **103**: 2333–2346.

633 Ravot G, Ollivier B, Magot M, Patel BKC, Fardeau ML, Garcia J-L. (1995). Thiosulfate
634 reduction, an important physiological feature shared by members of the order
635 Thermotogales. *Appl Environ Microbiol* **61**: 2053–2055.

636 Reysenbach A-L, Liu Y, Lindgren AR, Wagner ID, Sislak CD, Mets A, *et al.* (2013).
637 *Mesoaciditoga lauensis* gen. nov., sp. nov., a moderately thermoacidophilic member of
638 the order Thermotogales from a deep-sea hydrothermal vent. *Int J Syst Evol Micr* **63**:
639 4724–4729.

640 Schaefer BF. (2005). GEOCHEMISTRY: When Do Rocks Become Oil? *Science* **308**:
641 1267–1268.

642 Schut GJ, Boyd ES, Peters JW, Adams MWW. (2013). The modular respiratory
643 complexes involved in hydrogen and sulfur metabolism by heterotrophic
644 hyperthermophilic archaea and their evolutionary implications. *FEMS Microbiol Rev* **37**:
645 182–203.

646 Selby D. (2005). Direct radiometric dating of hydrocarbon deposits using Rhenium-
647 Osmium isotopes. *Science* **308**: 1293–1295.

648 Stoffels L, Krehenbrink M, Berks BC, Unden G. (2012). Thiosulfate reduction in
649 *Salmonella enterica* is driven by the proton motive force. *J Bact* **194**: 475–485.

650 Tan B, Jane Fowler S, Laban NA, Dong X, Sensen CW, Foght J, *et al.* (2015).
651 Comparative analysis of metagenomes from three methanogenic hydrocarbon-degrading
652 enrichment cultures with 41 environmental samples. *The ISME Journal* **9**: 2028–2045.

653 Vignais PM, Billoud B. (2007). Occurrence, classification, and biological function of
654 hydrogenases: an overview. *Chem Rev* **107**: 4206–4272.

655 Voordouw G, Grigoryan AA, Lambo A, Lin S, Park HS, Jack TR, *et al.* (2009). Sulfide
656 remediation by pulsed injection of Nitrate into a low temperature Canadian heavy oil
657 reservoir. *Environ Sci Technol* **43**: 9512–9518.

658 Wachter J, Hill S. (2016). Positive selection pressure drives variation on the surface-
659 exposed variable proteins of the pathogenic *Neisseria*. *PLoS ONE* **11**: e0161348.

660 Wang S, He J. (2013). Phylogenetically distinct bacteria involve extensive dechlorination
661 of Aroclor 1260 in sediment-free cultures. *PLoS ONE* **8**: e59178.

662 Wilhelms A, Larter SR, Head I, Farrimond P, di-Primio R, Zwach C. (2001).
663 Biodegradation of oil in uplifted basins prevented by deep-burial sterilization. *Nature*
664 **411**: 1034–1037.

665 Wolfe AJ. (2005). The acetate switch. *Microbiology and Molecular Biology Reviews* **69**:
666 12–50.

667 Zhaxybayeva O, Swithers KS, Foght J, Green AG, Bruce D, Detter C, *et al.* (2012).
668 Genome Sequence of the mesophilic thermotogales bacterium *Mesotoga prima* MesG1.
669 Ag. 4.2 reveals the largest thermotogales genome to date. *Genome Biol Evol* **4**: 700–708.

670

671

672 **Figure legends**

673 **Figure 1. Phylogenetic relationships among *Mesotoga* genomes based on (a)**
674 **16SrRNA genes, (b) core SNPs and (c) presence/absence of shared 500-bp genomic**
675 **fragments.** The 16S rRNA maximum likelihood phylogeny was estimated using RAxML
676 in Geneious v 10. For networks shown in (b) and (c), data were obtained using PanSeq
677 (Laing *et al.*, 2010). Core SNPs in (b) were required to be present in 14 of 18 genomes
678 (including SAGs), and genomic fragments were considered shared if they were at least
679 70% identical. The network in (c) was constructed using only genomes from isolates;
680 shared fragments were required to be present in all 9 genomes and be at least 70%
681 identical in sequence. Networks were calculated in SplitsTree using NeighborNet
682 algorithm (Huson and Bryant, 2006) from uncorrected distances. The three lineages are
683 indicated by W, US and A.

684

685 **Figure 2. Visualization of recombination events detected among *Mesotoga* genomes**
686 **from different lineages.** The genomes are color-coded according to lineage and arranged
687 on the circumference of the circle; W lineage, blue; US lineage, orange; A lineage, green.
688 The recombination events with predicted donor and recipient are shown as lines
689 connecting the two genomes and the locations of recombined regions, where line color
690 reflects the donor lineage and the width of the line is proportional to the length of the
691 recombinant region. Predicted events were required to be significant in at least 3 of 4
692 algorithms in RDP (Martin *et al.*, 2010). The diagram was generated using Circos
693 Version circos-0.69 (Krzywinski *et al.*, 2009).

694

695 **Figure 3. Model of energy generation pathway in *Mesotoga prima* during growth on**
696 **xylose and thiosulfate.** Glucose and xylose poly- and oligosaccharides are hydrolyzed by
697 various intracellular and interstitial glycosidases (GHs). Glucose oxidation occurs via the
698 glycolytic Embden-Meyerhof-Parnas pathway, whereas xylose is utilized via xylose
699 isomerase (*XylA*, Theba_1394), xylulose kinase (*XylB*, Theba_1395, Theba_2230, Theba

2429, Theba 2518, Theba 2544, Theba 2588), ribulose phosphate 3-epimerase (Theba_0639) and enzymes of the pentose-phosphate pathway. Specifically, xylose isomerase converts D-xylose to D-xylulose, which is phosphorylated by the set of xylulose kinases to D-xylulose 5-phosphate, and further to ribulose 5-phosphate by the ribulose-phosphate 3-epimerase. Both xylulose 5-phosphate and ribulose 5-phosphate produced by this pathway are common metabolic intermediate in the pentose-phosphate pathway. The xylose isomerase Theba_1394 was among the most highly transcribed genes during cultivation of *M. prima* on xylose and thiosulfate (Supplementary Table S5). Acetyl-CoA formation occurs by means of pyruvate-ferredoxin oxidoreductase (PFOR, Theba_1954). In the possible case of growth on acetate, its activation occurs by means of acetate kinase (ACKA, Theba_0428) and phosphotransacetylase (PTA, Theba_0782), acting in reverse. The model includes gene products hypothesized to be involved in thiosulfate reduction. $\underline{Na^+}$ refers to Na+ ions involved in generating the generating the sodium motive force. A: The FeFe hydrogenase (Theba_0443 and Theba_0461 – 0465) reduces NADH to form H_2, which is used as an electron donor for thiosulfate reduction catalyzed by SudAB (Theba_0076, Theba_0077). Mbx (Theba_1796-1808) and/or Rnf (Theba_1343-1348) complexes provide additional NADH along with the oxidation of excessive reducing equivalents (Fd_{red}) and generation of a sodium-motive force. B and C: other possible scenarios of H_2 oxidation and thiosulfate reduction.

Table 1. List of genomes analyzed. All genomes, except those of *Mesotoga prima, M. prima* PhosAc3 and *Mesotoga infera,* were sequenced as part of the current study.

Name and Source	Short Name	Genome Size	% GC	Ref. for description of sample site / accession no. in Genbank	% completeness of SAG[a]
Isolates					
Produced water from oil field B near Brooks, Alberta, Canada[b]				(Hulecki *et al.*, 2009)	
Mesotoga sp. Brooks.08.YT.4.2.5.1[c]	BR5.1	2,957,195	45.9	AYTX01000000	
Mesotoga sp. Brooks.08.YT.4.2.5.2	BR5.2	2,953,308	45.9	JPGZ00000000	
Mesotoga sp. Brooks.08.YT.4.2.5.4[c]	BR5.4	3,002,147	45.9	ATCT01000000	
Mesotoga sp. Brooks.08.YT.105.5.1	BR105.1	2,992,699	45.9	AYTW01000000	
Mesotoga sp. Brooks.08.YT.105.6.4	BR105.4	3,205,299	45.9	JWIM00000000	
Free water knockout fluids from oil field H near Stettler, Alberta[d]				(Eckford and Fedorak, 2002)	
Mesotoga sp. HF07.pep.5.2[c]	HF5.2	2,838,813	45.3	JFHJ01000000	
Mesotoga sp. HF07.pep.5.3	HF5.3	2,934,282	45.3	AYTV01000000	
Mesotoga sp. HF07.pep.5.4	HF5.4	2,968,642	45.3	JFHM01000000	
Sediments from Baltimore Harbour, Maryland, USA				(Holoman *et al.*, 1998)	
Mesotoga prima MesG1.Ag.4.2[e]	*M.prima*	2,974,229	45.5	NC_017934	
Mesotoga sp. BH458.6.3.2.1[f]	BH458	3,234,409	45.7	JFHL01000000	
Wastewater treatment plant, Tunisia				(Hania *et al.*, 2011)	
Mesotoga prima PhosAc3	PhosAc3	3,108,267	45.2	NZ_CARH01000000	

(continued)

Genomes assembled from single cell amplified genomes (SAGs)					
Produced water from oil field E near Medicine Hat, Alberta[g]				(Voordouw *et al.*, 2009)	
Mesotoga sp. 3PWK154PWL11	SC_PW.1	876,625	46.8	JMRN01000000	21%
Mesotoga sp. 3PWM13N19	SC_PW.2	1,886,634	45.8	JMRM01000000	78%
Mesotoga sp. 4PWA21	SC_PW.3	1,541,163	45.9	JMQL01000000	34%
Oil sands tailings pond sediments near Fort McMurray, Alberta				(Tan *et al.*, 2015)	
Mesotoga sp. NapDC	SC_NapDC	1,885,291	45.8	JNFM01000000	88%
Mesotoga sp. NapDC2	SC_NapDC2	1,337,305	45.6	JQSC01000000	53%
Mesotoga sp. NapDC3	SC_NapDC3	1,828,922	45.6	JWIP00000000	66%
Contaminated aquifer fluids from Colorado, USA				(Fowler *et al.*, 2012; Gieg *et al.*, 1999)	
Mesotoga sp. TolDC[h]	SC_TOLDC	2,257,992	46.1	AYSI01000000	74%

[a] Completeness of single cell genomes was calculated based on HMM hits to 119 single copy marker genes (see Supplementary Material).

[b]Belongs geologically to the Glauconitic formation.

[c]These genomes were sequenced using IonTorrent PGM; all other genomes and single cells were sequenced using Illumina MiSeq

[d]Belongs geologically to the Upper Mannville Group – Cretaceous age.

[e] The *M. prima* genome was sequenced by Zhaxybayeva *et al.* (2012).

[f] This is a sister-culture of the enrichment culture that yielded *M. prima* MesG1.Ag.4.2 (Nesbø *et al.*, 2006; 2012).

[g] Belongs geologically to the Western Canadian Sedimentary Basin.

[h] In addition to the SAG sequences, 39 *Mesotoga* fosmid clones prepared from the same culture were included in the assembly (http://hmp.ucalgary.ca/HMP/metagenomes/isolates.html).

Table 2. *Mesotoga* sequences recovered from publicly available metagenomes. For the sequences obtained from the IMG (JGI) database, only sequences classified as Thermotogae were downloaded. The predominant *Mesotoga* lineage in each metagenome is shown in boldface. Metagenomes dominated by sequences similar to *Mesotoga infera* were not included.

Metagenome origin[a]	IMG, Genbank or SRA accession nos.	*Mesotoga* sequences (no. contigs)	*Mesotoga* sequences with best match to lineage A, W or US (no. contigs)	% average pairwise identity (range)
Alberta				
MLSB tailings pond	IMG: 26785	491,657 bp (1,707)	**A: 485,638 bp (1,694)**	**99.8% (98.5-100%)**
			W: 2,455 bp (8)	93.9% (88.7-99.6)
			US: 634 bp (2)	86.4% (89.7-95%)
Oil reservoir E	IMG: 15764	5,190,293 bp (4,833)	**A: 3,137,228 bp (3,259)**	**98.6% (71.3-100%)**
			W: 195,159 bp (317)	94.5% (79.7-100%)
			US: 1,693,057 bp (1,308)	**95.3% (82.2-100%)**
USA				
California				
Alameda Naval Air Station: Soil contaminated with Chloroethene	IMG: 5776	407,588 bp (453)	A: 17,955 bp (23)	92% (85.3-99.5%)
			W: 47,017 bp (61)	93.4% (83-99.1%)
			US: 297,011 bp (360)	**94.2%(83.6-100%)**
Blank Spring hotspring sediment	IMG: 94476	4,060,664 bp (2,037)	**A: 920,911 bp (924)**	**91.6% (64.3-100%)**
			W: 226,161 bp (176)	85.4% (64.3-99.1%)
			US: 2,025,524 bp (825)	**89.3% (66.1-99.6 %)**
Long Beach Municipal wastewater AD_UKC097	IMG: 89744	2,339,863 bp (1,810)	**A: 390,020 bp (588)**	**96.1% (65-100%)**
			W: 710,377 bp (914)	**94.1% (64.3-100%)**
			US: 202,508 bp (207)	88.4% (63.6-100%)

(continued)

Illinois				
Decatur municipal wastewater AD_UKC034[b]	IMG: 89745	4,031,397 bp (2,218)	**A: 2,085,927 bp (1,162)** **W: 820,163 bp (588)** US: 271,756 bp (260)	**96.5% (65.1-100%)** **92.7% (64.6-100%)** 89.0% (65.0-100%)
New York Sulfidogenic MTBE-NYH[c]	IMG: 62988	5,195,092 bp[i] (1,727)	**A: 421,552 bp (261)** W: 209,518 bp (218) **US:1,825,787 bp (450)**	**85.8% (63.8-99.2%)** 84.6% (64.5-99.8%) **89.2% (65.9-100%)**
New Jersey Methanogenic MTBE-AKM[d]	IMG: 62224	2,464,953 bp (4,206)	A: 31,663 bp (104) W: 122,392 bp (320) **US: 1,154,398 bp (2,706)**	87.2% (66.8-100%) 93.8% (65.0-100%) **98.0% (66.5-100%)**
Sulfidogenic MTBE-AKS2[e]	IMG: 62223	4,971,880 (3,215)	A: 152,359 bp (276) W: 240,489 bp (307) **US: 2,665,479 bp (1,575)**	86.3% (65.9-99.3%) 86.6% (65.6-100%) **94.3% (65.8-100%)**
Boston Wastewater AD_UKC077[f]	IMG: 89805	5,651,655[i] (1,386)	A: 149,486 bp (212) **W: 1,379,334 bp (326)** US: 141,168 bp (150)	80.2% (63.4-100%) **88.2% (63.7-100%)** 77.6% (64.4-99.7%)
Hong Kong				
Wastewater AD_UKC109[f] 2015-03-06	IMG: 89888	4,071,190 bp (4,096)	A: 415,727 bp (615) **W: 1,189,447 bp (2,476)** US: 632,042 bp(902)	92.8% (64.7-100%) **93.6% (64.3-100%)** 91.9% (63-100%)
Wastewater AD_UKC119[f] 2015-01-26	IMG: 89894	5,462,335 bp (2,499)	**A: 2,347,962 bp (853)** W: 586,387 bp (497) **US: 1,521,681(1,071)**	**95.7% (65.4-100%)** 89.4% (64.3-100%) **93.0% (63.9-100%)**

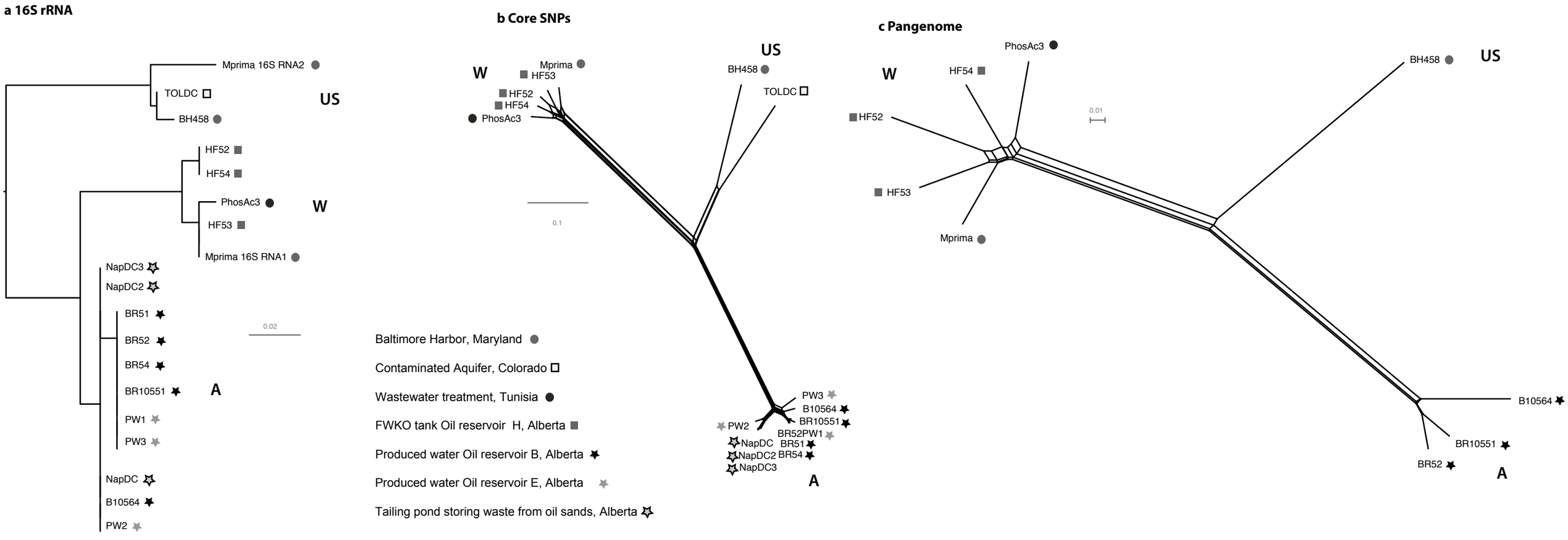
(continued)

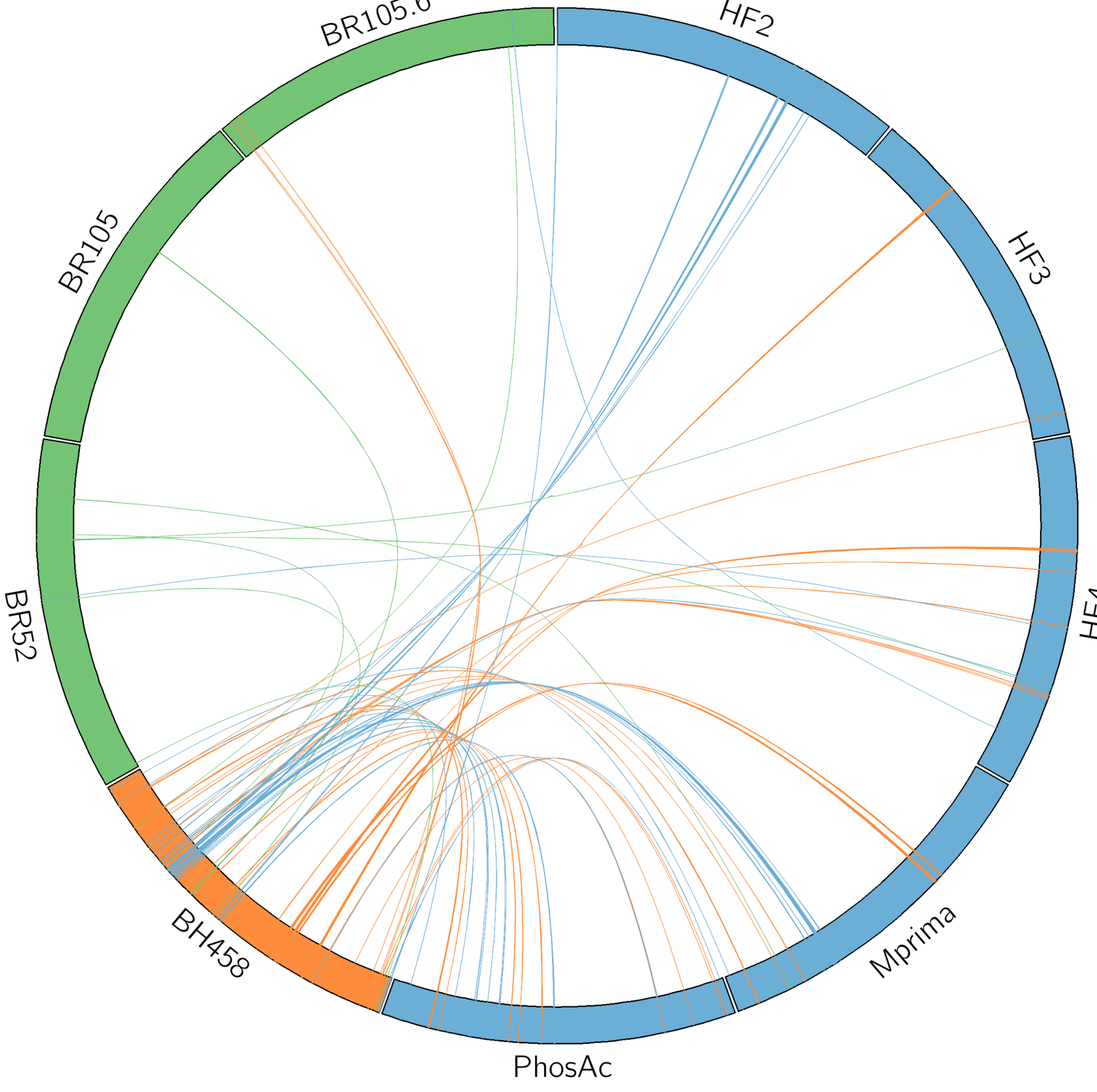
Metagenome-assembled genomes (MAGs)

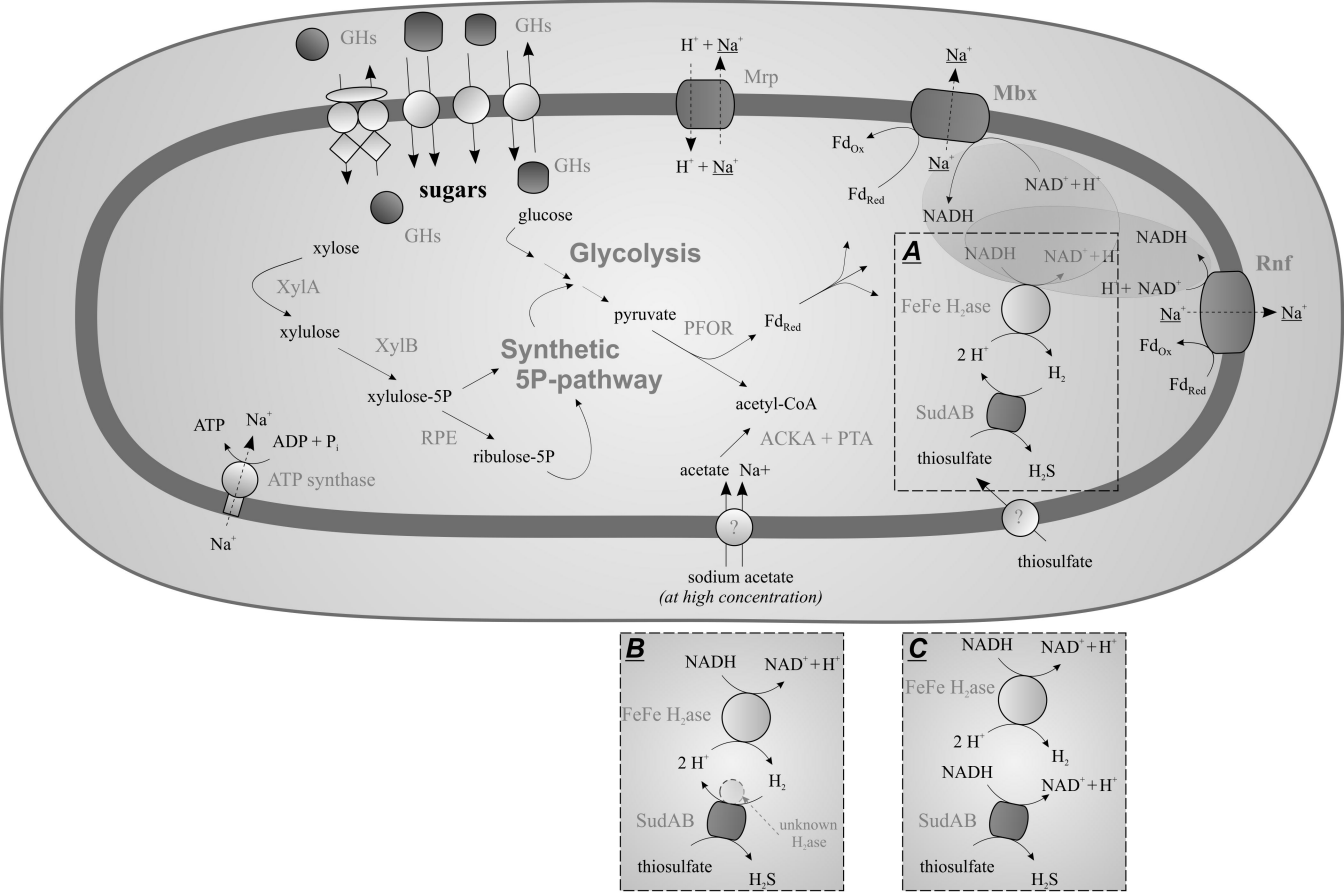
Alaska Oil reservoir LGGP01[g]	Genbank: GCA_001508515	1,712,609 bp (440)	**A: 1,470,927 bp (384)** W: 80,31 bp (30) US: 47,775 bp (21)	**98.5% (71.3-100%)** 92.3% (70.0-99.6%) 88.9% (71.9-100%)
LGGH01[g]	GCA_001508435	1,225,111 bp (267)	A: 48,008 bp (13) W: 63,609 bp (21) **US: 1,009,462 bp (233)**	89.9% (78.8-99.6%) 92.5% (80.4-100%) **94.5% (80.9-99.6)**
LGGW01[g]	GCA_001509115	1,622,264 bp (264)	A: 85,486 bp (20) W: 104,756 bp (25) **US: 1,139,439 (211)**	87.3% (67.8-99.6%) 92.3% (82.6-99.9%) **94.5% (83.6-99.8%)**
California Anaerobic digester in Oakland[h]	IMG: 81407 (Unclassified Thermotogales bacterium Bin 13)	3,480,910 bp (395)	**A: 2,287,852 bp (247)** W: 109,011 bp (48) US: 118,885 bp (33)	**93.6% (64.5-100%)** 83.1% (64.4-100%) 79.1% (65.8-96.5)
China PCB-fed mixed *Dehalococcoides* culture CG1 from sand and silt near Liangjiang River[i]	SRA: SRX392467	2,727,841bp (379)	**A: 2,226,7034 bp (345)** W: 22,520 bp (14) US: 40,010 bp (20	**98.0% (84.6-100%)** 90.2% (68.2-97.4%) 90.7% (78.2-99.8%)

a) *Mesotoga* sequences were identified by performing blastn searches using a database containing all the *Mesotoga* spp. genomes listed in Table 1. We used word size =11 and expected = e^{-10}. Sequences with matches were then sorted according to the *Mesotoga* lineage with the best match (A, W or US; see main text).
b) Nine metagenomes and three transcriptomes are available from the same system; only one was selected as representative.
c) Sulfidogenic MTBE-degrading enrichment culture microbial communities from New York harbour sediments
d) Methanogenic MTBE-degrading enrichment culture microbial communities from Arthur Kill sediments

e) Sulfidogenic MTBE-degrading enrichment culture microbial communities from Arthur Kill sediments. Two very similar metagenomes are available (Figure S3b); only one is included here.
f) Wastewater treatment anaerobic digesters. Additional metagenomes with similar *Mesotoga* lineage composition are available from this site and the ones included here were chosen as representatives.
g) Metagenome assembled genomes from (Hu *et al.*, 2016). LGGP01 and LGGH01 were from oil reservoir sample SB1 and LGGW01was from oil reservoir sample SB2.
h) Sludge microbial communities from wastewater, phosphite and CO_2-enriched.
i) Genome extracted and assembled by us


a 16S rRNA
Mprima 16S RNA2
TOLDC
BH458
US
HF52
HF54
PhosAc3
HF53
W
Mprima 16S RNA1
NapDC3
NapDC2
BR51
BR52
BR54
BR10551
A
PW1
PW3
NapDC
B10564
PW2
0.02
b Core SNPs
W
Mprima
HF53
HF52
HF54
PhosAc3
US
BH458
TOLDC
0.1
PW3
B10564
BR10551
BR52PW1
PW2
NapDC
BR51
NapDC2
BR54
NapDC3
A
Baltimore Harbor, Maryland
Contaminated Aquifer, Colorado
Wastewater treatment, Tunisia
FWKO tank Oil reservoir H, Alberta
Produced water Oil reservoir B, Alberta
Produced water Oil reservoir E, Alberta
Tailing pond storing waste from oil sands, Alberta
c Pangenome
PhosAc3
W
HF54
HF52
HF53
Mprima
0.01
BH458
US
B10564
BR10551
BR52
A


BR105.6
HF2
BR105
HF3
BR52
HF4
BH458
Mprima
PhosAc


GHs
GHs
GHs
GHs
sugars
GHs
H+ + Na+
Mrp
H+ + Na+
Na+
Mbx
Fd_Ox
Fd_Red
Na+
NAD+ + H+
NADH
glucose
xylose
XylA
xylulose
XylB
xylulose-5P
RPE
ribulose-5P
Glycolysis
Synthetic
5P-pathway
pyruvate
PFOR
Fd_Red
acetyl-CoA
ACKA + PTA
acetate Na+
A
NADH
NAD+ + H+
FeFe H2ase
2 H+
H2
SudAB
thiosulfate
H2S
NADH
H+ + NAD+
Na+
Na+
Fd_Ox
Fd_Red
Rnf
ATP
Na+
ADP + Pi
ATP synthase
Na+
?
?
thiosulfate
sodium acetate
(at high concentration)
B
NADH
NAD+ + H+
FeFe H2ase
2 H+
H2
SudAB
unknown
H2ase
thiosulfate
H2S
C
NADH
NAD+ + H+
FeFe H2ase
2 H+
H2
NADH
NAD+ + H+
SudAB
thiosulfate
H2S