ORIGINAL ARTICLE

Evidence for extensive gene flow and *Thermotoga* subpopulations in subsurface and marine environments

Camilla L Nesbø^{1,2}, Kristen S Swithers^{3,4}, Håkon Dahle⁵, Thomas HA Haverkamp¹, Nils-Kåre Birkeland⁵, Tatiana Sokolova⁶, Ilya Kublanov⁶ and Olga Zhaxybayeva^{7,8}

¹Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biology, University of Oslo, Blindern, Oslo, Norway; ²Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada; ³Department of Cell Biology, Yale School of Medicine, Yale University, New Haven, CT, USA; ⁴Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA; ⁵Department of Biology and Centre for Geobiology, University of Bergen, Bergen, Norway; ⁶Winogradsky Institute of Microbiology, Russian Academy of Sciences, Moscow, Russia; ⁷Department of Biological Sciences, Dartmouth College, Hanover, NH, USA and ⁸Department of Computer Science, Dartmouth College, Hanover, NH, USA

Oil reservoirs represent a nutrient-rich ecological niche of the deep biosphere. Although most oil reservoirs are occupied by microbial populations, when and how the microbes colonized these environments remains unanswered. To address this question, we compared 11 genomes of *Thermotoga maritima*-like hyperthermophilic bacteria from two environment types: subsurface oil reservoirs in the North Sea and Japan, and marine sites located in the Kuril Islands, Italy and the Azores. We complemented our genomes with *Thermotoga* DNA from publicly available subsurface metagenomes from North America and Australia. Our analysis revealed complex non-bifurcating evolutionary history of the isolates' genomes, suggesting high amounts of gene flow across all sampled locations, a conjecture supported by numerous recombination events. Genomes from the same type of environment tend to be more similar, and have exchanged more genes with each other than with geographically close isolates from different types of environments. Hence, *Thermotoga* populations of oil reservoirs do not appear isolated, a requirement of the 'burial and isolation' hypothesis, under which reservoir bacteria are descendants of the isolated communities buried with sediments that over time became oil reservoirs. Instead, our analysis supports a more complex view, where bacteria from subsurface and marine populations have been continuously migrating into the oil reservoirs and influencing their genetic composition. The *Thermotoga* spp. in the oil reservoirs in the North Sea and Japan probably entered the reservoirs shortly after they were formed. An Australian oil reservoir, on the other hand, was likely colonized very recently, perhaps during human reservoir development.

The ISME Journal (2015) 9, 1532–1542; doi:10.1038/ismej.2014.238; published online 12 December 2014

Introduction

The deep biosphere harbors a considerable portion of the Earth biomass with microbial cell numbers roughly equal to those in surface environments (Fry *et al.*, 2009; Schrenk *et al.*, 2010; Kallmeyer *et al.*, 2012). Because of the extremely low flux of nutrients and energy, this is also the environment with the slowest growth rates and the longest estimated division times for prokaryotic cells—perhaps up to

1000 years per generation (Jørgensen, 2011; Morono *et al.*, 2011). However, the cells are alive and exist in a dormant or low-activity state: anaerobic heterotrophic cells from such environments respond quickly to nutrients, reaching doubling times numbered in days not years (Morono *et al.*, 2011). Metagenomic analyses suggest that subsurface environments harbor distinct microorganisms adapted for life in the subsurface (see, for example, Biddle *et al.*, 2011; Castelle *et al.*, 2013) and metatranscriptomics show that organisms from all domains of life are active in the subsurface (Orsi *et al.*, 2013). Moreover, distribution of phage in the deep subsurface indicate ongoing viral production (Engelhardt *et al.*, 2014). How microorganisms come to inhabit subsurface sediments, however, remains an open question (Schrenk *et al.*, 2010): for example, they may be introduced during sedimentation or transported from either the basement or adjacent sediments.

Correspondence: CL Nesbø, Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo Blindern, PO Box 1066, Oslo N-0316, Norway.

E-mail: c.l.nesbo@bio.uio.no

or O Zhaxybayeva, Department of Biological Sciences, Dartmouth College, Hanover, NH 03755, USA.

E-mail: olgazh@dartmouth.edu

Received 28 July 2014; revised 3 November 2014; accepted 10 November 2014; published online 12 December 2014

Some of the most nutrient-rich subsurface environments are oil reservoirs, which contain large amount of hydrocarbons and fatty acids, and the effect of microbial activities on the former is well documented (Head *et al.*, 2003; Ollivier and Alazard, 2010). In these reservoirs, *indigenous* microorganisms, that is, those not introduced by recent human activities, are thought to live in the biofilms attached to surfaces in the lower zone of the reservoir, where there is an oil–water contact (Magot, 2005; Bennett *et al.*, 2013). Microbial lifestyle in oil reservoirs is postulated to be slow, because of low levels of electron acceptors and some nutrients, such as phosphorus (Ollivier and Alazard, 2010). Moreover, cell counts are similar to those in other active deep biosphere systems (Bennett *et al.*, 2013).

As the temperature and pressure during oil formation can approach or exceed the upper limits for life (>100–150 °C and up to 100 MPa; Planckaert, 2005), it is unlikely that microorganisms in an oil reservoir originate from the source rock (Stetter and Huber, 1999). Instead, as oil migrates to shallower sediments, the newly formed oil reservoirs become occupied by an indigenous microbiome. As for subsurface sediments in general, two alternative hypotheses explaining where these microorganisms originate from have been proposed. In the first, ‘burial and isolation’ hypothesis (a corollary from the paleosterilization model of oil biodegradation), the microorganisms buried with the sediments adapt to the new environment as the oil reservoir is formed (Wilhelms *et al.*, 2001). Under this hypothesis, the microbial community within an oil reservoir is isolated, and there is no subsequent migration of other microorganisms into the reservoir from adjacent areas. The support for this hypothesis comes from examination of cool reservoirs with no biodegraded oil. During their geological history, the reservoir sediments were buried at depths where the temperature exceeds the survival limit of hydrocarbon-degrading microorganisms, resulting in ‘paleosterilization’ (Wilhelms *et al.*, 2001). The conjecture is that new organisms have not colonized such oil reservoirs after their subsequent uplift and cooling, and that is why no oil biodegradation is presently observed. In the alternative ‘colonization’ hypothesis, subsurface microorganisms populate the oil reservoir after their formation (Stetter and Huber, 1999; Fry *et al.*, 2009). The microbiomes of some reservoirs may have assembled through a combination of these mechanisms.

Most oil reservoirs in the world have contained oil for several millions years (Myr) (Huang and Larter, 2005). For instance, oil generation from the Draupne formation, the main source rock in the North Sea, began ~88 Myr ago and peaked at ~78 Myr ago (Adda, 2012). The sediments of the Troll reservoir are shallow marine sandstones from the middle and late Jurassic (~145–150 Myr ago) (Gautier, 2005). If bacteria have resided in the reservoir since the sediment deposition, their populations have been

isolated for over 145 Myr. As a result, ~900–90 000 single-nucleotide polymorphisms (SNPs) per genome are expected to accumulate between the reservoir bacteria and their counterparts from the outside via mutation (such wide range of uncertainty is because of our poor knowledge of generation time and mutation rate during growth and dormancy; see Supplementary Methods for calculations.) If the reservoirs are not isolated and the microbes colonized the reservoir from surrounding subsurface populations at a later time, we would expect ongoing gene flow, which will be detectable as numerous recombination events between oil reservoirs of any age. Lastly, if the reservoir has been colonized after humans started its development, we would expect large tracts of genomic identity of the colonizers to bacteria from, for instance, adjacent marine environments.

The *Thermotoga* are a bacterial phylum that has been detected in and isolated from numerous oil reservoirs, both continental and offshore (Cappelletti *et al.*, 2014). Some lineages to date are found only in these environments (Ollivier and Alazard, 2010), whereas others appear to be physiologically adapted to the reservoir conditions (Slobodkin *et al.*, 1999; Bonch-Osmolovskaya *et al.*, 2003; Mongodin *et al.*, 2005). Thus, these organisms may be indigenous to oil reservoirs (Grassia *et al.*, 1996). However, an alternative hypothesis of human introduction during oil exploration is not ruled out (Ollivier and Alazard, 2010; Struchtemeyer *et al.*, 2011).

Bacteria from the genus *Thermotoga* are hyperthermophilic anaerobes and have been detected and isolated from oil reservoirs, as well as from geothermally heated sea floors and continental hot springs (Huber and Hannig, 2006). Analyses of fosmid clones from *Thermotoga* isolates from geothermally heated sea floors of the Mediterranean and Atlantic (*Thermotoga maritima* MSB8 and *Thermotoga* sp RQ2), and the Kubiki oil reservoir in Japan (*Thermotoga petrophila* and *Thermotoga naphthophila*) revealed frequent recombination involving *Thermotoga* spp. across geographic regions (Nesbø *et al.*, 2006), suggesting that these bacteria are not isolated. This raises the question of whether the subsurface isolates are recent invaders from geographically adjacent marine populations (or vice versa), or whether the two ecologically different habitats harbor distinct subpopulations adapted to their niches. To address this, we investigated the phylogeographic patterns of the *Thermotoga* spp. by analyzing the genomes of seven new isolates from the Troll oil reservoir in the North Sea and shallow marine hydrothermal vents at the Kuril Islands (Supplementary Table S1) together with the genomes of the *Thermotoga* isolates from the fosmid-clone study (Nesbø *et al.*, 2006) and three *Thermotoga*-containing metagenomes. We find evidence for subsurface and marine subpopulations, although the genomes from both habitat types show signs of

extensive DNA exchange across ecological boundaries. Therefore, our observations suggest that oil reservoir populations are not isolated and that the reservoirs were likely colonized from marine or already existing subsurface populations.

Materials and methods

Genomic data

Thermotoga spp. were isolated from oil production fluids of the Troll B and Troll C oil platforms, as described in Dippio *et al.* (2009), and from shallow marine hydrothermal vents in the Kuril islands, as described in Svetlichny *et al.* (1991). Detailed description of the isolation procedure, DNA isolation, genome and fosmid clone sequencing and assembly and annotation is provided in Supplementary Information. The genomes and fosmid clones are available in GenBank under accession numbers CP003408, CP003409, AJ1101000000, JSFG01000000, JSFI01000000, JSFH01000000, JSFJ01000000, KP239990 and KP229298. Description of the analysis and assembly of *Thermotoga* DNA from metagenomes can also be found in Supplementary Information.

Gene content and genome alignments

To determine the gene pool shared among all genomes, a bidirectional top-scoring BLASTP (Altschul, 1997) approach (E -value $<10^{-5}$) was used to identify the pairwise orthologs. Unique genes and genes shared only by pairs of genomes were identified using the *phylogenetic profile* tool in IMG (Markowitz *et al.*, 2014), with a percent identity of $>80\%$ and an E -value of $<10^{-5}$.

Pan-genome calculations were performed in Pan-seq (Laing *et al.*, 2010) using a fragment size of 500 bp and 80% identity cutoff for the analyses of TM-group genomes, and 90% identity cutoff for the analyses of TM-group and genomes assembled from metagenomes (to confidently exclude possible contamination). The data matrices of shared core SNPs and shared 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 colinear block (LCB) scores. LCB positions with gaps were removed in GBlocks (Castresana, 2000). The edited LCBs were concatenated in Geneious 6 (www.geneious.com).

Quartet decomposition (QD) analysis

Positional orthologs (referred to as gene families) were extracted from the LCBs. Only gene families present in at least four genomes were retained for further analyses. The gene families were aligned

in ClustalW version 2 (Larkin *et al.*, 2007), and phylogenetic trees were reconstructed from 100 bootstrap samples in RAxML version 7.3.6 (Stamatakis, 2006) under GTR + Γ model. Resulting trees were analyzed using the Quartet Decomposition server (Mao *et al.*, 2012), removing quartets with internal branch <0.02 substitutions/site or with any external branch 10 times longer than the internal branch. A detailed description of the analyses is in Supplementary Information.

Recombination detection

The relative rate of recombination to mutation and average recombination tract length were assessed using the *pairwise* program and likelihood look-up tables generated by the *complete* program in the LDhat package (McVean *et al.*, 2002; Jolley, 2004). Detection of recombinant fragments was carried out in RDP version 4.33 (Martin *et al.*, 2010) and LikeWind (Archibald and Roger, 2002). A detailed description of the analyses is in Supplementary Information.

Results and discussion

Limited genomic diversity within sampled sites

In the remainder of the article, we denote the newly sequenced (Supplementary Table S1) and other isolates with genomes closely related to *T. maritima*, *T. petrophila* and *T. naphthophila* as the 'TM-group' of *Thermotoga* spp.

Examination of the genomes for rearrangements, indels and SNPs reveal very low diversity among the Troll oil reservoir genomes (Supplementary Table S2). Whole-genome alignments resulted in three LCBs, with only a single 85 kb genome rearrangement. The nucleotide identity within each of the three LCBs ranges from 98.7% to 99.4%. Apart from variation in CRISPR (clustered regularly interspaced short palindromic repeat) regions (Supplementary Figure S1), the largest source of diversity is indels (3–23 nt) in 7 genes (Supplementary Table S3). On average, only 100 SNPs were detected between the Troll genomes from different platforms (Supplementary Table S2).

Troll genomes from platform B, *Thermotoga* sp. strains TBGT17.6.5 and TBGT17.6.6, differ only by seven SNPs that are located within a hypothetical gene flanking a CRISPR region. Comparison of the *Thermotoga* sp. TBGT17.6.5 genome with fosmid clones of two additional Troll B isolates also displayed only 6 and 4 SNPs over 42 148 and 33 295 bp of fosmid data, respectively. Similarly, there are only 54 SNPs between the two Troll genomes from the platform C. Mapping sequences from a Troll C metagenome (data not shown) onto the *Thermotoga* sp. CELL2 genome showed 99.6% identity over 364 180 bp (in contrast to 98.5% identity to the *T. maritima* MSB8 genome over 294 382 bp). Our low diversity observations are

concordant with metagenomic studies of the Troll reservoir (data not shown) and two other oil reservoirs in this region (Lewin *et al.*, 2014). Nevertheless, Troll isolates differ in fast-evolving CRISPR loci (Supplementary Figure S1) (Mongodin *et al.*, 2005), suggesting that these are distinct lineages that have encountered different exogenous phage and mobile DNA.

Among three genomes of the Kuril Islands isolates, *Thermotoga* sp. EMP and 2812B differ only by 23 SNPs (Supplementary Table S2) and a 693-bp insertion in the EMP_01477 gene encoding flagellar hook-length control protein. The third Kuril Islands genome (*Thermotoga* sp. Mc24) showed much higher divergence to the other two genomes, as well as to the rest of TM-group genomes (Supplementary Table S2).

Although we report overall low within-geographic-region diversity of the two sampled sites, this is not a universal feature of the TM-group (Supplementary Table S2). The genomes of *T. petrophila* and *T. naphthophila* from the Kubiki oil reservoir in Niigata, Japan (Takahata *et al.*, 2008), are only 96.7% identical. Unlike the Troll reservoir, the Kubiki oil reservoir has been flooded with water to enhance oil production (Takahata *et al.*, 2000, 2008) that could have recently introduced new genetic diversity. The Troll reservoir is also older than Kubiki oil reservoir, as the sediments of the latter were formed in early Pliocene and late Miocene, 5–7 Myr ago (Kawai and Totani, 1971; Chakhmankhchev *et al.*, 1996). Thus, the Troll microbiome may also have been isolated for a longer period of time and lost more diversity because of genetic drift. Consistent with the drift scenario, the genomes of the Troll reservoir isolates are smaller and encode fewer genes than any other known *Thermotoga* genome (Supplementary Table S1), a feature also observed in a *Thermotoga* sp. isolate from an Alaskan oil reservoir (Mongodin *et al.*, 2005).

High gene content similarity across all TM-group genomes

All of the analyzed genomes share >90% of their genes, and the genomes from the same type of environment (that is, either oil reservoir or marine sediments) share larger proportion of genes (Figure 1). As the four Troll oil reservoir genomes and two of Kuril Islands genomes exhibit very high within-group DNA sequence identity and gene content similarity, only one representative genome from each group was selected for all further analyses (*Thermotoga* sp. Cell2 and *Thermotoga* sp. 2812B). Genome-wide alignment of seven representative TM-group genomes (*T. maritima* MSB8, *T. petrophila* RKU1, *T. naphthophila* RKU10, *T. sp.* RQ2, *Thermotoga* sp. Cell2, *Thermotoga* sp. 2812B and *Thermotoga* sp. Mc24) resulted in 16 LCBs spanning 1 543 882 bp ('core' genome), including 83–89% of the individual genomes' DNA. The observed

divergence within the 'core' genome is on average 4% (Supplementary Table S2).

Examination of the accessory 889 066 bp of the TM-group pan-genome showed that individual genomes carry between 29 and 64 unique genes found only in one genome, accounting for 81 630 bp in total (Supplementary Table S4). Also, 134 genes are shared by only two genomes, spanning an additional 43 873 bp of the accessory pan-genome (Supplementary Table S4). Thus, the accessory pan-genome is mainly composed of genes shared by at least three genomes. When genomes are grouped according to the environment type, only two genes are specific to isolates from oil reservoirs (hypothetical proteins; CELL2_08715 and CELL2_08720 in *Thermotoga* sp. CELL2), whereas nine genes are specific to isolates from marine sediments (a phosphotransferase domain-containing (TM0559), a hypothetical (TM1012) and 7 CRISPR-associated proteins (TM1791, TM1793–TM1795 and TM1800–TM1802 in *T. maritima* MSB8)). Future biochemical characterization of these proteins is needed to elucidate any contributions they may have to niche adaptation in the two types of environment.

Lack of consistent phylogenetic signal suggests rampant recombination within the TM-group

In the 'shared DNA fragments' network, the genomes from 'oil reservoirs' and 'marine sediments' form separate groups, with the 'marine sediments' genomes further differentiated according to their geographic proximity (Figure 2a). The 'core SNPs' network (Figure 2b), on the other hand, does not show such clear pattern: *Thermotoga* sp. RQ2 clusters with *Thermotoga* sp. CELL2 (shared 'geographical proximity') and *T. maritima* MSB8 clusters with *Thermotoga* sp. 2812B (shared 'ecological niche'). In addition, phylogenetic histories reconstructed from typical marker genes (Supplementary Figure S2) and from a concatenation of the aligned genome LCBs were incongruent (see below). QD analysis (Zhaxybayeva *et al.*, 2006; Mao *et al.*, 2012) of phylogenetic histories of 1728 gene families present in at least four of the seven TM-group genomes reveals no support of any unique bifurcating phylogenetic tree by the majority of the gene families (Figure 2c and Supplementary Figure S3a). Moreover, evolutionary histories of 857 gene families were at least partially incongruent with the phylogenetic signal of the plurality of gene families. Further screening of gene families resulted in 69 and 25 gene families that strongly support 'ecological niche' and 'geographical proximity' divisions, respectively (Supplementary Figure S3b). In phylogenetic trees of 33 of the 69 (47%) 'ecology-specific' families, the observed groupings are likely because of recent recombination events, as the involved taxa have identical (or nearly identical) DNA sequences. Some of these genes could represent habitat-specific alleles spreading through the population via homologous recombination and selection, as

Figure 1 Pairwise comparisons of shared gene pool across 11 TM-group genomes. Genes are designated as shared if they are a bidirectional BLASTP hits (E -value $< 10^{-5}$). The genomes are classified according to geographic location (Atlantic vs Pacific) and environment type (oil reservoir vs marine vent). The heatmap indicates that the isolates from the same environment type share more genes than the isolates in geographic proximity. Comparisons between genomes from the same type environment are outlined in bold.

Figure 2 Phylogenetic relationships among seven TM-group genomes based on (a) presence/absence of shared 500bp genomic fragments, (b) core SNPs and (c) plurality gene families from the QD analysis. The genomes are classified according to geographic location (Atlantic vs Pacific) and environment type (oil reservoir vs marine vent). For networks (a) and (b), the data were obtained using PanSeq (Laing *et al.*, 2010). Core SNPs were required to be present in all 7 genomes, whereas the genomic fragments were considered shared if they were at least 80% identical. Networks were calculated in SplitsTree using NeighborNet algorithm (Huson and Bryant, 2006) from uncorrected distances (a, b) and quartet topologies supported by plurality of gene families identified in the QD analysis (c).

observed in *Sulfolobus* and *Vibrio* populations (Shapiro and Polz, 2014). In 27 of the 33 gene families, these recent recombination events are observed between *Thermotoga* spp. from oil reservoirs in North Sea and Japan, ~8000 km apart. Only 8 of the 25 'geography-specific' families (37%) show evidence of recent recombination, and in 4 cases recombination is between the two Kuril Islands isolates, *Thermotoga* sp. Mc24 and 2812B. Although some phylogenetic conflict may be a result of uncertainty of phylogenetic reconstruction, incongruence because of gene flow will manifest itself in recombination events recorded among the TM-group lineages.

Quantification of recombination events reveals high recombination to mutation ratios

In bacteria, recombination events are best described as a gene conversion process (McVean *et al.*, 2002). Thus, for each LCB in the alignment of seven TM-group genomes, we calculated the population mutation rate (θ) and the gene conversion parameter (γ) that can be interpreted as the population rate of recombination due to gene conversion between two distantly linked loci (McVean *et al.*, 2002). Across the TM-group the estimated average recombination tract length is 6800 bp (ranging between 2000 and 14600 bp), and the average θ and γ are 0.045 (0.019–0.058) and 2.9 (1.1–3.8), respectively (Supplementary Table S5). The resulting high γ/θ ratios of 24–109 is indicative of high levels of recombination, consistent with the phylogenetic analyses above. The high level of recombination suggests that the low genomic diversity of TM-group bacteria within geographic populations could in part be because of genome homogenization via recombination.

The search for specific recombination breakpoints resulted in 330 and 471 recombination events detected in LikeWind (Archibald and Roger, 2002) and RDP (Martin *et al.*, 2010), respectively (Figures 3 and 4 and Supplementary Table S6). As both algorithms rely on phylogenetic evidence, the number of recombination events is likely underestimated because of failure to detect recombination between sister taxa. On average, 67 recombination events per genome are predicted by RDP, ranging from 93 in *Thermotoga* sp. Mc24 to 48 in *Thermotoga* sp. CELL2 (Supplementary Table S6).

Recombination events were detected among all isolates regardless of their geographic location (Figure 3 and, Supplementary Table S6). However, surprisingly few recombination events were inferred among some genomes of geographically proximal isolates (that is, with *T. petrophila* and *T. naphthophila* as recipients and *Thermotoga* sp. 2812B as donor, and with *Thermotoga* sp. RQ2 as recipient and *T. maritima* MSB8 as donor). This was supported by the QD analysis where the two quartets with the highest support and lowest conflict cluster *T. maritima* MSB8 and *Thermotoga*

sp. 2812B. Interestingly, *Thermotoga* sp. RQ2 has obtained most DNA from the *T. naphthophila* lineage, whereas *T. naphthophila* RKU10 and *T. petrotoga* RKU1 have received most DNA from *Thermotoga* sp. Mc24. For the remaining genomes, most of the recombinant DNA comes from strains from the same type of environment. This was particularly pronounced for *Thermotoga* sp. CELL2 genome, which is predicted to have received much more DNA from the other 'oil reservoir' genomes (30 events) than from the geographically closer *T. maritima* MSB8 (2 events).

Combined evidence from the QD analysis, recombination detection and the examination of the pan-genome suggests high levels of reticulate evolution among TM-group members. Therefore, we infer that TM-group bacteria form a global interacting collective. Similar mixing across large geographic distances have been observed in several other taxonomic groups such as haloarchaea, *Prochlorococcus* and *Vibrio* (reviewed in Shapiro and Polz, 2014).

However, such global connectivity does not exclude the existence of local and ecology-driven subpopulations. We find that TM-group bacteria can be divided into two 'habitat-based' subpopulations ('marine sediments' and 'subsurface') that share more DNA with each other than they do with isolates of different habitat types from the same geographical region. These findings could be explained by either preferential gene exchange among bacteria from the same type of environment or preferential retention of acquired genes with a selective advantage in the ecological niche. The high level of recombination among genomes from oil reservoirs of different geological ages (Kubiki and Troll) is not compatible with isolation after burial. Although cells buried in the sediments may contribute to the population's genetic makeup, our data support the 'colonization' hypothesis, under which the oil reservoirs are colonized either from an already existing subsurface population or from surrounding marine populations. In addition, the observation of distinct habitat-based subpopulations attests to indigeneity of *Thermotoga* in Kubiki and Troll reservoirs.

Other sites harbor *Thermotoga* spp. populations with similar properties

To investigate whether the observed patterns of extensive recombination and low within-population diversity can be generalized for other TM-group populations, we searched for publicly available metagenomes that harbor TM-group bacteria, using as a criterion >90% identity to *Thermotoga* genus genes. We identified three such sample sites (Supplementary Table S7): San Juan basin coal bed production water (An *et al.*, 2013), a 'continental hot spring' fed from a subterranean reservoir (Great Boiling Spring (GBS), Nevada) (Costa *et al.*, 2009) and the off-shore Tuna oil reservoir in the

Figure 3 Visualization of recombination events detected among seven TM-group genomes. The genomes are color coded and arranged on the circumference of a circle. The recombination events with predicted donor and recipient are shown as lines connecting the two genomes and the locations of recombined regions. The line color reflects the donor lineage. The predicted events were required to be significant in three out of four algorithms in RDP (Martin *et al.*, 2010). The diagram was generated using Circos (Krzywinski *et al.*, 2009).

Figure 4 Detected recombination events in seven TM-group genomes. Sliding-window phylogenetic analysis of the alignment of concatenated LCBs was performed in the LikeWind program (Archibald and Roger, 2002) using 1000bp windows selected in 100bp increments. (a) Phylogenetic tree reconstructed from the whole alignment. (b, c) Two examples of significantly different tree topologies. (d) Across-the-alignment dynamics of $\Delta \log L$, the difference between the log-likelihood of the maximum likelihood (ML) tree of the 1000bp window and the ML tree reconstructed from the entire alignment. The x axis shows the position along the concatenated alignment. Peaks correspond to tree topologies of sliding alignment windows that are very different from the whole alignment tree topology depicted in (a). The parametric bootstrapping analysis (Archibald and Roger, 2002) indicated that values of $\Delta \log L > 12.7$ correspond to significantly different tree topologies, and the x axis crosses the y axis at that value. Position of the alignment is color coded by LCB.

Bass Strait, Gippsland Basin, Australia (Sutcliffe *et al.*, 2013).

Only 16 protein-coding sequences in the San Juan basin metagenome are TM-group like. In

phylogenetic analyses, nine of them cluster with their homologs from the geographically close GBS. Moreover, on 11 of 16 phylogenetic trees, one or more of the 'oil reservoir' isolates are the sister

taxon either to the CG7 homolog or to the group of CG7 and GBS homologs, supporting a close phylogenetic relationship among subsurface TM-group populations.

Three cellulolytic *in situ* enrichment cultures from the GBS site include >2000 *Thermotoga*-like contigs with $\geq 90\%$ similarity to the genomes used in this study (Supplementary Table S7), consistent with 16S rRNA analysis (Peacock *et al.*, 2013). Comparisons of *Thermotoga*-like sequences among the three cultures revealed that they are >98% identical across homologous contigs >5 kb in length, again demonstrating low within-site diversity for TM-group bacteria.

Our assembly of a 2.1-Mb draft genome from a GBS metagenome (referred as *Thermotoga* sp. GBS) and the already assembled draft genome of *Thermotoga* sp. A7A from the Tuna oil reservoir metagenome (Sutcliffe *et al.*, 2013) allowed us to repeat the genome-wide phylogenetic analyses with 9 TM-group genomes. On the 'core SNPs' network, *Thermotoga* sp. GBS is most closely linked to *Thermotoga* sp. CELL2 (Figure 5). The close relationship between *Thermotoga* sp. GBS and *Thermotoga* sp. CELL2, which were isolated ~ 7500 km apart in Nevada and the North Sea, respectively, demonstrates that *Thermotoga* sp. CELL2 does not represent a lineage specific to either oil or ancient North Sea sediments. A shallow ocean covered Nevada at the time the Troll oil reservoir sediments were laid down (Blakey, 2014), and thus these isolates could have originated from the same marine population and, according to the 'burial and isolation' hypothesis, stayed isolated ever since. However, the QD analysis identified 133 gene families that place *Thermotoga* sp. GBS and *Thermotoga* sp. CELL2 in different groups with >90% bootstrap support (Supplementary Figure

S4). Many of these gene families are likely a result of gene exchange with both marine and other subsurface bacteria, suggesting that contact among lineages from the two environments is an on-going process. Hence, regardless of how the bacteria entered the oil reservoir sediments, they do not appear to have stayed isolated.

The close connection of *Thermotoga* sp. A7A to the two genomes from 'marine sediments' (*T. maritima* MSB8 and *Thermotoga* sp. 2812B; Figure 5) suggests that the *Thermotoga* population has entered the Tuna reservoir recently from a marine environment. The temperature of the reservoir is reported to be 105 °C (Sutcliffe *et al.*, 2013), which is considerably higher than the proposed 80–90 °C threshold temperature for sustaining life in oil reservoirs (Stetter and Huber, 1999; Wilhelms *et al.*, 2001), suggesting that *Thermotoga* sp. A7A may be inhabiting the petroleum pipelines, where the temperature is lower. The opposite direction of the migration, from subsurface to the marine habitat, is supported by the position of *Thermotoga* sp. RQ2 in the 'core SNPs' network (Figure 5). We conjecture that this migration happened relatively recently, consistent with a <8 Myr (Abdel-Monem *et al.*, 1975) volcanic origin of the Azores.

Potential causes of low diversity

Not only do we observe low genomic diversity of *Thermotoga* within sites, but the 'core SNP' network shows that *Thermotoga* sp. GBS (Nevada; hot spring), *Thermotoga* sp. CELL2 (North Sea; oil reservoir) and *Thermotoga* sp. RQ2 (Azores; marine sediment) are on shorter branches, and hence have smaller number of underlying nucleotide substitutions (Figure 5; 98.2–98.5% genome-wide nucleotide identity and 11 254–14 446 core SNPs). Several factors may contribute to the observed low diversity. Given the slower growth rates (and therefore lower mutation supply rates) in the subsurface environments, the limited diversity may simply be because of insufficient time to accumulate substitutions. In addition, subsurface sediments, including oil reservoirs, are among the most stable environments on Earth both spatially and temporally, as temperature and pressure stay constant at a given depth (S Larter, personal communication). This stability may contribute to low within-population diversity, as environmental variability is one of the main drivers in maintaining genetic diversity of a population (Kassen, 2002). If the diversity of phages that can infect these bacteria is smaller than what is seen in mesophilic environments (Weinberger *et al.*, 2012), the selective advantage of rare genotypes would be lower. Finally, the inferred high recombination incidence is likely to further lower the divergence among TM-group genomes.

Why do we see such high levels of recombination among *Thermotoga* spp.? Hyperthermophiles, like *Thermotoga*, have to constantly repair their DNA

Figure 5 Phylogenetic relationships of nine TM-group genomes. The network was calculated using core SNPs as described in Figure 2 legend, except the included fragments were required to be at least 90% identical. All nodes of the network have at least 96% bootstrap support. The plurality network reconstructed from the QD analysis has the same topology (not shown).

because of the damaging effects of high temperatures (Grogan, 2000). Extensive recombination (repair) has been observed in other hyperthermophilic microorganisms such as *Pyrococcus furiosus* (Diruggiero *et al.*, 1997), *Sulfolobus islandicus* (Whitaker *et al.*, 2005) and *Persephonella* (Mino *et al.*, 2013). Thus, the high levels of recombination observed in *Thermotoga* might be a by-product of high levels of DNA repair (Johnston *et al.*, 2014).

Scenarios for gene dispersal across subsurface populations

Our analysis suggests larger gene flow among the subsurface populations than between subsurface and marine populations, but what would be the mechanism of gene flow across geographically isolated oil reservoirs? The gene flow may be mediated entirely via the surface, with preferential retention of genes with a selective advantage in oil reservoirs. Indeed, dispersal of thermophilic organisms from the subsurface has been suggested to occur via deep-sea hydrothermal plumes (Summit and Baross, 1999; Dick *et al.*, 2013). Alternatively, some of the contact may occur strictly within the subsurface. Under this scenario, the subsurface populations belong to a large, sparsely populated, slow-growing collective of subsurface thermophiles that exists in areas with favorable conditions (anoxic, 55–90 °C, organic material and so on). The existence of such subsurface communities has been inferred from detection of thermophiles in hydrothermal vent plumes associated with seismic events (Summit and Baross, 1999) and from direct sampling of crustal fluids (Huber *et al.*, 2006). Organisms from this collective may have colonized the oil reservoirs and GBS, as illustrated in Supplementary Figure S5. Under this scenario, the exchange between oil reservoirs is mediated by the exchange with microorganisms from surrounding sediments.

Concluding remarks

Our genomic and metagenomic analysis suggests that different subsurface populations of *Thermotoga* are not isolated and thus favors ‘colonization’ over the ‘burial and isolation’ scenario. Buried marine *Thermotoga* may instead be a substantial source of marine immigrants entering the subsurface collective. The phylogenetic analyses further suggest that some *Thermotoga* have likely colonized their oil reservoirs a long time ago, whereas others have entered reservoirs recently, probably during the commercial development. In both cases, gene flow appears to be an ongoing process. Although some gene flow likely involves dispersal via the surface (for example, burial, oil seeps and hydrothermal vent plumes), followed by subsequent selection for genes beneficial for subsurface life, we suggest that within-subsurface gene exchange is nonnegligible.

Future investigations are needed to determine how such gene flow is achieved.

The level of diversity among *Thermotoga* spp. genomes is strikingly low when compared with, for instance, marine *Vibrio* populations, in which individual cells may carry up to 200 genes unique to their genomes (Cordero and Polz, 2014). This might be a result of the extreme environment *Thermotoga* spp. inhabit. As discussed above, low diversity may be generated by several alternative, but not mutually exclusive, mechanisms. Investigating genome diversity in *Thermotogae* living at less extreme temperatures, such as *Mesotoga* and *Petrotoga* lineages, may help elucidate the effect of temperature on population-level genome diversity and phylogeography.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

This work is supported by Dartmouth College start-up funds to OZ and by a Research Council of Norway award (Project No. 180444/V40) to CLN. We thank Dr Steve Larter and Dr W Ford Doolittle for comments and discussions.

References

- Abdel-Monem AA, Fernandez LA, Boone GM. (1975). K-Ar ages from the eastern Azores group (Santa Maria, Miguel and the Formigas Islands). *Lithos* **16**: 247–254.
- Adda GW. (2012). Hydrocarbon generation and migration from Jurassic source rocks in the northern North Sea. *Master of Sciences Thesis. Norwegian University of Science and Technology*.
- Altschul S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.
- An D, Caffrey SM, Soh J, Agrawal A, Brown D, Budwill K *et al.* (2013). Metagenomics of hydrocarbon resource environments indicates aerobic taxa and genes to be unexpectedly common. *Environ Sci Technol* **47**: 10708–10717.
- Archibald JM, Roger AJ. (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.
- Bennett B, Adams JJ, Gray ND, Sherry A, Oldenburg TBP, Huang H *et al.* (2013). The controls on the composition of biodegraded oils in the deep subsurface – Part 3. The impact of microorganism distribution on petroleum geochemical gradients in biodegraded petroleum reservoirs. *Org Geoch* **56**: 94–105.
- Biddle JF, White JR, Teske AP, House CH. (2011). Metagenomics of the subsurface Brazos-Trinity Basin (IODP site 1320): comparison with other sediment and pyrosequenced metagenomes. *ISME J* **5**: 1038–1047.

- Blakey R (2014). Global and regional paleogeographic maps. <http://cpgeosystems.com/paleomaps.html> (Accessed 15 July 2014).
- Bonch-Osmolovskaya EA, Miroshnichenko ML, Lebedinsky AV, Chernyh NA, Nazina TN, Ivoilov VS et al. (2003). Radioisotopic, culture-based, and oligonucleotide microchip analyses of thermophilic microbial communities in a continental high-temperature petroleum reservoir. *Appl Environ Microbiol* **69**: 6143–6151.
- Cappelletti M, Zannoni D, Postec A, Ollivier B. (2014). Members of the order Thermotogales: From microbiology to hydrogen production. In: Zannoni D, Philippis RD (eds) *Microbial BioEnergy: Hydrogen Production, Advances in Photosynthesis and Respiration* vol. 38. Springer Netherlands: Dordrecht, pp 197–224.
- Castelle CJ, Hug LA, Wrighton KC, Thomas BC, Williams KH, Wu D et al. (2013). Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nat Commun* **4**: 2120.
- Castresana J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540–552.
- Chakhmankhchev A, Suzuki N, Suzuki M, Takayama K. (1996). Biomarker distributions in oils from the Akita and Niigata Basins, Japan. *Chem Geol* **133**: 1–14.
- Cordero OX, Polz MF. (2014). Explaining microbial genomic diversity in light of evolutionary ecology. *Nat Rev Microbiol* **12**: 263–273.
- Costa KC, Navarro JB, Shock EL, Zhang CL, Soukup D, Hedlund BP. (2009). Microbiology and geochemistry of great boiling and mud hot springs in the United States Great Basin. *Extremophiles* **13**: 447–459.
- Darling AE, Mau B, Perna NT. (2010). progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* **5**: e11147.
- Dick GJ, Anantharaman K, Baker BJ, Li M, Reed DC, Sheik CS. (2013). The microbiology of deep-sea hydrothermal vent plumes: ecological and biogeographic linkages to seafloor and water column habitats. *Front Microbiol* **4**: 124.
- 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 Micro* **59**: 2991–3000.
- Diruggiero J, Santangelo N, Nackerdien Z, Racvel J, Robb FT. (1997). Repair of extensive ionizing-radiation DNA damage at 95 degrees C in the hyperthermophilic archaeon *Pyrococcus furiosus*. *J Bact* **179**: 4643–4645.
- Engelhardt T, Kallmeyer J, Cypionka H, Engelen B. (2014). High virus-to-cell ratios indicate ongoing production of viruses in deep subsurface sediments. *ISME J* **8**: 1503–1509.
- Fry JC, Horsfield B, Sykes R, Cragg BA, Heywood C, Kim GT et al. (2009). Prokaryotic populations and activities in an interbedded coal deposit, including a previously deeply buried section (1.6–2.3 km) above ~ 150 Ma basement rock. *Geomicrobiol J* **26**: 163–178.
- Gautier D. (2005). Kimmeridgian Shales total petroleum system of the North Sea Graben Province. *US Department of the Interior, US Geological Survey, Bulletin 2204-C* 1–29.
- Grassia GS, McLean KM, Glénat P, Bauld J, Sheehy AJ. (1996). A systematic survey for thermophilic fermentative bacteria and archaea in high temperature petroleum reservoirs. *FEMS Microbiol Ecol* **21**: 47–58.
- Grogan DW. (2000). The question of DNA repair in hyperthermophilic archaea. *Trends Microbiol* **8**: 180–185.
- Head IM, Jones M, Larter SR. (2003). Biological activity in the deep subsurface and the origin of heavy oil. *Nature* **426**: 344–352.
- Huang H, Larter SR. (2005). Biodegradation of petroleum in subsurface geological reservoirs. In Ollivier B, Magot M (eds) *Petroleum Microbiology*. ASM Press: Washington, DC, pp 91–121.
- Huber JA, Johnson HP, Butterfield DA, Baross JA. (2006). Microbial life in ridge flank crustal fluids. *Environ Microbiol* **8**: 88–99.
- Huber R, Hannig M. (2006). Thermotogales. In Huber R, Hannig M (eds) *The Prokaryotes*. Springer New York: New York, NY, pp 899–922.
- Huson DH, Bryant D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* **23**: 254–267.
- Johnston C, Martin B, Fichant G, Polard P, Claverys J-P. (2014). Bacterial transformation: distribution, shared mechanisms and divergent control. *Nat Rev Microbiol* **12**: 181–196.
- Jolley KA. (2004). The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. *Mol Biol Evol* **22**: 562–569.
- Jørgensen BB. (2011). Deep seafloor microbial cells on physiological standby. *Proc Natl Acad Sci USA* **108**: 18193–18194.
- Kallmeyer J, Pockalny R, Adhikari RR, Smith DC, D'Hondt S. (2012). Global distribution of microbial abundance and biomass in seafloor sediment. *Proc Natl Acad Sci USA* **109**: 16213–16216.
- Kassen R. (2002). The experimental evolution of specialists, generalists, and the maintenance of diversity. *J Evol Biol* **15**: 173–190.
- Kawai K, Totani S. (1971). Relationship between crude-oil properties and geology in some oil and gas fields in the Niigata basin, Japan. *Chem Geol* **8**: 219–246.
- Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D et al. (2009). Circos: An information aesthetic for comparative genomics. *Genome Res* **19**: 1639–1645.
- Laing C, Buchanan C, Taboada EN, 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.
- Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.
- Lewin A, Johansen J, Wentzel A, Kotlar HK, Drabløs F, Valla S. (2014). The microbial communities in two apparently physically separated deep subsurface oil reservoirs show extensive DNA sequence similarities. *Environ Microbiol* **16**: 545–558.
- Magot M. (2005). Indigenous microbial communities in oil fields. In Ollivier B, Magot M (eds) *Petroleum Microbiology*. ASM Press: Washington, DC, pp 21–33.
- Mao F, Williams D, Zhaxybayeva O, Poptsova M, Lapierre P, Gogarten J et al. (2012). Quartet decomposition server: a platform for analyzing phylogenetic trees. *BMC Bioinformatics* **13**: 123.
- Markowitz VM, Chen I-MA, 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–D567.

- Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefevre P. (2010). RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**: 2462–2463.
- McVean G, Awadalla P, Fearnhead P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**: 1231–1241.
- Mino S, Makita H, Tomohiro T, Miyazaki J, Kato S, Watanabe H *et al.* (2013). Biogeography of *Persephonella* in deep-sea hydrothermal vents of the Western Pacific. *Front Microbiol* **4**: 1–12.
- Mongodin EF, Hance IR, DeBoy RT, Gill SR, Daugherty S, Huber R *et al.* (2005). Gene transfer and genome plasticity in *Thermotoga maritima*, a model hyperthermophilic species. *J Bact* **187**: 4935–4944.
- Morono Y, Terada T, Nishizawa M, Ito M, Hillion F, Takahata N *et al.* (2011). Carbon and nitrogen assimilation in deep subseafloor microbial cells. *Proc Natl Acad Sci USA* **108**: 18295–18300.
- Nesbø CL, Dlutek M, Doolittle WF. (2006). Recombination in *Thermotoga*: implications for species concepts and biogeography. *Genetics* **172**: 759–769.
- Ollivier B, Alazard D. (2010). The oil reservoir ecosystem. In Timmis KN (ed) *Handbook of Hydrocarbon and Lipid Microbiology*. Springer: Berlin/Heidelberg, pp 2262–2268.
- Orsi WD, Edgcomb VP, Christman GD, Biddle JF. (2013). Gene expression in the deep biosphere. *Nature* **499**: 205–208.
- Peacock JP, Cole JK, Murugapiran SK, Dodsworth JA, Fisher JC, Moser DP *et al.* (2013). Pyrosequencing reveals high-temperature cellulolytic microbial consortia in Great Boiling Spring after in situ lignocellulose enrichment. *PLoS One* **8**: e59927.
- Planckaert M. (2005). Oil reservoirs and oil production. In Ollivier B, Magot M (eds) *Petroleum Microbiology*. ASM Press: Washington, DC, pp 3–19.
- Schrenk MO, Huber JA, Edwards KJ. (2010). Microbial provinces in the subseafloor. *Ann Rev Mar Sci* **2**: 279–304.
- Shapiro BJ, Polz MF. (2014). Ordering microbial diversity into ecologically and genetically cohesive units. *Trends Microbiol* **22**: 235–247.
- Slobodkin AI, Jeanthon C, L'Haridon SXP, Nazina T, Miroshnichenko M, Bonch-Osmolovskaya E. (1999). Dissimilatory reduction of Fe (III) by thermophilic bacteria and archaea in deep subsurface petroleum reservoirs of Western Siberia. *Curr Microbiol* **39**: 99–102.
- Stamatakis A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.
- Stetter KO, Huber R. (1999). The role of hyperthermophilic prokaryotes in oil fields. In Bell CR, Brylinsky M, Johnson–Green P (eds). *Microbial Biosystems: New Frontiers, Proceedings of the 8th International Symposium on Microbial Ecology*. Atlantic Canada Society for Microbial Ecology: Halifax, Canada, pp 369–375.
- Struchtemeyer CG, Davis JP, Elshahed MS. (2011). Influence of the drilling mud formulation process on the bacterial communities in thermogenic natural gas wells from the Barnett Shale. *Appl Environ Microbiol* **77**: 4744–4753.
- Summit M, Baross JA. (1999). Thermophilic subseafloor microorganisms from the 1996 North Gorda Ridge eruption. *Deep Sea Res II* **45**: 2751–2766.
- Sutcliffe B, Midgley DJ, Rosewarne CP, Greenfield P, Li D. (2013). Draft genome sequence of *Thermotoga maritima* A7A reconstructed from metagenomic sequencing analysis of a hydrocarbon reservoir in the Bass Strait, Australia. *Genome Announc* **1**: e00688–13.
- Svetlichny VA, Sokolova TG, Gerhardt M, Kostrikina NA, Zavarzin GA. (1991). Anaerobic extremely thermophilic carboxydophilic bacteria in hydrotherms of Kuril Islands. *Microb Ecol* **21**: 1–10.
- Takahata Y, Nishijima M, Hoaki T, Maruyama T. (2000). Distribution and physiological characteristics of hyperthermophiles in the Kubiki oil reservoir in Niigata, Japan. *Appl Environ Microbiol* **66**: 73–79.
- Takahata Y, Nishijima M, Hoaki T, Maruyama T. (2008). *Thermotoga petrophila* sp. nov. and *Thermotoga naphthophila* sp. nov., two hyperthermophilic bacteria from the Kubiki oil reservoir in Niigata, Japan. *Int J Syst Evol Micr* **51**: 1901–1909.
- Weinberger AD, Wolf YI, Lobkovsky AE, Gilmore MS, Koonin EV. (2012). Viral diversity threshold for adaptive immunity in prokaryotes. *mBio* **3**: e00456–12.
- Whitaker RJ, Grogan DW, Taylor WJ. (2005). Recombination shapes the natural population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*. *Mol Biol Evol* **22**: 2354–2361.
- Wilhelms A, Larter SR, Head I, Farrimond P, di-Primio R, Zwach C. (2001). Biodegradation of oil in uplifted basins prevented by deep-burial sterilization. *Nature* **411**: 1034–1037.
- Zhaxybayeva O, Gogarten JP, Charlebois RLR, Doolittle WF, Papke RT. (2006). Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res* **16**: 1099–1108.

Supplementary Information accompanies this paper on The ISME Journal website (<http://www.nature.com/ismej>)