Genomic analysis of the mesophilic Thermotogae genus *Mesotoga* reveals phylogeographic structure and genomic determinants of its distinct metabolism

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

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

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

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

4 InnoTech Alberta, Edmonton, Alberta, Canada T6N 1E4

5 Winogradsky Institute of Microbiology, Federal Research Center of Biotechnology, Russian Academy of Sciences, Moscow, Russia

6 Norwegian Veterinary Institute, Oslo, Norway.

*Corresponding Authors: nesbo@ualberta.ca

Department of Biological Sciences, CW 405 Biological Sciences Bldg., 11455 Saskatchewan Drive , University of Alberta, Edmonton, Alberta, Canada, T6G 2E9

Running title: Comparative genomic analysis of *Mesotoga*.

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

Originality-Significance Statement

This study comprises one of the first whole-genome-based phylogeographic analyses of anaerobic mesophiles, and our data suggest that such microbes are more restricted by geography than are thermophiles (and mesophilic aerobes). This is likely to be a general trait for similar anaerobic organisms – and therefore broadly relevant to and testable in other environments. Moreover, *Mesotoga* bacteria are part of the largely understudied subsurface ecosystem that has relatively recently been recognized as a new and important biosphere. Understanding the forces responsible for the distribution of organisms in the subsurface, as well as the identification of genes responsible for *Mesotoga*'s distinct metabolism, will contribute to the understanding of these communities.

Summary

The genus *Mesotoga*, the only described mesophilic *Thermotogae* lineage, is common in mesothermic anaerobic hydrocarbon-rich environments. Besides mesophily, *Mesotoga* displays lineage-specific phenotypes, such as no or little H_2 production and dependence on sulfur-compound reduction, which may influence its ecological role. We used comparative genomics of 18 *Mesotoga* strains (pairwise 16S rRNA identity > 99%) 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. We propose that *Mesotoga* accomplish H_2 oxidation and thiosulfate reduction using a sulfide dehydrogenase and a hydrogenase-complex and that a pyruvate:ferredoxin oxidoreductase acquired from *Clostridia* is responsible for oxidizing acetate. Phylogenetic analysis revealed three distinct *Mesotoga* lineages (89.6-99.9% average nucleotide identity [ANI] within lineages, 79.3-87.6% ANI between lineages) having different geographic distribution patterns and high levels of intra-lineage recombination but little geneflow between lineages. Including data from metagenomes, phylogeographic patterns suggest that geographical separation historically has been more important for *Mesotoga* than hyperthermophilic *Thermotoga* and we hypothesize that distribution of *Mesotoga* is constrained by their anaerobic lifestyle. Our data also suggest that recent anthropogenic activities and environments (e.g., wastewater treatment, oil exploration) have expanded *Mesotoga* habitats and dispersal capabilities.

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; Hania *et al.*, 2011; Nesbø *et al.*, 2012; Hania *et al.*, 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 in 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 (Zhaxybayeva *et al.*, 2012; Pollo *et al.*, 2015). 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 high levels of geneflow among all *Thermotoga* spp. genomes investigated, and that populations from similar environments have exchanged more genes than geographically close populations from different environments. For instance, *Thermotoga* genomes from

oil reservoirs in Japan and in the North Sea, as well as from a continental hot spring in North America, have exchanged more genes through homologous recombination than they have with genomes from geographically closer marine vents. Moreover, the phylogeographic analysis of *Thermotoga* genomes suggested that oil reservoirs were colonized from subsurface populations rather than being buried with the sediments that mature into oil reservoirs reservoirs (a corollary of the paleosterilization hypothesis; (Wilhelms *et al.*, 2001)) (Nesbø *et al.*, 2015). 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 thermophilic Thermotogae. For instance, while growth of most thermophilic 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 include in our analysis *Mesotoga* sequences from publicly available metagenomes. We compare our findings from the mesophilic *Mesotoga* to the patterns previously observed in the hyperthermophilic *Thermotoga* (Nesbø *et al.*, 2015)

and infer that geographic separation has had more 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.

Results

Genome sequences

We generated draft genomes for eight newly isolated *Mesotoga* strains from two oil reservoirs (H and B) in Alberta Canada and one *Mesotoga* strain from a PCB-degrading enrichment culture from Baltimore Harbor, Maryland USA (Table 1). Seven partial single cell amplified genomes (SAGs) were obtained from cells sorted from produced water from an Albertan oil reservoir (PW), a naphtha-degrading enrichment culture inoculated with sediments from an Albertan oil sands tailings pond (NAPDC), and a toluene-degrading enrichment culture inoculated with sediments from a contaminated aquifer in Colorado USA (TOLDC). We also included in our analyses the draft genome of PhosAc3, previously isolated in Tunisia (Hania *et al.*, 2015) and the closed genome of *M. prima* (Zhaxybayeva *et al.*, 2012) from Baltimore Harbor.

Table 1

The pan-genome of the *Mesotoga* isolate genomes 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 (Fig. S1; see Supporting Information for additional details of the

pan-genome and within-sample site diversity). In pairwise comparisons, the genomes shared on average 77% of their genes (Supporting Table S1).

Phylogenetic analysis reveals three distinct *Mesotoga* lineages.

The 16S rRNA genes of all 17 genomes had ≥99% identity to *M. prima*; phylogenetic trees revealed three distinct lineages (Fig. 1a). Genome networks based on core single nucleotide polymorphisms (SNPs) also had topologies consistent with the 16S rRNA gene phylogeny, with three distinct lineages being evident (Fig. 1b). Two lineages have a widespread geographical distribution: the World lineage (W; found in all regions represented) and the US lineage found in Baltimore Harbor and Colorado in the USA. The Alberta (A) lineage was observed in the Albertan samples only. Interestingly, *M. prima* has one 16S rRNA gene from the W lineage and one from the US lineage, suggesting one copy has been acquired laterally.

Figure 1

Very little reticulate evolution was observed among the three groups (Fig. 1b), and the A lineage in particular showed very little connection with the other two groups, suggesting that the three lineages have evolved independently for a relatively long time. In agreement with this, the ANI within groups ranged from 89.6-99.9%, while ANI between lineages ranged from 79.3-87.6% (Supporting Table S2). The same pattern was observed for the pangenome, with most lateral connections occurring within groups (Fig. 1c). Moreover, genomes from isolates of the same lineages share more genes in

comparative analyses: average 86% within W and 92% within A (Supporting Table S1). 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 70% of genes, W and US share 76%, and A and US share 75% of their genes.

A high level of recombination was detected, with the majority (> 200) of recombination events involving genomes from the same lineage (Fig. S2). For the W and A lineages, respectively, the average recombination tract length was estimated to be 36,000 – 56,000 bp and 17,000–23,000 bp; the population mutation rate (θ) was estimated to be 0.022 and 0.013, and the population recombination rate (γ) to be 1.8 (range 1.5–2.2) and 1.5 (1.3–1.7). The resulting high γ/θ ratios of ~82–115 indicate high levels of recombination and are similar to estimates for *Thermotoga* spp. (Nesbø *et al.*, 2015).

Phylogenetic analysis identified 52 regions where recombination likely occurred between lineages: 39 regions showed evidence of recombination between *Mesotoga* sp. BH458 and the W lineage, eight regions suggested recombination between *Mesotoga* sp. BH458 and the A lineage, and only five regions showed possible recombination between A- and W-lineage genomes (Fig. 2). The regions with recombination involving the A lineage were short (range 230–530 bp) and the sequences more divergent, whereas several of the fragments involving the W lineage and *Mesotoga* sp. BH458 were > 5 kb (average 3000 bp, range 260–20,000). Multiple recombination events in the same locus will eventually result in shorter recombinant fragments being detected (see, e.g. (Mau *et al.*, 2006)). Taken together with the >10 kb length of the recombinant fragments detected in the within-lineage analysis, this difference in recombinant-fragment-length suggests

that recombination events between the W lineage and *Mesotoga* sp. BH458 are more recent than those involving the A lineage. Very high levels of recombination were observed for a few genes. Among these is Theba_0319 in *M. prima*, the fourth most highly expressed gene (Supporting Table S3) that encodes the OmpB protein (Petrus *et al.*, 2012), a major component of the toga structure of Thermotogae.

Figure 2

Comparison to metagenomes and phylogeographic patterns of the three *Mesotoga* lineages

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 Supporting Information).

Table 2

Recent range expansion of the W-lineage: *Mesotoga* sequences with high similarity to the W lineage were identified using BLASTN searches in several wastewater treatment systems confirming its wide distribution in these environments (Table 2). A network

including population genomes (PGs) of *Mesotoga* contigs (with > 90% sequence identity to W isolate genome) from three metagenomes dominated by W lineage sequences (Long Beach, Boston and Hong Kong, Table 2) revealed no geographical structuring.

Isolation by distance can explain the distribution of US genomes: The metagenome data expanded the observed distribution of the US lineage. As expected, metagenome IMG 15764 from Albertan oil reservoir E (the source of *Mesotoga* sp. SC_PW1-3) contained sequences with high identity to the A lineage. However, it also contained many sequences with high identity to the US lineage (Table 2), and sequence binning yielded two *Mesotoga* metagenome-assembled genomes (MAGs): one most similar to US-genomes (Fig. S3b) and one with a mix of sequences from the A lineage and US lineage (not shown).

The network of US-*Mesotoga* including PGs composed of contigs from metagenomes in Table 2 (with sequence identity > 80% to US-isolate genomes) revealed three groups (Fig. S3b) where PGs from New York and Blank Spring (California) form a cluster that does not contain any of the genomes sequenced in this study (Table 1). The clustering of remaining genomes correlates with both geography and environment type: the MAG assembled from oil reservoir E (Alberta), two MAGs from an Alaskan oil reservoir (Hu *et al*., 2016) , 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), having substantial numbers of A-lineage sequences detected in wastewater metagenomes. For this lineage, MAGs were available from the same oil reservoir in Alaska where we observed the US-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 (Fig. S3c).

Distinct metabolism in mesophilic Thermotogae

We also examined the newly available genomes for metabolic insights, which may be linked to *Mesotoga*'s lower growth temperatures and may influence the role(s) *Mesotoga* play in their environments.

***Mesotoga*-specific genes:** Comparison of the *Mesotoga* isolate genomes to other Thermotogae genomes in IMG revealed 200 *M. prima* genes found in all *Mesotoga* genomes (including the more distantly related *Mesotoga infera* not included in the phylogenomic analyses), but in no other Thermotogae genomes. The majority of these genes were hypothetical proteins (N=119, Supporting Table S3). When *Mesotoga*-specific genes with a predicted function were classified according to Clusters of Orthologous Groups (COG) categories, the largest category was 'Amino Acid

metabolism and transport' with 11 genes, most of which were dipeptidases (COG4690, N=6).

***Mesotoga-specific genes related to O_2 exposure**:* Several *Mesotoga*-specific genes are predicted to be involved in oxygen radical defense (Supporting Table S4). 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.

Reducing equivalents and thiosulfate reduction: *Mesotoga*'s core metabolism differs from that of other characterized Thermotogae. While growth of most Thermotogae is stimulated by, but not dependent upon, the presence of thiosulfate, sulfur, or other reduced sulfur compounds in laboratory medium (Ravot *et al.*, 1995; Boileau *et al.*, 2016), reduction of sulfur compounds appears to be essential for growth of *Mesotoga* in pure culture (Hania *et al.*, 2011; 2013; Fadhlaoui *et al.*, 2017). The first description of *M. prima* (Nesbø *et al.*, 2012) reported that growth was only slightly stimulated by the presence of thiosulfate or sulfur. However, here we observed growth of this isolate only in the presence of sulfur or thiosulfate (Supporting Table S5 and Table S6), confirming that this is a general trait of *Mesotoga* spp. Additionally, while other Thermotogae produce H_2 (and H_2S if grown with partially reduced sulfur compounds), *Mesotoga* spp. produce large amounts of H_2S and no or little H_2 (Supporting Table S5).

To reconcile these observations with genomic data, transcriptome analysis was performed using a culture of *M. prima* grown with 0.5% yeast extract, xylose and thiosulfate. RNAseq analysis revealed high expression of Theba_0443 (RPKM of 3650;

Supporting Tables S3 and S4) encoding a Fe-hydrogenase homologous to the one used by *Kosmotoga 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 (Supporting 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 (Fig. 3). Both genes were transcribed at moderate levels in *M. prima* grown with thiosulfate (RPKM 341 and 243, respectively), whereas 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: $E°$` [$S_2O_3^{2-}$/ $HS^-$ + $SO_3^{2-}$] = -402 mV and $E°$` [NAD+/ NADH] = -320 mV. The $E°$` 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 $E°`$ $[2H^+ / H_2]$ = -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). However, a cluster of five highly transcribed genes (Theba_0461 – 0465, RPKM 1203-3697, Supporting Table S4) encodes 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 (Fig. 3). We hypothesize that this complex is involved in intracellular synthesis of molecular hydrogen for thiosulfate reduction by SudAB coupled to NADH oxidation (formed by Mbx and/or Rnf complexes, see below and Fig. 3). The Hnd genes have homologs in other Thermotogae, however, similar genomic context is observed only in genomes of other *Mesotoga* and *Kosmotoga* spp. (Supporting Table S7).

Figure 3

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 proton-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 complex involved.

Acetate and xylose utilization: Growth on acetate was reported for *Mesotoga* PhosAc3 (Hania *et al.*, 2015), and we observed weak stimulation of growth of its close relative *M. prima* by acetate (day 5-10 in Supporting Fig. S4 and Table S6). (Nobu *et al.*, 2015) suggested that Ca. "*Mesotoga acetoxidans*", a MAG closely related to *M. infera,* oxidizes acetate by using a novel pathway even though the genes comprising the pathway are conserved in all Thermotogae genomes. Yet, this phenotype is uncommon among Thermotogae and has been reported only for *Pseudothermotoga lettingae* (Balk *et al.*, 2002). Instead, many Thermotogae are inhibited by acetate, including one of *Mesotoga*'s closest relatives, *K. olearia* (Dipippo *et al.*, 2009). Our search for *Mesotoga*-specific genes that may be responsible for their observed growth on acetate revealed a candidate gene encoding a bacterial homodimeric pyruvate:ferredoxin oxidoreductase (PFOR; Theba_1954), with close homologs only found in *Kosmotoga pacifica* (Jiang *et al.*, 2017) and *Mesoaciditoga lauensis* (Reysenbach *et al.*, 2013) . Unfortunately, the description of these two species did not investigate growth on acetate. The *pfor* gene is distantly related to the archaeal multi-subunit-type used by other Thermotogae (Ragsdale, 2003) and almost all its close homologs fall within the *Clostridia* (Supporting Fig. S4). Genes having 97-99% identity to *pfor* from *M. infera*, and 83-85% identity to the *M. prima* homolog, were found in both the metagenome and metatranscriptome published by

(Nobu *et al.*, 2015) (locus tag JGI12104J13512_10052834 and JGI11944J13513_10066464) but were not included in their model. We propose that PFOR may work with the acetate kinase (Theba_0428 in *M. prima*) and phosphotransacetylase (Theba_0782 in *M. prima*) found in all Thermotogae to enable *Mesotoga* to grow on acetate. At high extracellular acetate concentrations we suggest that PFOR shifts the balance favoring the production of pyruvate from acetyl-CoA (i.e. serves 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) (Fig. 3). Several possible xylulose kinase genes were found co-localized with genes encoding xylosidases, sugar transporters, and kinases, suggesting their synergetic activities in xylan hydrolysis, xylose import, and utilization.

Discussion

Mesotoga have conserved core genomes and diverse pangenomes

The comparative analysis of the *Mesotoga* genomes revealed higher levels of diversity in genome content than observed in the hyperthermophilic Thermotogae. Whereas *Thermotoga* spp. share > 90% of their genes in pairwise comparisons (Nesbø *et al.*, 2015), *Mesotoga* genomes from the same lineage share on average 86% - 92% of their genes. (Nesbø *et al.*, 2015) suggested that high levels of recombination may be partly responsible for homogenizing *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 larger proportion of variable accessory genes. Perhaps

more cryptic niches are available in low- versus high-temperature subsurface environments (McInerney *et al.*, 2017), or *Mesotoga* may have larger effective population sizes than the hyperthermophiles (Andreani *et al.*, 2017).

Comparing the nucleotide divergence within the core genomes revealed 'species' level divergence between the three lineages detected (ANI < 87%), while ANI within the A and W lineage was very high at 98.5% and 97.5%, respectively. In comparison, the ANI among the *Thermotoga* genomes investigated by (Nesbø *et al.*, 2015) was 95.3%. Thus *Mesotoga* spp., particularly those from the W-lineage, appear to have more conserved core genomes and more diverse pangenomes than their hyperthermophilic relatives.

Three *Mesotoga* lineages with distinct phylogeographies: isolation by distance, range expansion, and burial with isolation

The networks calculated for both the core and the pangenome gave the same overall topology as that observed in the 16S rRNA tree with three distinct groups. The low level of recombination observed among these three groups suggests they have evolved independently for a relatively long time. The observation of several recent recombination events between the W and US lineages, which currently co-exist in at least one location (i.e., Baltimore Harbor), demonstrates that recombination between lineages is possible. We therefore suggest that the three *Mesotoga* lineages have evolved independently due to geographical, not genetic, isolation. This is contrary to the patterns of geneflow observed in *Thermotoga* spp. genomes, where environment type was more important than geographic separation in determining level of geneflow (Nesbø *et al.*, 2015). Although it

may seem counterintuitive that mesophilic *Mesotoga* would be more affected by geographical separation than hyperthermophilic *Thermotoga*, this may be a consequence of their anaerobic metabolism. (Chakraborty *et al.*, 2018) showed that bacteria are dispersed out of deep hot subsurface oil reservoirs and into the ocean through hydrocarbon seeps, and this might serve as a major route of migration between these environments. Moreover, temperature gradients associated with hydrothermal systems are often very sharp (Dick *et al.*, 2013), and hyperthermophilic *Thermotoga* cells will therefore quickly become inactive if they enter cold aerobic ocean water (Fig. S6). Mesophilic *Mesotoga* cells will, however, more likely enter oxygenated environments having a suitable temperature before they reach a new optimal anaerobic site and therefore may more often succumb to oxygen exposure, limiting viable dispersal and gene exchange (Fig. S6). In support of this, many *Mesotoga*-specific genes appear to be involved in O_2 or H_2O_2 detoxification.

Within the three lineages we see patterns consistent with different phylogeographic histories. Comparing the isolate genomes to *Mesotoga* sequences in metagenomes, the US-lineage shows patterns consistent with isolation by distance. Moreover, the US-lineage has an intermediate position between the A- and W-lineages when considering ANI, gene content, and recombination, which may be due to this lineage co-existing with both W and A genomes (e.g. Baltimore Harbor, Oil field E).

Members of the widespread W-lineage show high identity in their core genomes, large pan-genomes, and no indication of geographical structuring, indicative of a recent range expansion (Choudoir *et al.*, 2017). To date, W lineage *Mesotoga* have been detected only at sites heavily influenced by human activities (e.g., drilling,

contamination), suggesting an anthropogenic role in their dispersal and possibly selective pressure on these genomes. Interestingly, one of the W-lineage-specific genes (Theba_0620, Supplemental material) is involved in synthesis of poly-gamma glutamate, which has been implicated in survival under harsh conditions and may have contributed to the wide distribution of this lineage.

The A lineage is more isolated from the other lineages (Fig. 1 and 2), which might suggest that this clade evolved in isolation since the formation of oil reservoir sediments in Alberta 55–120 Ma (Schaefer, 2005; Selby, 2005; Head *et al.*, 2014). The high similarity of the MAGs from the Alaskan oil field to the Albertan genomes and MAGs from the A and US lineages (Fig. S3) could be due to these oil reservoir sediments being laid down around the same time (~100 Ma (Hu *et al.*, 2016). However, the position of these MAGs in the genome networks could also be explained by these oil reservoirs being colonized by the same subsurface population, as suggested for *Thermotoga* spp. (Nesbø *et al.*, 2015). Additional oil reservoir genomes are needed to resolve this question and also to determine if the A-lineage is indeed indigenous to oil reservoirs.

Also this lineage has likely experienced recent dispersal events due to human activities: MAGs from a polluted river bank in Liangjiang, China (Wang and He, 2013) and waste water from Oakland (California) showed very high identity to *Mesotoga* sp. SC_NapDC from a northern Albertan oil sands tailings pond. In fact, these genomes show the second highest level of pairwise identity among the A lineage genomes (Fig. S3d), suggesting recent dispersal, possibly due to human activities in the last decades.

Distinct metabolism in mesophilic Thermotogae.

The *Mesotoga* genomes and transcriptome also elucidated the genetic background for their distinct energy metabolism compared to thermophilic Thermotogae bacteria, i.e. the strict need for sulfur or thiosulfate and no or little H_2 production, but rather H_2S production unless in co-culture with a sulfate reducer (Fadhlaoui *et al.*, 2017). (Fadhlaoui *et al.*, 2017) suggested that *Mesotoga'*s inability to ferment sugars is mainly due to its lack of a bifurcating hydrogenase. However, *K. olearia* also lacks this enzyme and ferments pyruvate, producing large amounts of hydrogen using the homolog of *M. prima*'s only Fe-hydrogenase (Pollo *et al.*, 2017). In the model in Fig. 3 panel A, we therefore instead suggest this is accomplished by utilizing a bifurcated hydrogenase, SudAB, Mbx and Rnf.

The model shown in Fig. 3 panel A accounts for the observed dependence of *M. prima* on sulfur or thiosulfate for growth, the lack of H_2 production, and involves proteins previously implicated in hydrogen and sulfur metabolism. Importantly, however, currently there are no known enzymes that couple H_2 oxidation and thiosulfate/sulfur reduction. It is therefore possible that *M. prima* SudAB uses NADH as the electron donor and is much more effective than the hydrogenase which results in almost no H_2 as growth product (Fig. 3 panel C).

Alternatively, thiosulfate reduction coupled to H_2 oxidation (i.e., the postulated role of SudAB; Fig. 3 panel A) may be performed solely by one of the highly-transcribed hypothetical *Mesotoga* proteins with no match to genes in *Kosmotoga* and other Thermotogae, or in combination with SudAB (Fig. 3 panel B). Several candidate genes listed in Supporting Table S4 encode proteins with unknown functions. Functional studies of these genes, as well as the gene products shown in Fig. 3, are needed to assess

their role, if any, in thiosulfate reduction. Additional genes that may be involved in thiosulfate uptake and electron transfer are also discussed in Supporting Information. Interestingly, PhosAc3 and *M. infera* were found to reduce only elemental sulfur (Hania *et al.*, 2011; 2013) whereas the strains isolated by us also reduce thiosulfate. These differences may reflect selection during isolation; all the isolates obtained in the current study were from enrichment cultures containing thiosulfate, whereas PhosAc3 and *M. infera* were enriched with sulfur. This suggests that the sulfur-compound-preference may be a variable and flexible phenotype in *Mesotoga* populations.

We also observed gene content differences that probably are directly linked to *Mesotoga*'s lower growth temperature. The higher abundance of genes associated with 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 two of the genes possibly involved in O_2 or H_2O_2 defense (Theba_0075, Catalase and Theba_2399, Rubrerythrin) were highly expressed (top 5% of expressed genes; Supporting Table S3 and S4), 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.

Conclusion

Our genomic analysis suggests that the lower growth temperature of *Mesotoga* spp. compared to the hyperthermophilic *Thermotoga* has likely influenced *Mesotoga* phylogeography, with geographic separation historically having a greater influence than genetic separation, possibly due to the damaging effects of oxygen exposure during dispersal (Fig. S6). Whether this is a general feature of strictly anaerobic organisms remains to be resolved. There is also some indication of possible ecotype differentiation among the *Mesotoga* lineages, with the US lineage being common in communities degrading aromatic pollutants (PCB, toluene) 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 analysis including data from metagenomes also suggests that anthropogenic activities have expanded *Mesotoga*'s habitats and also enhanced its dispersal capabilities (Fig. S6), with inferred recent long-distance dispersal events involving anthropogenic environments and/or activities.

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

Experimental procedures

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). All nine available isolates were selected for genome sequencing. 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 SAGs. Detailed descriptions of isolation procedures, DNA extraction, genome assembly and annotation are provided in Supporting 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 $< e^{-50}$. For additional details on

search parameters and information on assembly of draft genomes from metagenomic sequences or contigs see Supporting Information.

Genome content and genome alignments

Shared genes and genome specific 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 lineage-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 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 *et al.*, 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 and Roger, 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 libraries were performed as described by (Pollo *et al.*, 2017). RNAseq analysis was carried out in CLC Genomics Workbench version 7.0.4 as described by (Pollo *et al.*, 2017). The transcriptome has been submitted to GenBank's SRA archive with accession number PRJNA495810.

H_2 and H_2S measurements

Standard gas chromatographic analysis of culture headspace gas was performed using an Agilent CP4900 Micro Gas Chromatograph to detect H_2 production by the cultures, as described in Supporting Information. Dissolved sulfide concentrations were measured using a VACUettes® Visual High Range Kit (Chemetrics), following the manufacturer's recommendations.

Acknowledgements

This work was supported by a Research Council of Norway award (project no. 180444/V40) to C.L.N. and by a Genome Canada grant (Hydrocarbon Metagenomics Project) to J.F. The work of IVK was supported by the Russian Science Foundation grant # 18-44-04024. We thank Dr. Alexander Lebedinsky for constructive criticism and

helpful suggestions.

Conflict of Interest Statement

The authors declare no conflict of interest.

References

Andreani, N.A., Hesse, E., and Vos, M. (2017) Prokaryote genome fluidity is dependent on effective population size. *ISME J.* **11**: 1719–1721.

Archibald, J.M. and Roger, A.J. (2002) Gene conversion and the evolution of euryarchaeal chaperonins: a Maximum Likelihood-based method for detecting conflicting phylogenetic signals. *J Mol Evol* **55**: 232–245.

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

Bhandari, V. and Gupta, R.S. (2014) The phylum Thermotogae. In, *The Prokaryotes*. Springer, Berlin, Heidelberg, Berlin, Heidelberg, pp. 989–1015.

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

Chakraborty, A., Ellefson, E., Li, C., Gittins, D., Brooks, J.M., Bernard, B.B., and Hubert, C.R.J. (2018) Thermophilic endospores associated with migrated thermogenic hydrocarbons in deep Gulf of Mexico marine sediments. *ISME J.* **8**: 1–12.

Choudoir, M.J., Panke-Buisse, K., Andam, C.P., and Buckley, D.H. (2017) Genome surfing as driver of microbial genomic diversity. *Trends Microbiol* **8**: 624–636.

Darling, A.E., Mau, B., and Perna, N.T. (2010) ProgressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**: e11147.

Dick, G.J., Anantharaman, K., Baker, B.J., Li, M., Reed, D.C., and Sheik, C.S. (2013) The microbiology of deep-sea hydrothermal vent plumes: ecological and biogeographic linkages to seafloor and water column habitats. *Front Microbiol* **4**: 124.

Dipippo, J.L., Nesbø, C.L., Dahle, H., Doolittle, W.F., Birkland, N.-K., and Noll, K.M. (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, R.E. and Fedorak, P.M. (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, W.B., 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, S.K., May, H.D., and Sowers, K.R. (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.

Fowler, S.J., Dong, X., Sensen, C.W., Suflita, J.M., and Gieg, L.M. (2012) Methanogenic toluene metabolism: community structure and intermediates. *Environ Microbiol* **14**: 754–764.

Gieg, L.M., Kolhatkar, R.V., McInerney, M.J., Tanner, R.S., Harris, S.H., Sublette, K.L., and Suflita, J.M. (1999) Intrinsic bioremediation of petroleum hydrocarbons in a gas condensate-contaminated aquifer. *Environ Sci Technol* **33**: 2550–2560.

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

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

Hagen, W.R., Silva, P.J., Amorim, M.A., Hagedoorn, P.L., Wassink, H., Haaker, H., and Robb, F.T. (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, W.B., 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.

Hania, W.B., Ghodbane, R., Postec, A., Brochier-Armanet, C., Hamdi, M., Fardeau, M.-L., and Ollivier, B. (2011) Cultivation of the first mesophilic representative (“mesotoga”) within the order Thermotogales. *Syst Appl Microbiolo* **34**: 581–585.

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

Head, I.M., Gray, N.D., and Larter, S.R. (2014) Life in the slow lane; biogeochemistry of biodegraded petroleum containing reservoirs and implications for energy recovery and carbon management. *Front Microbiol* **5**: 297.

Holoman, T.R., Elberson, M.A., Cutter, L.A., May, H.D., and Sowers, K.R. (1998) Characterization of a defined 2,3,5, 6-tetrachlorobiphenyl-ortho-dechlorinating microbial community by comparative sequence analysis of genes coding for 16S rRNA. *Appl Environ Microbiol* **64**: 3359–3367.

Hu, P., Tom, L., Singh, A., Thomas, B.C., Baker, B.J., Piceno, Y.M., et al. (2016) Genome-resolved metagenomic analysis reveals roles for candidate phyla and other microbial community members in biogeochemical transformations in oil reservoirs. *mBio* **7**: e01669–15–12.

Hulecki, J.C., Foght, J.M., Gray, M.R., and Fedorak, P.M. (2009) Sulfide persistence in oil field waters amended with nitrate and acetate. *J Ind Microbiol Biotechnol* **36**: 1499–1511.

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

Jiang, L., L'Haridon, S., Jebbar, M., Xu, H., Alain, K., and Shao, Z. (2017) Complete genome sequence and whole-genome phylogeny of Kosmotoga pacifica type strain SLHLJ1T from an East Pacific hydrothermal sediment. *Stand Genomic Sci* **12**: 1–9.

Johnson, M.R., Conners, S.B., Montero, C.I., Chou, C.J., Shockley, K.R., and Kelly, R.M. (2005) The *Thermotoga maritima* phenotype is impacted by syntrophic interaction with *Methanococcus jannaschii* in hyperthermophilic coculture. *Appl Environ Microbiol* **72**: 811–818.

Jolley, K.A., Wilson, D.J., Kriz, P., McVean, G., McVean, and Maiden, M.C.J. (2004) The Influence of Mutation, Recombination, Population History, and Selection on Patterns of Genetic Diversity in *Neisseria meningitidis*. *Mol Biol Evol* **22**: 562–569.

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

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

Lee, W., Lewandowski, Z., Nielsen, P.H., and Hamilton, W.A. (1995) Role of sulfate□reducing bacteria in corrosion of mild steel: A review. *Biofouling* **8**: 165–194.

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

Mau, B., Glasner, J.D., Darling, A.E., and Perna, N.T. (2006) Genome-wide detection and analysis of homologous recombination among sequenced strains of Escherichia coli. *Genome Biol* **7**: R44.

McInerney, J.O., McNally, A., and O'Connell, M.J. (2017) Why prokaryotes have pangenomes. *Nat Microbiol* **2**: 1–5.

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

Mulkidjanian, A.Y., Galperin, M.Y., Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2008) Evolutionary primacy of sodium bioenergetics. *Biology Direct* **3**: 13.

Müller, V., Imkamp, F., Biegel, E., Schmidt, S., and Dilling, S. (2008) Discovery of a ferredoxin:NAD+□oxidoreductase (Rnf) in *Acetobacterium woodii*. *Ann NY Acad Sci* **1125**: 137–146.

Nesbø, C.L., Bradnan, D.M., Adebusuyi, A., Dlutek, M., Petrus, A.K., Foght, J., et al. (2012) *Mesotoga prima* gen. nov., sp. nov., the first described mesophilic species of the Thermotogales. *Extremophiles* **16**: 387–393.

Nesbø, C.L., Dlutek, M., Zhaxybayeva, O., and Doolittle, W.F. (2006) Evidence for existence of "mesotogas," members of the order Thermotogales adapted to low-temperature environments. *Appl Environ Microbiol* **72**: 5061–5068.

Nesbø, C.L., Kumaraswamy, R., Dlutek, M., Doolittle, W.F., and Foght, J.M. (2010) Searching for mesophilic Thermotogales bacteria: "mesotogas" in the wild. *Appl Environ Microbiol* **76**: 4896–4900.

Nesbø, C.L., S Swithers, K., Dahle, H., Haverkamp, T.H.A., Birkeland, N.-K., Sokolova, T., et al. (2015) Evidence for extensive gene flow and *Thermotoga* subpopulations in subsurface and marine environments. *ISME J.* **9**: 1532–1542.

Nobu, M.K., Narihiro, T., Rinke, C., Kamagata, Y., Tringe, S.G., Woyke, T., and Liu, W.-T. (2015) Microbial dark matter ecogenomics reveals complex synergistic networks in a methanogenic bioreactor. *ISME J.* **9**: 1710–1722.

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

Petrus, A.K., Swithers, K.S., Ranjit, C., Wu, S., Brewer, H.M., Gogarten, J.P., et al. (2012) Genes for the major structural components of Thermotogales species' togas revealed by proteomic and evolutionary analyses of OmpA and OmpB homologs. *PLoS ONE* **7**: e40236.

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

Pollo, S.M.J., Adebusuyi, A.A., Straub, T.J., Foght, J.M., Zhaxybayeva, O., and Nesbø, C.L. (2017) Genomic insights into temperature-dependent transcriptional responses of Kosmotoga olearia, a deep-biosphere bacterium that can grow from 20 to 79 °C. *Extremophiles* **21**: 963–979.

Pollo, S.M.J., Zhaxybayeva, O., and Nesbø, C.L. (2015) Insights into thermoadaptation and the evolution of mesophily from the bacterial phylum *Thermotogae*. *Can J Microbiol* **61**: 655–670.

Ragsdale, S.W. (2003) Pyruvate ferredoxin oxidoreductase and its radical intermediate. *Chem Rev* **103**: 2333–2346.

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

Reysenbach, A.-L., Liu, Y., Lindgren, A.R., Wagner, I.D., Sislak, C.D., Mets, A., and Schouten, S. (2013) *Mesoaciditoga lauensis* gen. nov., sp. nov., a moderately thermoacidophilic member of the order Thermotogales from a deep-sea hydrothermal vent. *Int J Syst Evol Microbiol* **63**: 4724–4729.

Schaefer, B.F. (2005) GEOCHEMISTRY: When Do Rocks Become Oil? *Science* **308**: 1267–1268.

Schut, G.J., Boyd, E.S., Peters, J.W., and Adams, M.W.W. (2013) The modular respiratory complexes involved in hydrogen and sulfur metabolism by heterotrophic hyperthermophilic archaea and their evolutionary implications. *FEMS Microbiol Rev* **37**: 182–203.

Selby, D. (2005) Direct Radiometric Dating of Hydrocarbon Deposits Using Rhenium-Osmium Isotopes. *Science* **308**: 1293–1295.

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

Tan, B., Jane Fowler, S., Laban, N.A., Dong, X., Sensen, C.W., Foght, J., and Gieg, L.M. (2015) Comparative analysis of metagenomes from three methanogenic hydrocarbon-

degrading enrichment cultures with 41 environmental samples. *ISME J.* **9**: 2028–2045.

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

Voordouw, G., Grigoryan, A.A., Lambo, A., Lin, S., Park, H.S., Jack, T.R., et al. (2009) Sulfide remediation by pulsed injection of nitrate into a low temperature canadian heavy oil reservoir. *Environ Sci Technol* **43**: 9512–9518.

Wang, S. and He, J. (2013) Phylogenetically Distinct Bacteria Involve Extensive Dechlorination of Aroclor 1260 in Sediment-Free Cultures. *PLoS ONE* **8**: e59178.

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

Wolfe, A.J. (2005) The acetate switch. *Microbiol Mol Biol Rev* **69**: 12–50.

Zhaxybayeva, O., Swithers, K.S., Foght, J., Green, A.G., Bruce, D., Detter, C., et al. (2012) Genome sequence of the mesophilic thermotogales bacterium *Mesotoga prima* MesG1. Ag. 4.2 reveals the largest thermotogales genome to date. *Genome Biol Evol* **4**: 700–708.

Figure legends

Fig. 1. Phylogenetic relationships among *Mesotoga* genomes based on (a) 16SrRNA genes, (b) core SNPs and (c) presence/absence of shared 500-bp genomic fragments. The 16S rRNA maximum likelihood phylogeny was estimated using RAxML in Geneious v 10. For networks shown in (b) and (c), data were obtained using PanSeq (Laing *et al.*, 2010). Core SNPs in (b) were required to be present in 14 of 18 genomes (including SAGs), and genomic fragments were considered shared if they were at least 70% identical. The network in (c) was constructed using only genomes from isolates; shared fragments were required to be present in all 9 genomes and be at least 70% identical in nucleotide sequence. Networks were calculated in SplitsTree using NeighborNet algorithm (Huson and Bryant, 2006) from uncorrected distances. The isolates cluster into the same three lineages in (a), (b) and (c) and are named based on their geographical distribution. The World (W) lineage occurs in all regions represented. The US lineage is found in locations in the USA and the Alberta (A) lineage was observed in the Albertan samples only.

Fig. 2. Visualization of recombination events detected among *Mesotoga* genomes from different lineages. The genomes are color-coded according to lineage (see text and Fig. 1) and arranged on the circumference of the circle: W lineage, blue; US lineage, orange; A lineage, green. Only isolate genomes were included in this analysis. A single representative genome (BR5.2) selected from the three highly similar genomes comprising the BR population (as described in Supporting Material) was included in the analysis. The recombination events with predicted donor and recipient are shown as lines connecting the two genomes; the locations of recombined regions, where line color reflects the donor lineage and the width of the line is proportional to the length of the recombinant region. The diagram was generated using Circos Version circos-0.69 (Krzywinski *et al.*, 2009).

Fig. 3. Model of energy generation pathway in *Mesotoga prima* during growth on xylose and thiosulfate. Glucose and xylose poly- and oligosaccharides are hydrolyzed by various intracellular and interstitial glycosidases (GHs). Glucose oxidation occurs via the glycolytic Embden-Meyerhof-Parnas pathway, whereas xylose is utilized via xylose 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 intermediates in the pentose phosphate pathway. Xylose isomerase (Theba_1394) was among the most highly transcribed genes during cultivation of *M. prima* on xylose and thiosulfate (Supporting Table S3). 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 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* and *M. prima* PhosAc3, 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	Estimated % 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.*, 2015)	
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				(Gieg *et al.*, 1999; Fowler *et al.*, 2012)	
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 Supporting 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] The source was 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; accessed October 2018).

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 extracted from 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)** W: 2,455 bp (8) US: 634 bp (2)	**99.8% (98.5-100%)** 93.9% (88.7-99.6) 86.4% (89.7-95%)
Oil reservoir E	IMG: 15764	5,190,293 bp (4,833)	**A: 3,137,228 bp (3,259)** W: 195,159 bp (317) **US: 1,693,057 bp (1,308)**	**98.6% (71.3-100%)** 94.5% (79.7-100%) **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) W: 47,017 bp (61) **US: 297,011 bp (360)**	92% (85.3-99.5%) 93.4% (83-99.1%) **94.2%(83.6-100%)**
Blank Spring hotspring sediment	IMG: 94476	4,060,664 bp (2,037)	**A: 920,911 bp (924)** W: 226,161 bp (176) **US: 2,025,524 bp (825)**	**91.6% (64.3-100%)** 85.4% (64.3-99.1%) **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)** **W: 710,377 bp (914)** US: 202,508 bp (207)	**96.1% (65-100%)** **94.1% (64.3-100%)** 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%)**

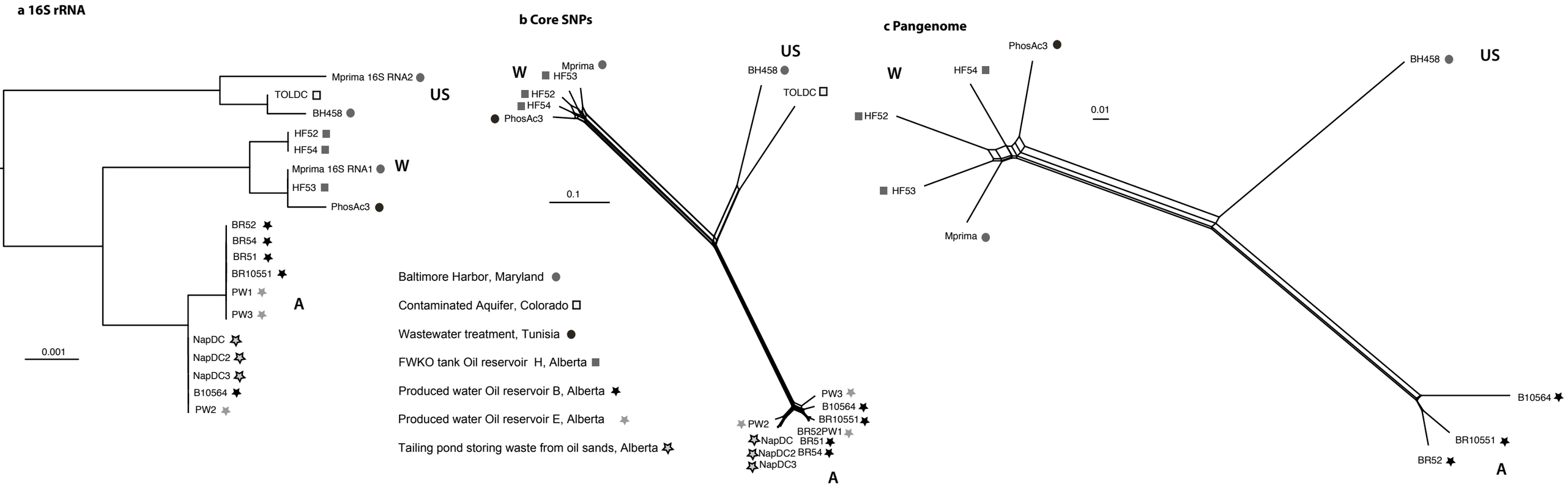
(continued)

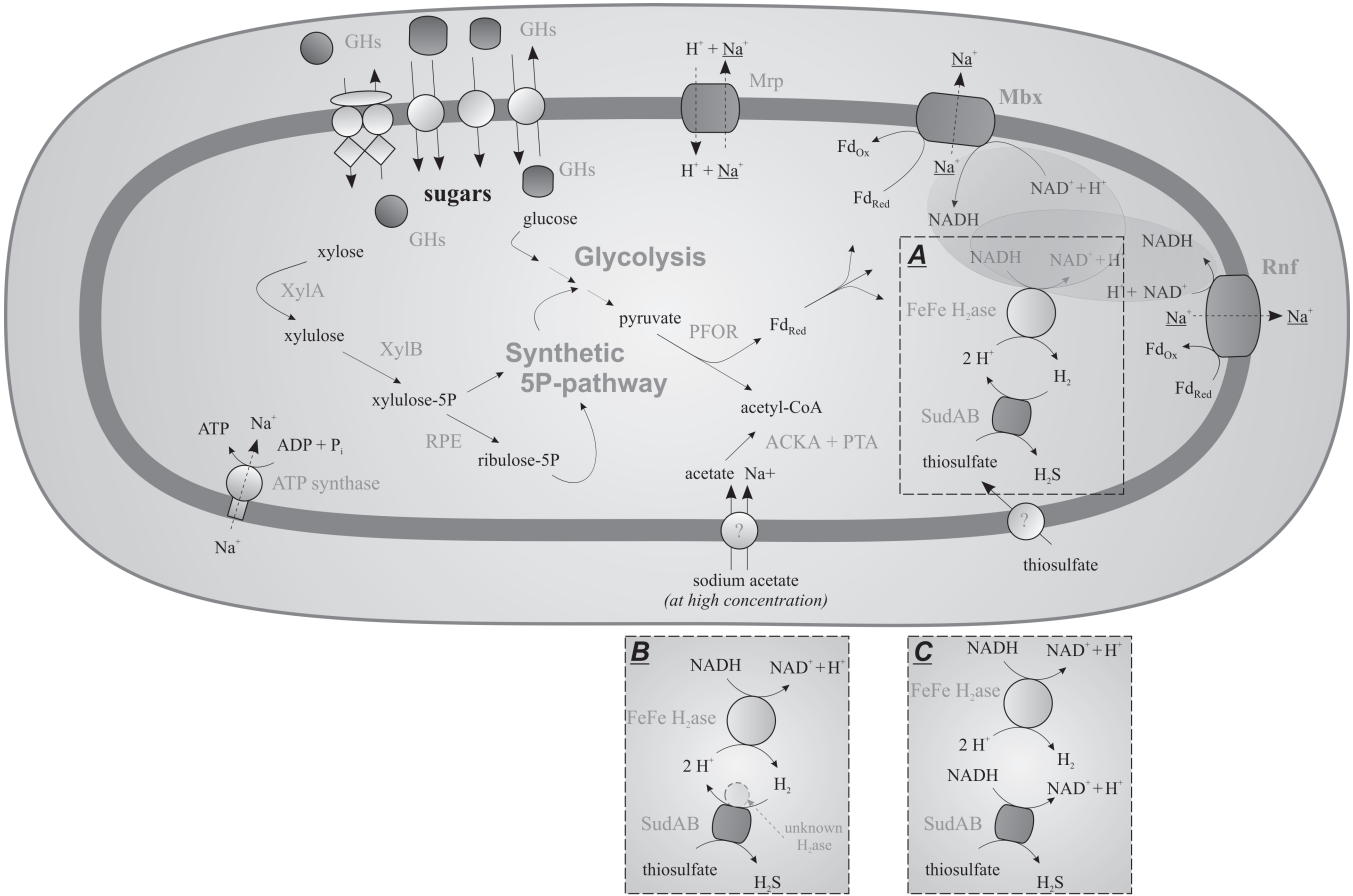
Metagenome-assembled genomes (MAGs)

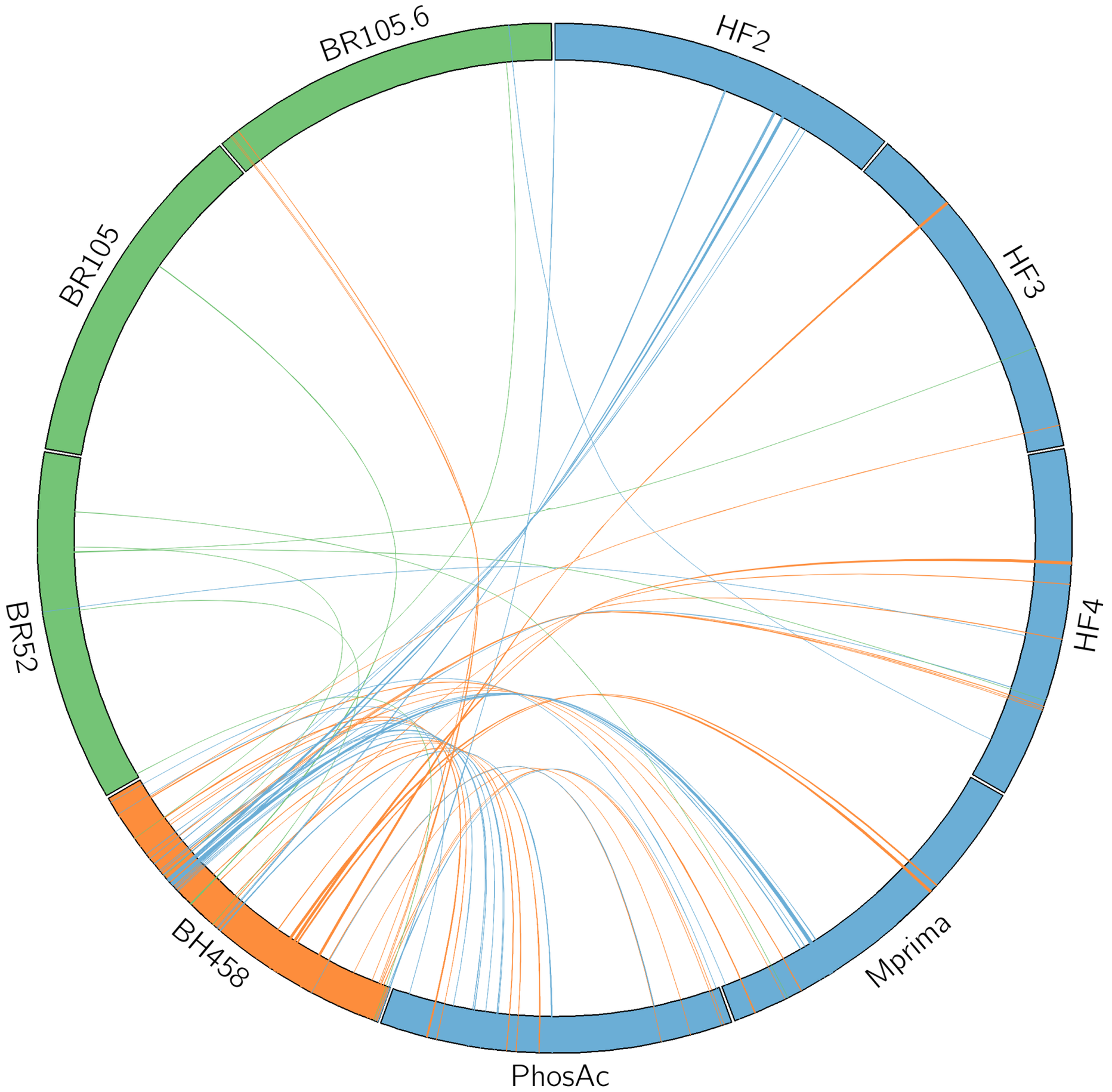
Alaska Oil reservoir LGGP01[g]	GenBank: GCA_001508515	1,712,609 bp (440)	**A: 1,470,927 bp (384)** W: 80,31 bp (30) US: 47,775 bp (21)	**98.5% (71.3-100%)** 92.3% (70.0-99.6%) 88.9% (71.9-100%)
LGGH01[g]	GCA_001508435	1,225,111 bp (267)	A: 48,008 bp (13) W: 63,609 bp (21) **US: 1,009,462 bp (233)**	89.9% (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 (Fig. S3b); only one is included here.
f) Wastewater treatment anaerobic digesters. Additional metagenomes with similar *Mesotoga* lineage composition are available from these sites 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
Mprima 16S RNA1
HF53
PhosAc3
W
BR52
BR54
BR51
BR10551
PW1
PW3
A
NapDC
NapDC2
NapDC3
B10564
PW2
0.001
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


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


BR105.6
HF2
HF3
HF4
Mprima
PhosAc
BH458
BR52
BR105