New Asgard archaea capable of anaerobic hydrocarbon cycling

Kiley W. Seitz[1], Nina Dombrowski[1,3], Laura Eme[2], Anja Spang[2,3], Jonathan Lombard[2], Jessica R. Sieber[4], Andreas P. Teske[5], Thijs J.G. Ettema[2], and Brett J. Baker[1*]

1. Department of Marine Science, University of Texas Austin, Port Aransas, TX 78373 2. Uppsala University, Uppsala Sweden 3. NIOZ, Royal Netherlands Institute for Sea Research, and Utrecht University, The Netherlands 4. University Minnesota Duluth, MN 5. Department of Marine Sciences, University of North Carolina, Chapel Hill, NC *Corresponding author

Large reservoirs of natural gas in the oceanic subsurface sustain a complex biosphere of anaerobic microbes, including recently characterized archaeal lineages that extend the potential to mediate hydrocarbon oxidation (methane and butane) beyond the Methanomicrobia. Here we describe a new archaeal phylum, Helarchaeota, belonging to the Asgard superphylum with the potential for hydrocarbon oxidation. We reconstructed Helarchaeota genomes from hydrothermal deep-sea sediment metagenomes in hydrocarbon-rich Guaymas Basin, and show that these encode novel methyl-CoM reductase-like enzymes that are similar to those found in butane-oxidizing archaea. Based on these results as well as the presence of several alkyl-CoA oxidation and Wood-Ljungdahl pathway genes in the Helarchaeota genomes, we suggest that members of the Helarchaeota have the potential to activate and subsequently anaerobically oxidize short-chain hydrocarbons. These findings link a new phylum of Asgard archaea to the microbial utilization of hydrothermally generated hydrocarbons, and extend this genomic blueprint further through the archaeal domain.

Short-chain alkanes, such as methane and butane, are abundant in marine sediments and play an important role in carbon cycling with methane concentrations of ~1 Gt being processed globally through anoxic microbial communities[1–3]. Until recently, archaeal methane cycling was thought to be limited to Euryarchaeota[4]. However, additional archaeal phyla, including Bathyarchaeota[5] and Verstraetarchaeota[6], have been shown to contain proteins with homology to the activating enzyme methyl-coenzyme M reductase (Mcr) and corresponding pathways for methane utilization. Furthermore, new lineages within the Euryarchaeota belonging to *Candidatus* Syntrophoarchaeum spp., have been shown to use methyl-CoM reductase-like enzymes for anaerobic butane oxidation[7]. Similar to methane oxidation in many ANME-1 archaea, butane oxidation in Syntrophoarchaeum is proposed to be enabled through a syntrophic interaction with sulfur reducing bacteria[7]. Metagenomic reconstructions of genomes recovered from deep-sea sediments from near 2000 m depth in Guaymas Basin (GB) in the Gulf of California have revealed the presence of additional uncharacterized alkyl methyl-CoM reductase-like enzymes in metagenome-assembled genomes within the Methanosarcinales (Gom-Arc1)[8]. GB is characterized by hydrothermal alterations that transform large amounts of organic carbon into methane, polycyclic aromatic hydrocarbons (PAHs), low-molecular weight alkanes and organic acids allowing for diverse microbial communities to thrive (Supplementary Table 1)[8–11].

Recently, genomes of novel clade of uncultured archaea, referred to as the Asgard superphylum that includes the closest archaeal relatives of eukaryotes, have been recovered from anoxic environments around the world[12–14]. Diversity surveys in anoxic marine sediments show that Asgard archaea appear to be globally distributed[9,11,12,13]. Based on phylogenomic analyses, Asgard archaea have been divided into four distinct phyla: Lokiarchaeota, Thorarchaeota, Odinarchaeota and Heimdallarchaeota, with the latter possibly representing the closest relatives of eukaryotes[12]. Supporting their close relationship to eukaryotes, Asgard archaea possess a wide repertoire of proteins previously thought to be unique to eukaryotes known as eukaryotic signature proteins (ESPs)[17]. These ESPs include homologs of eukaryotic proteins, which in eukaryotes are involved in ubiquitin-mediated protein recycling, vesicle formation and trafficking, endosomal sorting complexes required for transport (ESCRT)-mediated multivesicular body formation as well as cytokinetic abscission and cytoskeleton formation[18]. Asgard archaea have been suggested to possess heterotrophic lifestyles and are proposed to play a role in carbon degradation in sediments; however, several members of the Asgard archaea also have genes that code for a complete Wood-Ljungdahl pathway and are therefore interesting with regard to carbon cycling in sediments[14,19].

Here we present the first evidence of metagenome assembled genomes (MAGs), recovered from Guaymas Basin deep-sea hydrothermal sediments, which represent a new Asgard phylum with the metabolic potential to perform anaerobic hydrocarbon degradation using a methyl-CoM reductase-like homolog.

Results

Identification of Helarchaeota genomes from Guaymas Basin sediments. We recently obtained over ~280 gigabases of sequencing data from 11 samples taken from various sites and depths at Guaymas Basin hydrothermal vent sediments[20]. *De novo* assembly and binning of metagenomic contigs resulted in the reconstruction of over 550 genomes (>50% complete)[20]. Within these genomes we detected a surprising diversity of archaea, including >20 phyla, which appear to represent up to 50% of the total microbial community in some of these samples[20]. Preliminary phylogeny of the dataset using 37 concatenated ribosomal proteins revealed two draft genomic bins representing a new lineage in the Asgard archaea. These draft genomes, referred to as Hel_GB_A and Hel_GB_B, were re-assembled and re-binned resulting in final bins that were 82 and 87% complete and had a bin size of 3.54 and 3.84 Mbp, respectively (Table 1). An in-depth phylogenetic analysis consisting of 56 concatenated ribosomal proteins was used to confirm the placement of these final bins form a distant sister-group with the Lokiarchaeota (Figure 1a). Hel_GB_A percent abundance ranged from 3.41×10^{-3}% to 8.59 $\times 10^{-5}$% and relative abundance from 8.43 to 0.212. Hel_GB_B percent abundance ranged from 1.20×10^{-3}% to 7.99 $\times 10^{-5}$% and relative abundance from 3.41 to 0.22. For both Hel_GB_A and Hel_GB_B the highest abundance was seen at the site the genomes bins were recovered from. These numbers are comparable to other Asgard archaea isolated form these sites[20]. Hel_GB_A and Hel_GB_B had a mean GC content of 35.4% and 28%, respectively, and were recovered from two distinct environmental samples, which share similar methane-supersaturated and strongly reducing geochemical conditions (concentrations of methane ranging from 2.3-3 mM, dissolved inorganic carbon ranging from 10.2-16.6 mM, sulfate near 21 mM and sulfide near 2 mM; Supplementary Table 1) but differed in temperature (28^{o}C and 10^{o}C, respectively, Supplementary Table 1)[19].

Phylogenetic analyses of a 16S rRNA gene sequence (1058 bp in length) belonging to Hel_GB_A confirmed that they are related to Lokiarchaeota and Thorarchaeota, but are phylogenetically distinct from either of these lineages (Figure 1b). A comparison to published Asgard archaeal 16S rRNA gene sequences indicate a phylum level division between the Hel_GB_A sequence and other Asgard archaea[22] (Supplementary Table 2). A search for ESPs in both bins revealed that they contained a similar suite compared to those previously identified in Lokiarchaeota, which is consistent with their distant phylogenomic relationship (Figure 2). These lineages are relatively distantly related as evidenced by their difference in GC content and relatively low pairwise sequence identity of proteins. An analysis of the average amino acid identity (AAI) showed that Hel_GB_A and Hel_GB_B shared 1477 genes with and AAI of 51.96%. When compared to Lokiarcheota_CR4, Hel_GB_A share 634 out of orthologous genes 3595 and Hel_GB_B had 624 orthologous genes out of 3157. Helarchaeota bins showed the highest AAI similarity to Odinarchaeota LCB_4 (45.9%); however, it contained fewer orthologous genes (574 out of 3595 and 555 out of 3157 for Hel_GB_A and Hel_GB_B, respectively). Additionally, the Hel_GB bins differed from Lokiarchaeota in their total gene number, for example Hel_GB_A possessed 3595 genes and CR_4 possessed 4218; this difference is consistent with the larger estimated genome size for Lokiarchaeum CR_4 compared to Hel_GB_A (~5.2 Mbp to ~4.6 Mbp) (Supplementary Table 3, Supplementary Methods). These results add support to the phylum level distinction observed for Hel_GB_A and Hel_GB_B in both the ribosomal protein and 16s rRNA phylogenetic trees. We propose the name Helarchaeota after Hel, the Norse goddess of the underworld and Loki's daughter for this lineage.

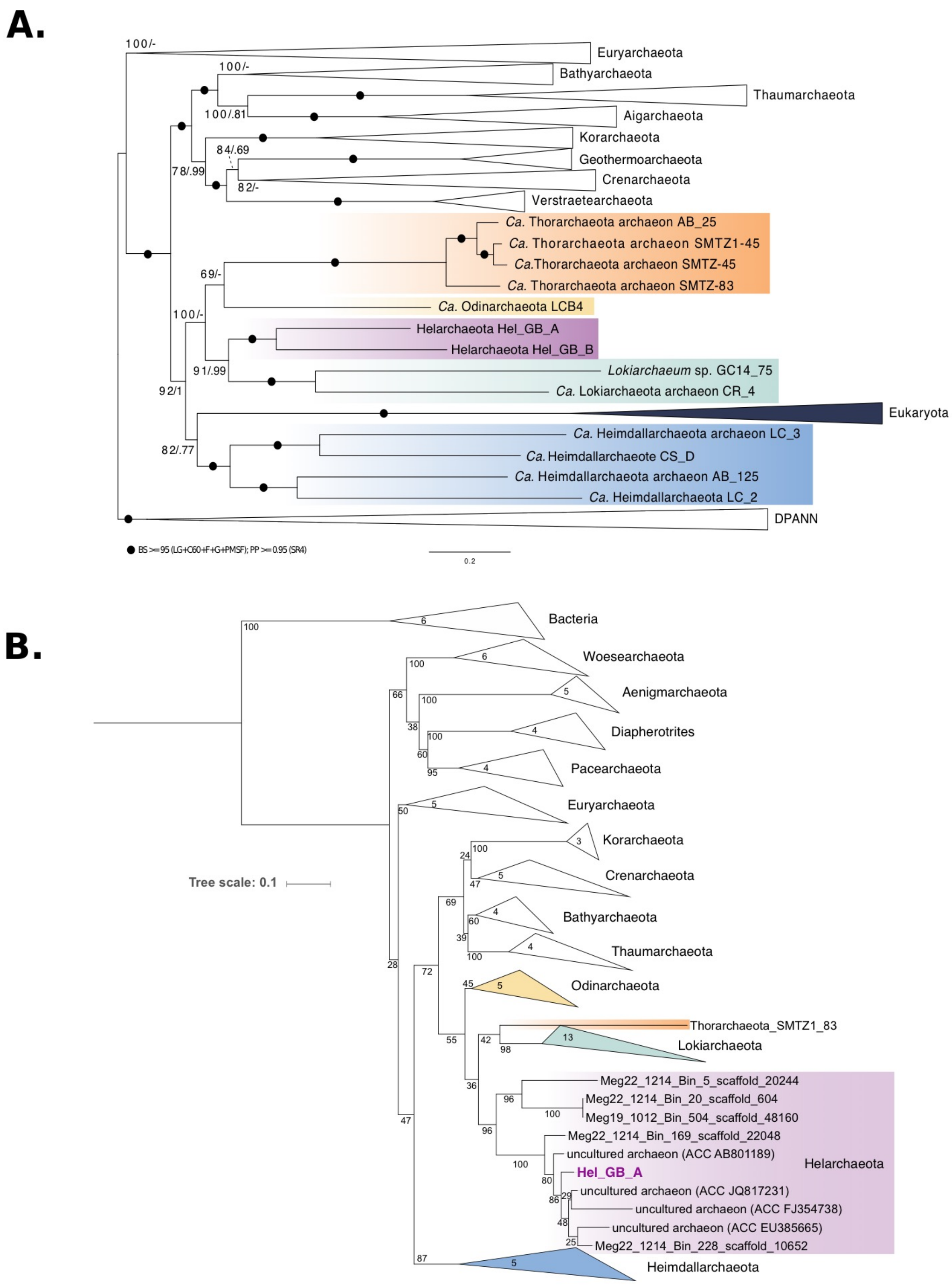

Figure 1. Phylogenomic position of Helarchaeota within the Asgard archaea superphylum (A) Phylogenomic analysis of 56 concatenated ribosomal proteins identified in Helarchaeota bins. Black circles indicate Bootstrap values greater than 95 (LG+C60+F+G+PMSF); Posterior Probability >= 0.95 (SR4). (B) Maximum-likelihood phylogenetic tree of 16S rRNA gene sequences thought to belong to Helarchaeota.

The phylogeny was generated using RAxML (GTRGAMMA model and number of bootstraps determined using the extended majority-rule consensus tree criterion). The purple box shows possible Helarchaeota sequences from GB data, as well as closely related published sequences and sequences form newly identified Helarchaeota bins (identified as Megxx_xxxx_Bin_xxx_scaffold_xxxxx). Number of sequences is depicted in the closed branches.

Metabolic analysis of Helarchaeota. To reconstruct the metabolic potential of these archaea, the Helarchaeota proteomes were compared to several functional protein databases[20] (Figure 3a). Like many archaea in marine sediments[23], Helarchaeota may be able to utilize organic carbon as they possess a variety of extracellular peptidases and carbohydrate degradation enzymes that include the β-glucosidase, α-L-arabinofuranosidase and putative rhamnosidase, among others (Supplementary Table 4 and 5). Degraded organic substrates can then be metabolized via glycolysis and an incomplete TCA cycle from citrate to malate and a partial gamma-aminobutyric acid shunt (Figure 3a, Supplementary Table 4). Both Helarchaeota bins are missing fructose-1,6-bisphosphatase and have few genes coding for the pentose phosphate pathway. Genes encoding for the bifunctional enzyme 3-hexulose-6-phosphate synthase/6-phospho-3-hexuloisomerase (hps-phi) were identified in Hel_GB_B suggesting they may be using the ribulose monophosphate (RuMP) pathway for formaldehyde anabolism. Genes coding for acetate-CoA ligase (both APM and ADP-forming) and an alcohol dehydrogenase (*adhE*) were identified in both genomes suggesting that the organisms may be capable of both fermentation and production of acetyl-CoA using acetate and alcohols (Supplementary Table 4). Like in Thorarchaeota and Lokiarchaeota, these genomes possess the large subunit of type IV Ribulose bisphosphate carboxylase[19,24]. Additionally, the Helarchaeota genomes encode for the catalytic subunit of the methanogenic type III ribulose bisphosphate carboxylase used for C-fixation[24]. Helarchaeota are metabolically distinct from Lokiarchaeota as both Hel_GB draft genomes appear to lack a complete TCA cycle as genes coding for citrate synthase and malate/lactate dehydrogenase are absent. Both genomes also likely produce acetyl-CoA using glyceraldehyde 3-phosphate dehydrogenase which is absent in Lokiarchaeota[19] (Supplementary Table 4). Helarchaeota genomes lack genes that code for enzymes involved in dissimilatory nitrogen and sulfur metabolism. Assimilatory genes including *sat, cysN and cysC* were found in Hel_GB_B however these genes were not identified in Hel_GB_A. This absence may be indicative of species-specific characteristics of their genomes or could be a results of genome incompleteness. Additional genomes of members of the Helarchaeota will help to fully understand the diversity of these pathways across the whole phylum.

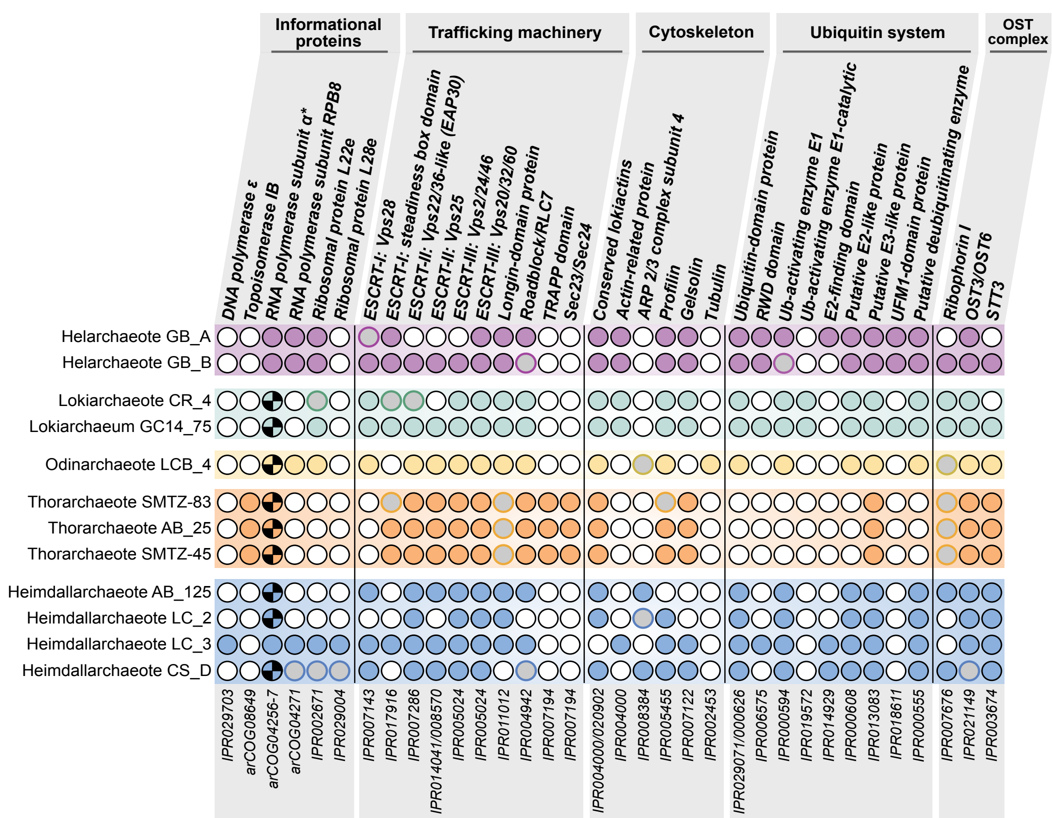

Figure 2. Distribution of eukaryotic signature proteins (ESPs) in Helarchaeota and other Asgard archaea. Numbers under each column correspond to the InterPro accession number (IPR) and Archaeal Clusters of Orthologous Genes (arcCOG) IDs that were searched for. Full circles refer to cases in which a homologue was found in the respective genomes. Empty circles with black outlines represent the absence of the ESP. The checkered pattern in the RNA polymerase subunit alpha represents the fact that the proteins were split, while the fused proteins are represented by the full circles. Grey circles with borders in any other color represent cases where the standard profiles were not found but potential homologs where detected. In the Roadblock proteins, potential homologs were detected but the phylogeny could not support the close relationship of any of these copies to the Asgard archaea group closest to eukaryotes. In the Ub-activating enzyme E1 represents homologs found clustered appropriately with its potential orthologs in the phylogeny but the synteny of this gene with other ubiquitin-related proteins in the genome is uncertain.

Interestingly, both Helarchaeota genomes have *mcrABG*-containing gene clusters encoding putative methyl-CoM reductase-like enzymes (Figure 3b, Supplementary Figure 2)[4,5,7]. Phylogenetic analyses of both the A subunit of methyl-CoM reductase-like enzymes (Supplementary Figure 2) as well as the concatenated A and B subunits (Figure 3b) revealed that the Helarchaeota sequences are distinct from those involved in methanogenesis and methane oxidation but cluster with homologs from butane oxidizing Syntrophoarchaea[7] and Bathyarchaeota with high statistical support (rapid bootstrap support/single branch test bootstrap support/posterior probability of 99.8/100/1; Figure 3b) excluding the distant homolog of *Ca.* Syntrophoarchaeum caldarius (OFV68676). Analysis of the Helarchaeota mcrA alignment confirmed that amino acids present at their active sites are similar to those identified on

Bathyarchaeota and Syntrophoarchaeum methyl-CoM reductase-like enzymes (Supplementary Figure 3). In Syntrophoarchaeum, the methyl-CoM reductase-like enzymes have been suggested to activate butane to butyl-CoM[7]. It is proposed that this process is then followed by the conversion of butyl-CoM to butyryl-CoA; however, the mechanism of this reaction is still unknown. Butyryl-CoA can then be oxidized to acetyl-CoA that can be further feed into the Wood-Ljungdahl pathway to produce CO_2[7]. While some n-butane is detected in Guaymas Basin sediments (usually below 10 micromolar), methane is the most abundant hydrocarbon (Supplementary Table 1) followed by ethane and propane (often reaching the 100 micromolar range); thus, a spectrum of short-chain alkanes could potentially be metabolized by Helarchaeota[26].

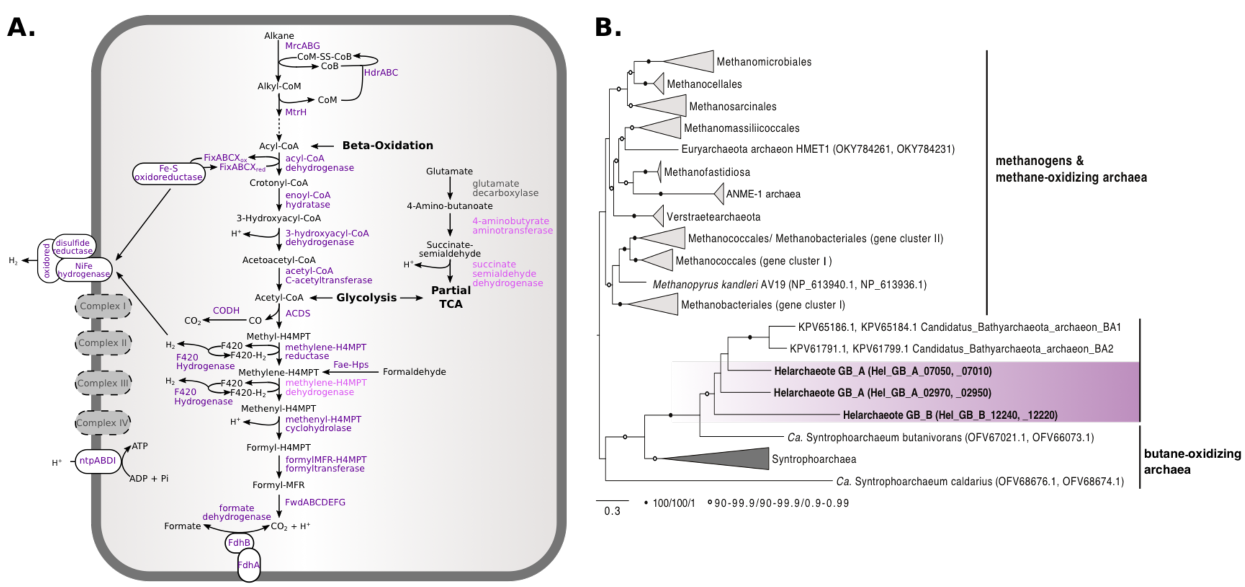

Figure 3. Metabolic inference of Helarchaeota and phylogenetic analyses of concatenated McrAB proteins. (A) Enzymes shown in dark purple are present in both genomes, those shown in light purple are present in a single genome and ones in grey are absent. (B) The tree was generated using IQ-tree with 1000 ultrafast bootstraps, single branch test bootstraps and posterior predictive values from the Bayesian phylogeny. White circles indicate bootstrap values of 90-99.9/90-99.9/0.9-0.99 and black filled circles

indicate values of 100/100/1. The tree was rooted arbitrarily between the cluster comprising canonical McrAB homologs and divergent McrAB homologs, respectively. Scale bars indicate the average number of substitutions per site.

Proposed hydrocarbon degradation pathway for Helarchaeota. Next, we searched for genes encoding enzymes potentially involved in hydrocarbon utilization pathways including propane and butane oxidation. Along with the methyl-CoM reductase-like enzyme that could convert alkane to alkyl-CoM, Helarchaeota possess heterodisulfide reductase subunits ABC (*hdrABC*) which is needed to recycle the CoM and CoB heterodisulfides after this reaction occurs (Figure 3 and 4)[7,8]. The conversion of alkyl-CoM to acyl-CoA is currently not understood in archaea capable of butane oxidation. Novel alkyl-binding versions of methyltransferases would be required to convert alkyl-CoM to butyl-CoA or other acyl-CoAs, as discussed for *Ca*. S. butanivorans[7]. Genes coding for methyltransferases were identified in both Helarchaeota genomes, including a likely tetrahydromethanopterin S-methyltransferase subunit H (MtrH) homolog (Figure 4; Supplementary Table 4). Short-chain acyl-CoA could be oxidized to acetyl-CoA using the beta-oxidation pathway via a short-chain acyl-CoA dehydrogenase, enoyl-CoA hydratase, 3-hydroxyacyl-CoA dehydrogenase and acetyl-CoA acetyltransferase, candidate enzymes for all of which are present in the Helarchaeota genomes and are also found in genomes of other Asgard archaea (Figure 4)[19]. Along with these enzymes, genes coding for the associated electron transfer systems, including an Fe-S oxidoreductase and all subunits of the electron transfer flavoprotein (ETF) complex were identified in Helarchaeota (Figure 4). Acetyl-CoA produced by beta-oxidation might be further oxidized to CO_2 via the Wood-Ljungdahl pathway, using among others the classical 5,10-methylene-tetrahydromethanopterin reductase (Figure 3a and 4).

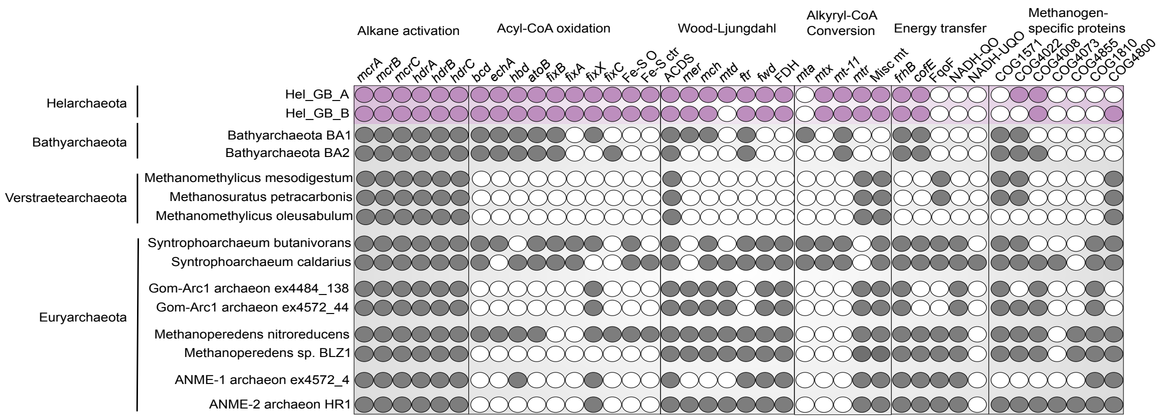

Figure 4. Comparison of Helarchaeota alkane metabolism to other alkane oxidizing and methanogenic archaea. Alkane metabolism of Helarchaeota compared to Bathyarchaeota and *Ca*. Syntrophoarchaeum sp., Verstraetearchaeota, GoM-Arc1 sp., ANME-1 sp. and ANME-2 sp. A list of genes and corresponding contig identifiers can be found in Supplementary Table 4.

Three possible energy-transferring mechanisms for Helarchaeota. To make anaerobic alkane oxidation energetically favorable, it must be coupled to the reduction of an internal electron acceptor or transferred to a syntrophic partner that can perform this reaction[7,26,27]. We could not identify an internal electron sink or any canonical terminal reductases used by ANME archaea (such as iron, sulfur or nitrogen), leading to the conclusion that a syntrophic partner organism would be necessary to enable growth on short-chain hydrocarbons. However, we could not identify any obvious syntrophic partner organisms based on co-occurrence analyses of abundance profiles of metagenomic datasets generated in this study[20].

An evaluation of traditional energy transferring mechanisms showed that our Helarchaeota bins lack genes coding for NADH:ubiquinone oxidoreductase, F_{420}-dependent oxidoreductase, $F_{420}H_2$:quinone oxidoreductase and NADH:quinone oxidoreductase that were identified in *Ca.* S. butanivorans (Figure 4)[7]. These electron-carrying proteins are important for energy transfer across the cell membrane and are common among syntrophic organisms[2,28,29]. Helarchaeota also lack genes coding for pili or cytochromes that are generally associated with electron transfer to a bacterial partner, as demonstrated for different ANME archaea[26,30]. Therefore, Helarchaeota may use a thus far unknown approach for energy conservation. Below we analyzed potential energy-transferring mechanisms that might be involved in syntrophic interactions between Helarchaeota and potential partner organisms.

A possible candidate for energy transfer to a partner may be formate dehydrogenase because substrate exchange in form of formate has previously been described to occur between methanogens and sulfur-reducing bacteria[27]. Helarchaeota genomes code for the alpha and beta subunits of a membrane-bound formate dehydrogenase (EC. 1.2.1.2) that could facilitate this transfer (Figure 2, Supplementary Table 4). However, to our knowledge formate transfer has not been shown to mediate methane oxidation. Alternatively, Helarchaeota may possess a novel redox-active complex. In both Helarchaeota bins, a gene cluster was found encoding three proteins that were identified as members of the HydB/Nqo4-like superfamily, Oxidored_q6 superfamily and a Fe-S disulfide reductase with a FlpD domain (mvhD) (Figure 5a). An analysis of these three proteins showed that each possessed transmembrane motifs (Figure 5b, and Supplementary Methods). While the membrane association of the disulfide reductase/FlpD needs to be confirmed, interactions with the other two membrane-associated subunits may allow for the bifurcated electrons to be transferred across the membrane.

Finally, hydrogen production and release was also considered as possible electron sink for Helarchaeota. We identified several hydrogenases and putative Fe-S disulfide reductase-encoding genes in the Helarchaeota genomes. Subsequent phylogenetic analyses revealed that the majority of these hydrogenases represent small and large subunits of group IIIC hydrogenases (methanogenic F_{420}-non-reducing hydrogenase (*mvh*)) that are usually involved in bifurcating electrons from hydrogen (Supplementary Figure 4, Supplementary Table 4). In contrast, while homologs belonging to the above mentioned Oxidored_q6 superfamily protein family are often found to be associated with group IV hydrogenases, canonical membrane-bound group IV-hydrogenases could not be identified in the genomes of the Helarchaeota. Altogether, this indicates that hydrogen could play a central role in energy metabolism of Helarcharota, but the absence of a classical membrane-bound hydrogenase makes it unlikely that hydrogen is the major syntrophic electron carrier.

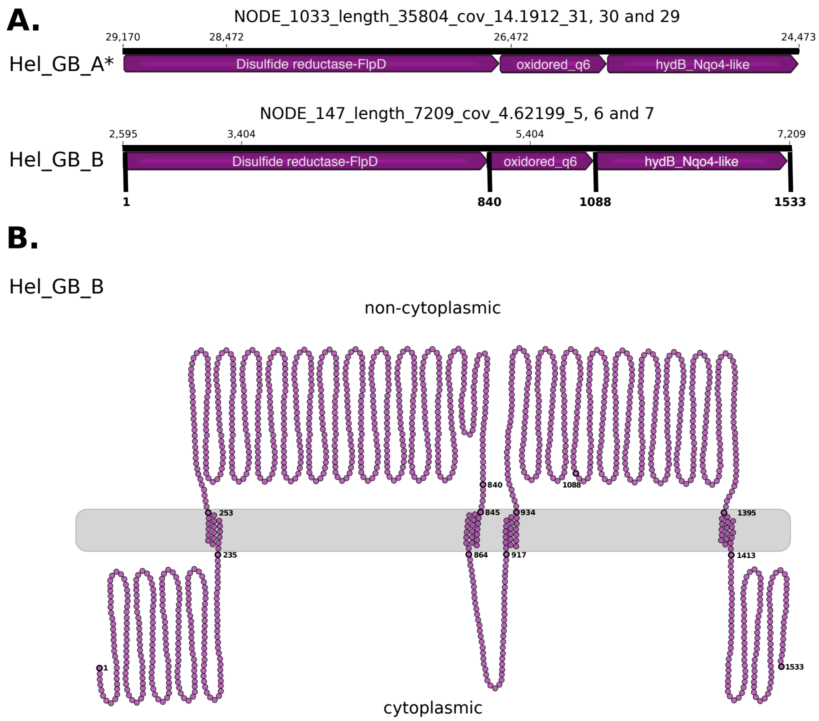

Figure 5. Depiction of a gene cluster found in both Helarchaeota genomes that consists of genes that encode for a possible energy-transferring complex. (A) In Hel_GB_A the complex was found on the reverse strand but has been oriented in the forward direction for clarity (asterisk). Arrows indicate the length of the reading frame. Gene names were predicted by various databases (Supplementary Methods). Small numbers located above the arrows refer to the nucleotide position for the full contig. Bold numbers on Hel_GB_B refer to the amino acid number of the whole complex. (B) Figure depicts the membrane motifs identified on NODE_147_length_7209_cov_4.62199_5, 6 and 7 using various programs (Supplementary methods). Each circle represents a single amino acid. Bold circles represent amino acids at the start of the protein, the start and end of the transmembrane sites, and the end of the complex. Numbering corresponds to the amino acid numbers of Hel_GB_B in panel (A). A full loop represents 50 amino acids and does not reflect the secondary structure of the complex.

Discussion

Historically methanogenesis and anaerobic methane oxidation were regarded as the only examples of anaerobic archaeal short-chain alkane metabolism. The enzymes acting in these pathways were considered to be biochemically and phylogenetically unique and limited to lineages within the Euryarchaeota[4]. This study represents the discovery of a novel phylum and the first indications for anaerobic short-chain alkane oxidation using a MCR-like homolog in the Asgard archaea. Since the presence of these *mcr* genes is restricted to Helarchaeota among the known Asgard archaea[19], these genes were likely transferred to Helarchaeota and do not constitute an ancestral trait within the Asgard superphylum. Based on current phylogenetic analysis, the Helarchaeota *mcr* gene cluster may have been horizontally acquired from either Bathyarchaeota or *Ca*. Syntrophoarchaeum (Fig. 1b, Supplementary Figure 3). Due to this close relationship, we based our analysis of Helarchaeota's ability to perform anaerobic short-chain hydrocarbon oxidation on the pathway proposed for Ca. Syntrophoarchaeum. Helarchaeota probably utilize a similar short-chain alkane as a substrate in lieu of methane, but given the low butane concentrations at our site it may not be an exclusive substrate.

Our comparison to *Ca.* S. butanivorans shows a consistent presence in genes necessary for this metabolism including a complete Wood-Ljungdahl pathway, acyl oxidation pathway and internal electron transferring systems. These electron-transferring systems are essential housekeeping components that act as electron carriers for oxidation reactions. Interestingly, in the Wood-Ljungdahl pathway identified in *Ca.* S. butanivorans, the bacterial enzyme is 5,10-methylene-tetrahydrofolate reductase (met) is thought to be substituting for the missing 5,10-methylene-tetrahydromethanopterin reductase (mer)[7]. In contrast, Helarchaeota encode the canonical archaeal-type mer. To render anaerobic butane oxidation energetically favorable, it must be coupled to the reduction of an electron acceptor such as nitrate, sulfate or iron[7,26,27]. In ANME archaea that lack genes for internal electron acceptors, methane oxidation is enabled through the transfer of electrons to a syntrophic partner organism. In Syntrophoarchaeum, syntrophic butane oxidation is thought to occur through the exchange of electrons via pili and/or cytochromes with sulfate-reducing bacteria[7]. Helarchaeota do not appear to encode any of the systems traditionally associated with syntrophy and no partner was identified in this study. Thus, further research is needed to identify possible bacterial partners.

Furthermore, the hypothesis for Helarchaeota growth through the anaerobic oxidation of short-chain alkanes remains to be confirmed as the genomes of members of this group do not encode canonical routes for electron transfer to a partner bacterium. However, we identified the genetic potential for potential enzymes that may be involved in transfer of electrons. Some methanogenic archaea use formate for syntrophic energy transfer to a syntrophic partner; therefore, the reverse reaction has been speculated to be energetically feasible for methane oxidation[27]. If this is true, the presence of a membrane-bound formate dehydrogenase in the Helarchaeota genomes may support this electron-transferring mechanism, however to our knowledge this has never been shown for an ANME archaea so far. Alternatively, the type 3 NiFe-hydrogenases encoded by Helarchaeota may be involved in transfer of hydrogen to a partner organism. For example, we identified a protein complex distantly related to the *mvh-hdr* of methanogens for electron transfer (Supplementary material). *Mvh-hdr* structures have been proposed to be potentially used by non-obligate hydrogenotrophic methanogens for energy transfer, but the directionality of hydrogen exchange could easily be reversed[2]. These methanogens form syntrophic associations with fermenting, H_2-producing bacteria, lack dedicated cytochromes or pili and use the *mvh-hdr* for electron bifurcation[2]. The detection of a hydrophobic region in the *mvh-hdr* complex led to the suggestion that this complex could be membrane bound and act as mechanism for electron transfer across the membrane; however, a transmembrane association has never been successfully shown[2]. While the membrane association of the disulfide reductase/FlpD needs to be confirmed, we were able to detect several other transmembrane motifs in the associated proteins that could potentially allow electron transfer in form of hydrogen to an external partner. Thus, while we propose that the most likely explanation for anaerobic short-chain alkane oxidation in Helarchaeota is via a syntrophic interaction with a partner, additional experiments are needed to confirm this working hypothesis.

The discovery of alkane-oxidizing pathways and possible syntrophic interactions in a new phylum of Asgard archaea indicates a much wider phylogenetic range for hydrocarbon utilization. Based on their phylogenetic distribution, the Helarchaeota *mcr* operon may have been horizontally transferred from either Bathyarchaeota or Syntrophoarchaeum. However, the preservation of a horizontally transferred pathway indicative of a competitive advantage; it follows that gene transfers among different archaeal phyla reflect alkane oxidation as a desirable metabolic trait. The discovery of the alkyl-CoM reductases and alkane-oxidizing pathways among the Asgard archaea indicates ecological roles for these still cryptic organisms, and opens up a wider perspective on the evolution and expansion of hydrocarbon-oxidizing pathways throughout the archaeal domain.

Methods

Sample collection and processing. Samples analyzed here are part of a study that aims to characterize the geochemical conditions and microbial community of Guaymas Basin (GB) hydrothermal vent sediments (Gulf of California, Mexico)[31,32]. The two genomic bins discussed in this paper, Hel_GB_A and Hel_GB_B, were obtained from sediment core samples collected in December 2009 on *Alvin* dives 4569_2 and 4571_4 respectively[21]. Immediately after the dive, freshly recovered sediment cores were separated into shallow (0-3 cm), intermediate (12-15 cm) and deep (21-24 cm) sections for further molecular and geochemical analysis, and frozen at -80^{o}C on the ship until shore-based DNA extraction. Hel_GB_A was recovered from the intermediate sediment (~28^{o}C) and Hel_GB_B was recovered from shallow sediment (~10^{o}C) from a nearby core (Supplementary Table 1); the sampling context and geochemical gradients of these hydrothermally influenced sediments are published and described in detail[21,31].

DNA was extracted from sediment samples using the MO BIO – PowerMax Soil DNA Isolation kit and sent to the Joint Genome Institute (JGI) for sequencing. A lane of Illumina reads (HiSeq–2500 1TB, read length of 2x151 bp) was generated for both samples. A total of 226,647,966 and 241,605,888 reads were generated for samples from dives for 4569-2 and 4571-4, respectively. Trimmed, screened, paired-end Illumina reads were assembled using the megahit assembler using a range of Kmers (See Supplementary Methods).

Genome reconstruction. The contigs from the JGI assembled data were binned using ESOM[33], MetaBAT[34] and CONCOCT[35] and resulting bins were combined using DAS Tool (version 1.0)[36] (See Supplementary Methods). CheckM lineage_wf (v1.0.5) was run on bins generated from DAS_Tool and 577 bins showed an completeness > 50% and were characterized further[37]. 37 Phylosift[38] identified marker genes were used for preliminary phylogenetic identification of individual bins (Supplementary Table 6). Thereby, we identified two genomes, belonging to a previously uncharacterized phylum within the Asgard archaea, which we named Helarchaeota. To improve the quality of these two Helarchaeota bins (increase the length of the DNA fragments and lower total number), we used Metaspades to reassemble the contigs in each individual bin producing scaffolds. Additionally, we tried to improve the overall assemblies by reassembling the trimmed, screened, paired-end Illumina reads provided by JGI using both IDBA-UD and Metaspades (Supplementary Methods). Binning procedures (using scaffolds longer than 2000 bp) as previously described in Supplementary Methods for the original bins were repeated with these new assembles. All bins were compared to the original Helarchaeota bins using blastn[39] for identification. Mmgenome[40] and CheckM[37] were used to calculate genome statistics (i.e. contig length, genome size, contamination and completeness). The highest quality Helarchaeota bin from each sample was chosen for further analyses. For the 4572-4 dataset, the best bin was generated using the Metaspades reassembly on the trimmed data and for the 4569-2 dataset the best bin was recovered using the Metaspades reassembly on the original Hel bin contigs. The final genomes were further cleaned by GC content, paired-end connections, sequence depth and coverage using Mmgenome[40]. CheckM was rerun on cleaned bins to estimate the Hel_GB_A to be 82% and Hel_GB_B to be 87% complete and both bins were characterized by a low degree of contamination (between 1.4-2.8% with no redundancy) (Table 1)[37]. Genome size was estimated to be 4.6 Mbp for Hel_GB_A and 4.1 for Hel_GB_B and was calculated using percent completeness and bin size to extrapolate the likely size of the complete genome. CompareM[41] was used to analysis differences between Helarchaeota bins and published Asgard bins using the command python comparem aai_wf --tmp_dir tmp/ --file_ext fa -c 8 aai_compair_loki aai_compair_loki_output.

16S rRNA gene analysis. Neither bin possessed a 16S rRNA gene sequence[38], and to uncover potentially unbinned 16S rRNA gene sequences from Helarchaeota, all 16S rRNA gene sequences obtained from samples 4569_2 and 4571_4 were identified using JGI-IMG annotations, regardless of whether or not the contig was successfully binned. These 16S rRNA gene sequences were compared using blastn[39] (blastn -

outfmt 6 -query Hel_possible_16s.fasta –db New_Hel_16s -out Hel_possible_16s_blast.txt -evalue 1E-20) to newly acquired 16S rRNA gene sequences from MAGs recovered from preliminary data from new GB sites. A 37 Phylosift[38] marker genes tree was used to assign taxonomy to these MAGs. We were able to identify five MAGs that possessed 16s and that formed a monophyletic group with our Hel_GB bins (Supplementary Table 2; Megxx in Figure 2). Of the unbinned 16S rRNA gene sequences one was identified as likely Helarchaeota sequence. The contig was retrieved from the 4572_4 assembly (designated Ga0180301_10078946) and was 2090 bp long and encoded for an 16S rRNA gene sequence that was 1058 bp long. Given the small size of this contig relative to the length of the 16S rRNA gene none of the other genes on the contig could be annotated. Blastn[39] comparison to published Asgard 16S rRNA gene sequences was performed using the following command: blastn -outfmt 6 -query Hel_possible_16s.fasta –db Asgrad_16s -out Hel_possible_16s_blast.txt -evalue 1E-20 (Supplementary Table 2). The GC content of each 16S rRNA gene sequence was calculated using the Geo-omics script length+GC.pl (https://github.com/Geo-omics/scripts/blob/master/AssemblyTools/length%2BGC.pl). For a further phylogenetic placement, the 16S rRNA gene sequences were aligned to the SILVA database (SINA v1.2.11) using the SILVA online server[42] and Geneious (v10.1.3)[43] was used to manually trim sequences. The alignment also contained 16S rRNA gene sequences from the new, preliminary Helarchaeota bins. The cleaned alignment was used to generated a maximum-likelihood tree with RAxML as follows: "/raxmlHPC-PTHREADS-AVX -T 20 -f a -m GTRGAMMA -N autoMRE -p 12345 -x 12345 -s Nucleotide_alignment.phy -n output" (Figure 1b).

Phylogenetic analysis of ribosomal proteins. For a more detailed phylogenetic placement, we used BLASTp[44] to identify orthologs of 56 ribosomal proteins in the two Helarchaeota bins, as well as from a selection of 130 representative taxa of archaeal diversity and 14 eukaryotes. The full list of marker genes selected for phylogenomic analyses is shown in Supplementary Table 7. Individual protein datasets were aligned using mafft-linsi[45] and ambiguously aligned positions were trimmed using BMGE (-m BLOSUM30)[46]. Maximum likelihood (ML) individual phylogenies were reconstructed using IQtree v. 1.5.5[47] under the LG+C20+G substitution model with 1000 ultrafast bootstraps that were manually inspected. Trimmed alignments were concatenated into a supermatrix, and two additional datasets were generated by removing eukaryotic and/or DPANN homologues to test the impact of taxon sampling on phylogenetic reconstruction. For each of these concatenated datasets, phylogenies were inferred using ML and Bayesian approaches. ML phylogenies were reconstructed using IQtree under the LG+C60+F+G+PMSF model[48]. Statistical support for branches was calculated using 100 bootstraps replicated under the same model. To test robustness of the phylogenies, the dataset was subjected to several treatments. For the 'full dataset' (i.e., with all 146 taxa), we tested the impact of removing the 25% fastest-evolving sites, as within a deep phylogenetic analysis, these sites are often saturated with multiple substitutions and, as a result of model-misspecification can manifest in an artifactual signal[50–52]. The corresponding ML tree was inferred as described above. Bayesian phylogenies were reconstructed with Phylobayes for the dataset "without DPANN" under the LG+GTR model. Four independent Markov chain Monte Carlo chains were run for ~38,000 generations. After a burn-in of 20%, convergence was achieved for three of the chains (maxdiff < 0.29). The initial supermatrix was also recoded into 4 categories, in order to ameliorate effects of model misspecification and saturation[52] and the corresponding phylogeny was reconstructed with Phylobayes, under the CAT+GTR model. Four independent Markov chain Monte Carlo chains were run for ~49,000 generations. After a burn-in of 20 convergence was achieved for all four the chains (maxdiff < 0.19). All phylogenetic analyses performed are summarized in Supplementary Table 8, including maxdiff values and statistical support for the placement of Helarchaeota, and of eukaryotes.

Phylogenetic analysis of McrA and concatenated McrA and McrB proteins. McrA homologs were aligned using mafft-linsi [45], trimmed with trimAL[53] and the final alignment consisting of 528 sites was subjected to phylogenetic analyses using v. 1.5.5 [47] with the LG+C60+R+F model. Support values were estimated using 1000 ultrafast boostraps[54] and SH-like approximate likelihood ratio test[55], respectively. Sequences for McrA and B were aligned separately with mafft-linsi [45] and trimmed using trimAL Subsequently, McrA and McrB encoded in the same gene cluster, were concatenated yielding a total alignment of 972 sites. Bayesian and Maximum likelihood phylogenies were inferred using IQtree v. 1.5.5 [47] with the mixture model LG+C60+R+F and PhyloBayes v. 3.2[56] using the CAT-GTR model. For Maximum likelihood inference, support values were estimated using 1000 ultrafast boostraps[54] and SH-like approximate likelihood ratio test[55], respectively. For Bayesian analyses, four chains were run in parallel, sampling every 50 points until convergence was reached (maximum difference < 0.07; mean difference < 0.002). The first 25% or the respective generations were selected as burn-in. Phylobayes posterior predictive values were mapped onto the IQtree using sumlabels from the DendroPy package[57]. The final trees were rooted artificially between the canonical Mcr and divergent Mcr-like proteins, respectively.

Metabolic Analyses. Gene prediction for the two Helarchaeota bins was performed using prodigal[58] (V2.6.2) with default settings and Prokka[59] (v1.12) with the extension '–kingdom archaea'. Results for both methods were comparable and yielded a total of 3,574-3,769 and 3,164-3,287 genes for Hel_GB_A and Hel_GB_B, respectively, with Prokka consistently identifying fewer genes. Genes were annotated by uploading the protein fasta files from both methods to KAAS (KEGG Automatic Annotation Server) for complete or draft genomes to assign orthologs[60]. Files were run using the following settings: prokaryotic option, GhostX and bi-directional best hit (BBH)[60]. Additionally, genes were annotated by JGI-IMG[61] to confirm hits using two independent databases. Hits of interest were confirmed using blastp on the NCBI webserver[44]. The dbCAN[62] and MEROPS[63] webserver were run using default conditions for identification of carbohydrate degrading enzymes and peptidases respectively. Hits with e-values lower than e^-20 were discarded. In addition to these methods an extended search was used to categorize genes involved in butane metabolism, syntrophy and energy transfer.

Identified genes predicted to code for putative alkane oxidation proteins were similar to those described from *Candidatus* Syntrophoarchaeum spp.. Therefore, a blastp[44] database consisting of proteins predicted to be involved in the alkane oxidation pathway of *Ca.* Syntrophoarchaeum was created in order to identify additional proteins in Helarchaeota, which may function in alkane oxidation. Positive hits were confirmed with blastp[44] on the NCBI webserver and compared to the annotations from JGI-IMG[61], Interpro[64], PROKKA[59] and KAAS[60] annotation. Genes for *mcrABG* were further confirmed by a HMMER[65] search to a published database using the designated threshold values[66] and multiple MCR trees (see Methods). To confirm that the contigs with the *mcrA* gene cluster were not missbined, all other genes on these contigs were analyzed for their phylogenetic placement and gene content. The prodigal protein predictions for genes on the contigs with *mcrA* operons were used to determine directionality and length of the potential operon.

To identify genes that are involved in electron and hydrogen transfer across the membrane, a database was created of known genes relevant in syntrophy that were download from NCBI. The protein sequences of the two Helarchaeota genomes were blasted against the database to detect relevant hits (E-value ≥ e ^-10). All hits were confirmed using the NCBI webserver, Interpro, JGI-IMG and KEGG. Hydrogenases were identified by a HMMER search to published database using the designated threshold values[67]. Hits were confirmed with comparisons against JGI annotations and NCBI blasts, the HydDB database[68] and a manual database made from published sequences[69,70]. All detected hydrogenases were used to generate two phylogenetic trees, one for proteins identified as small subunits and one for large subunits in order to properly identify the different hydrogenase subgroups. Hydrogenases that are part of the proposed complex were then further analyzed to evaluate if this was a possible operon by looking for possible transcription factors and binding motifs (Supplementary Methods).

ESP Identification. Gene prediction for the two Helarchaeota bins was performed using prodigal[58] (V2.6.2) with default settings. All the hypothetical proteins inferred in both Helarchaeaota were used as seeds against InterPro[64], arCOG[71] and nr using BLAST[44]. The annotation table from Zaremba-Niedzwiedzka, *et al*. 2017. was used as a basis for the comparison[12]. The IPRs (or in some cases, the arCOGs) listed in the Zaremba-Niedzwiedzka, et al. 2017 were searched for in the Helarchaeota genomes[12] and the resulting information was used to complete the presence/absence table. When something that had previously been detected in an Asgard bin was not found in a Helarchaeota bin using the InterPro/arCOG annotations, BLASTs were carried out using the closest Asgard seeds to verify the absence. In some cases, specific analyses were used to verify the homology or relevance of particular sequences. The details for each individual ESP are depicted in supplementary materials.

Data Availability. The raw reads from the metagenomes described in this study are available at JGI under the IMG genome ID 3300014911 and 3300013103 for samples 4569-2 and 4571-4, respectively. Genome sequences are available at NCBI under the accession numbers SAMN09406154 and SAMN09406174 for Hel_GB_A and Hel _GB_B respectively. Both are associated with BioProject PRJNA362212.

Tables

Table 1. Bin statistics for Helarchaeota Bins. Degree of completeness, contamination and heterogeneity was determined using CheckM[37].

SeqID	Hel_GB_A	Hel_GB_B
Completeness (%)	82.4	86.92
Contamination (%)	2.8	1.40
Strain heterogeneity (%)	0	0
Scaffold number	333	182
GC content (%)	35.40	28.00
N50 (bp)	15,161	28,908
Length total (Mbp)	3.84	3.54
Estimated Genome size (Mbp)	4.6	4.1
Longest contig (bp)	52,512	72,379
Mean contig (bp)	11,531	19,467

1. G E Claypool & Kvenvolden, and K. A. Methane and other Hydrocarbon Gases in Marine Sediment. *Annu. Rev. Earth Planet. Sci.* **11**, 299–327 (1983).
2. Thauer, R. K., Kaster, A.-K., Seedorf, H., Buckel, W. & Hedderich, R. Methanogenic archaea: ecologically relevant differences in energy conservation. *Nat. Rev. Microbiol.* **6**, 579–591 (2008).
3. Reeburgh, W. S. Oceanic Methane Biogeochemistry. *Chem. Rev.* **107**, 486–513 (2007).
4. Spang, A., Caceres, E. F. & Ettema, T. J. G. Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science* **357**, eaaf3883 (2017).
5. Evans, P. N. *et al.* Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science* **350**, 434–438 (2015).

6. Vanwonterghem, I. *et al.* Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. *Nat. Microbiol.* **1**, 16170 (2016).
7. Laso-Pérez, R. *et al.* Thermophilic archaea activate butane via alkyl-coenzyme M formation. *Nature* **539**, 396–401 (2016).
8. Dombrowski, N., Seitz, K. W., Teske, A. P. & Baker, B. J. Genomic insights into potential interdependencies in microbial hydrocarbon and nutrient cycling in hydrothermal sediments. *Microbiome* **5**, 106 (2017).
9. Bazylinski, D. A., Farrington, J. W. & Jannasch, H. W. Hydrocarbons in surface sediments from a Guaymas Basin hydrothermal vent site. *Org. Geochem.* **12**, 547–558 (1988).
10. Teske, A., Callaghan, A. V. & LaRowe, D. E. Biosphere frontiers of subsurface life in the sedimented hydrothermal system of Guaymas Basin. *Front. Microbiol.* **5**, (2014).
11. Von Damm, K. L., Edmond, J. M., Measures, C. I. & Grant, B. Chemistry of submarine hydrothermal solutions at Guaymas Basin, Gulf of California. *Geochim. Cosmochim. Acta* **49**, 2221–2237 (1985).
12. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
13. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).
14. Seitz, K. W., Lazar, C. S., Hinrichs, K.-U., Teske, A. P. & Baker, B. J. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *ISME J* **10**, 1696–1705 (2016).

15. Jørgensen, S. L., Thorseth, I. H., Pedersen, R. B., Baumberger, T. & Schleper, C. Quantitative and phylogenetic study of the Deep Sea Archaeal Group in sediments of the Arctic mid-ocean spreading ridge. *Front. Microbiol.* **4**, (2013).

16. Jorgensen, S. L. *et al.* Correlating microbial community profiles with geochemical data in highly stratified sediments from the Arctic Mid-Ocean Ridge. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E2846-2855 (2012).

17. Hartman, H. & Fedorov, A. The origin of the eukaryotic cell: a genomic investigation. *Proc. Natl. Acad. Sci.* **99**, 1420–1425 (2002).

18. Eme, L., Spang, A., Lombard, J., Stairs, C. & J. G. Ettema, T. *Archaea and the origin of eukaryotes*. **15**, (2017).

19. Spang, A. *et al.* A renewed syntrophy hypothesis for the origin of the eukaryotic cell based on comparative analysis of Asgard archaeal metabolism. *Nat. Microbiol.* **Submitted**,

20. Dombrowski, N., Teske, A. P. & Baker, B. J. Extensive metabolic versatility and redundancy in microbially diverse, dynamic Guaymas Basin hydrothermal sediments. *Nat. Commun.* **9:4999**, (2018).

21. McKay, L. *et al.* Thermal and geochemical influences on microbial biogeography in the hydrothermal sediments of Guaymas Basin, Gulf of California. *Environ. Microbiol. Rep.* **8**, 150–161 (2016).

22. Yarza, P. *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12**, 635–645 (2014).

23. Lazar, C. S. *et al.* Environmental controls on intragroup diversity of the uncultured benthic archaea of the miscellaneous Crenarchaeotal group lineage naturally enriched in

anoxic sediments of the White Oak River estuary (North Carolina, USA). *Environ. Microbiol.* **17**, 2228–2238 (2015).

24. Tabita, F. R., Satagopan, S., Hanson, T. E., Kreel, N. E. & Scott, S. S. Distinct form I, II, III, and IV Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and structure/function relationships. *J. Exp. Bot.* **59**, 1515–1524 (2007).

25. Dowell, F. *et al.* Microbial Communities in Methane- and Short Chain Alkane-Rich Hydrothermal Sediments of Guaymas Basin. *Front. Microbiol.* **7**, (2016).

26. Krukenberg, V. *et al.* Candidatus Desulfofervidus auxilii, a hydrogenotrophic sulfate-reducing bacterium involved in the thermophilic anaerobic oxidation of methane. *Environ. Microbiol.* **18**, 3073–3091 (2016).

27. Stams, A. J. M. & Plugge, C. M. Electron transfer in syntrophic communities of anaerobic bacteria and archaea. *Nat. Rev. Microbiol.* **7**, 568–577 (2009).

28. Meuer, J., Kuettner, H. C., Zhang, J. K., Hedderich, R. & Metcalf, W. W. Genetic analysis of the archaeon Methanosarcina barkeri Fusaro reveals a central role for Ech hydrogenase and ferredoxin in methanogenesis and carbon fixation. *Proc. Natl. Acad. Sci.* **99**, 5632–5637 (2002).

29. Kunow, J., Linder, D., Stetter, K. O. & Thauer, R. K. F420H2: quinone oxidoreductase from Archaeoglobus fulgidus. *Eur. J. Biochem.* **223**, 503–511 (1994).

30. Wegener, G., Krukenberg, V., Riedel, D., Tegetmeyer, H. E. & Boetius, A. Intercellular wiring enables electron transfer between methanotrophic archaea and bacteria. *Nature* **526**, 587–590 (2015).

31. McKay, L. J. *et al.* Spatial heterogeneity and underlying geochemistry of phylogenetically diverse orange and white Beggiatoa mats in Guaymas Basin hydrothermal sediments. *Deep Sea Res. Part Oceanogr. Res. Pap.* **67**, 21–31 (2012).

32. Meyer, S. *et al.* Microbial habitat connectivity across spatial scales and hydrothermal temperature gradients at Guaymas Basin. *Front. Microbiol.* **4**, (2013).

33. Dick, G. J. *et al.* Community-wide analysis of microbial genome sequence signatures. *Genome Biol.* **10**, R85 (2009).

34. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).

35. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, nmeth.3103 (2014).

36. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation, and scoring strategy. *bioRxiv* 107789 (2017). doi:10.1101/107789

37. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* gr.186072.114 (2015). doi:10.1101/gr.186072.114

38. Darling, A. E. *et al.* PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**, e243 (2014).

39. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

40. Karst, S. M., Kirkegaard, R. H. & Albertsen, M. mmgenome: a toolbox for reproducible genome extraction from metagenomes. *bioRxiv* 059121 (2016). doi:10.1101/059121

41. https://github.com/dparks1134/CompareM.

42. Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–1829 (2012).

43. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinforma. Oxf. Engl.* **28**, 1647–1649 (2012).

44. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

45. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

46. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).

47. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. Iq-tree: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

48. Wang, H.-C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst. Biol.* syx068 (2017).

49. Jeffroy, O., Brinkmann, H., Delsuc, F. & Philippe, H. Phylogenomics: the beginning of incongruence? *Trends Genet.* **22**, 225–231 (2006).

50. Lartillot, N. & Philippe, H. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **363**, 1463–72 (2008).

51. Brown, M. W. M. *et al.* Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. *Proc. R. Soc. B Biol. Sci.* **280**, 20131755 (2013).

52. Susko, E. & Roger, A. J. On reduced amino acid alphabets for phylogenetic inference. *Mol. Biol. Evol.* **24**, 2139–2150 (2007).

53. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinforma. Oxf. Engl.* **25**, 1972–1973 (2009).

54. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).

55. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).

56. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).

57. Sukumaran, J. & Holder, M. T. DendroPy: a Python library for phylogenetic computing. *Bioinforma. Oxf. Engl.* **26**, 1569–1571 (2010).

58. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

59. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinforma. Oxf. Engl.* **30**, 2068–2069 (2014).

60. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–W185 (2007).

61. Markowitz, V. M. *et al.* IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* **40**, D115–D122 (2012).

62. Yin, Y. *et al.* dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **40**, W445–W451 (2012).

63. Rawlings, N. D., Barrett, A. J. & Bateman, A. MEROPS: the peptidase database. *Nucleic Acids Res.* **38**, D227–D233 (2010).

64. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinforma. Oxf. Engl.* **30**, 1236–1240 (2014).

65. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).

66. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 13219 (2016).

67. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. **7**, (2016).

68. Søndergaard, D., Pedersen, C. N. S. & Greening, C. HydDB: A web tool for hydrogenase classification and analysis. *Sci. Rep.* **6**, 34212 (2016).

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

70. Vignais, P. M., Billoud, B. & Meyer, J. Classification and phylogeny of hydrogenases1. *FEMS Microbiol. Rev.* **25**, 455–501

71. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales. *Life* **5**, 818–840 (2015).

Acknowledgements

This study was supported in part by an Alfred P. Sloan Foundation fellowship (FG-2016-6301) and National Science Foundation Directorate of Biological Sciences (award 1737298) to BJB. Sampling in Guaymas Basin and post-cruise work was supported by NSF Awards OCE-0647633 and OCE-1357238 to APT, respectively. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 provided to ND.

Author contributions

KWS, TJGE, ND and BJB conceived the study. KWS, ND, and BJB analyzed the genomic data. APT collected and processed samples. KWS, AS, and LE performed phylogenetic analyses. JL analyzed ESPs. KWS, AS, JRS, APT, BJB handled the metabolic inferences. BJB and KWS wrote the manuscript with inputs from all authors.