Microbes in Guaymas Basin (Gulf of California) hydrothermal sediments thrive on hydrocarbons and sulfur and experience steep, fluctuating temperature and chemical gradients. The functional capacities of communities inhabiting this dynamic habitat are largely unknown. Here, we reconstructed 551 genomes from hydrothermally influenced, and nearby cold sediments belonging to 56 phyla (40 uncultured). These genomes comprise 22 unique lineages, including five new candidate phyla. In contrast to findings from cold hydrocarbon seeps, hydrothermal-associated communities are more diverse and archaea dominate over bacteria. Genome-based metabolic inferences provide first insights into the ecological niches of these uncultured microbes, including methane cycling in new Crenarchaeota and alkane utilization in ANME-1. These communities are shaped by a high biodiversity, partitioning among nitrogen and sulfur pathways and redundancy in core carbon-processing pathways. The dynamic sediments select for distinctive microbial communities that stand out by expansive biodiversity, and open up new physiological perspectives into hydrothermal ecosystem function.
Microbial communities inhabit every environment and are comprised of a multitude of different phyla, the majority of which are uncultured. Among these environments, marine sediments contain abundant and phylogenetically diverse microbial communities. High diversity has been suggested to emerge as a strategy for survival of microbes under fluctuating environmental conditions. While single-gene surveys allow us to address the phylogenetic diversity of microbial communities, metagenomic analyses provide a connection between biodiversity and the functional potential encoded within sedimentary communities.

Guaymas Basin (GB; Gulf of California, Mexico) is a young, active seafloor-spreading center characterized by high water column productivity and fast sedimentation rates, leading to the accumulation of massive layers of organic-rich sediments that cover the hydrothermal spreading center and ridge flanks. The emplacement of hot basalt sills into organic-rich sediment transforms buried organic matter into CO₂, H₂S, low-molecular-weight organic acids, ammonia, and hydrocarbons such as methane, ethane and benzene. These compounds migrate to the sediment surface with rising vent fluids, where they fuel hydrocarbon-degrading microbial communities. Among all hydrothermally generated hydrocarbons, methane has received considerable interest as greenhouse gas shaping global climate.

Porewater methane reaches millimolar concentrations while ethane ranges from 40-100 µM. Also present in these sediments are propane, n-butane and pentane, which accumulate at lower concentrations compared to methane. Altogether, hydrocarbons represent lucrative carbon sources for the resident microbial community. Additionally, hydrothermal circulation and seawater in-mixing provide the upper sediments with electron acceptors, among which sulfate is widely available in millimolar porewater concentrations and rarely depleted within hydrothermal sediment cores. In-situ microelectrode surveys detect small oxygen peaks within hydrothermal sediments near the mat-covered surface. These results are consistent with short-term dynamics of hydrothermal flow within minutes and hours. Additionally, short-term dynamics overlay with longer-term hydrothermal activity changes over months and years.

GB sediments have been shown to host diverse microbial communities with distinct roles in carbon cycling. In particular, microbial consortia perform the anaerobic oxidation of methane (AOM) in a syntrophic interaction consisting of anaerobic methane-oxidizing archaea (ANME) and bacterial sulfate reducers, typically Deltaproteobacteria, but including other thermophilic bacterial lineages, such as Candidatus Desulfurovibrio auxilii. Anaerobic hydrocarbon degraders include Ca. Syntrophoarchaeum, which oxidizes butane in a syntrophic interaction with Ca. Desulfurovibrio auxilli, or the butane- and propane oxidizing isolate BuSS, belonging to the Desulfosarcina-Desulfococcus cluster. Other common archaeal lineages include Marine Benthic Group D and Bathychaetota, while bacterial phyla include Proteobacteria (Delt-, Epsilon- and Gammaproteobacteria), Bacteroidetes and Chloroflexi as well as several candidate phyla. Within the GB hydrothermal area in the southern spreading center, a high degree of microbial community connectivity exists among hydrothermal vent sites and sediments within a few hundred meters. A core microbiome is shared between microbial communities of GB hydrothermal sediments, as well as with hot, oil-rich sediments. These results are consistent with the hypothesis that microbial assemblages from hydrothermal sediments are phylogenetically distinct from those in the surrounding region and host a greater metabolic diversity.

Results
Phylogenetic diversity in Guaymas Basin sediments. To examine the biodiversity of microbial communities inhabiting GB sediments, we sampled and sequenced eleven sediments covering different sampling locations, depths (0–24 cm), temperatures (3–60 °C) and geochemical regimes. We selected samples, represented by core 4567_28, that are not influenced by hydrothermal activity (temperature ~3 °C) and occur interlaced with hydrothermal hot spots within the spreading center. All other samples are characterized by steep thermal gradients, reflected by in-situ temperatures ranging from 4 °C to 60 °C. Dense mats of filamentous Gammaproteobacteria (family Beggiatoaceae) covered hydrothermal sediments from dive 4569, with an orange mat dominating core 4569_9 and a white mat at the adjacent core 4569_2. Core 4569_4 was collected from the periphery of this hydrothermal hotspot and did not contain visible mats (Fig. 1). Porewater methane, sulfate, dissolved inorganic carbon (DIC) and sulfide co-occurred throughout these cores (Supplementary Information), consistent with hydrothermal circulation and in-mixing of seawater-derived electron acceptors. Cores 4571_4 and 4488_9 represent hot and oily sediments with yellow-white sulfur precipitates on the surface (Fig. 1). Among the hydrothermal cores, 4488_9 stands out by steep thermal gradients (~150 °C at 30 cm depth), high sample temperature (~60 °C), sulfate depletion at shallow depths, and accumulation of non-methane hydrocarbons (Supplementary Figure 1, Supplementary Data 1). After sequence assembly, we reconstructed 551 draft genomes via tetranucleotide and coverage binning. These metagenome-assembled genomes (MAGs), simplified as ‘genomes’ throughout the manuscript, represent medium-quality MAGs and were >50% complete and <10% contaminated (301 genomes >70% and 61 genomes >90% complete; Supplementary Data 2, 3).

Each genome was classified by constructing a phylogenetic tree using 37 single-copy, protein-coding marker genes (Supplementary Data 4). Overall, the 551 genomes (247 archaea and 304 bacteria) represented 16 cultured and 40 uncultured, candidate phyla that comprise a substantial number of new microbial lineages, many of which branch basal to those previously described (Fig. 2, Supplementary Figure 2). These lineages are phylogenetically distinct from those in the southern spreading center and host a greater metabolic diversity. Therefore, we sequenced a total of ~4 billion genomic reads from eleven samples (two of which were from cool, background sediments) from GB. Altogether, these data add 22 branches to the tree of life and enabled us to determine the genetic repertoire and metabolic versatility of these extreme hydrothermal communities.
GB-BP1-3 for archaeal and bacterial phyla, respectively. The placement of these five phyla was confirmed by comparing the average amino acid identity (AAI) of genomes within a phylum to genomes of all other phyla (Supplementary Data 6). Within each new phylum, GB-AP1,2 shared an AAI of ~44 and ~96% and GB-BP1-3 of ~54, ~72 and ~60%, respectively, and were more similar to themselves than to other genomes (~43% AAI summarized across all genomes). While the genomes of GB-AP1 shared a low AAI, we did not detect any lineage with a closer similarity. Two 16S rRNA gene sequences recovered from GB-BP1 clustered with the uncultivated lineage MAT-CR-M4-B07\(^\text{30}\), which was previously detected in the Kazan mud volcano or Guerrero Negro hypersaline mats (Supplementary Figure 3). In total, we defined 24 archaeal and 37 bacterial groups (or ‘clusters’) for closer analysis (see Methods section, Supplementary Data 3 and Fig. 2). Archaeal genomes were represented by Bathyarchaeota (\(n = 41\)), Thermoproteales (\(n = 40\)) and Thermoplasmata (\(n = 36\)), and bacteria belonged to Deltaproteobacteria (\(n = 39\)), Gammaproteobacteria (\(n = 39\)), and Bacteroidetes (\(n = 27\)) (Supplementary Data 3). Additionally, we detected several candidate lineages, including Asgard archaea (\(n = 9\)), Verstraetearchaeota (\(n = 7\)), the bacterial CPR superphylum (\(n = 6\)). Overall, more genomes were recovered from hydrothermal (average of ~60 genomes per sample) than from background sediments (average ~9 genomes per sample; Supplementary Data 3). We detected only one archaeal (Bathyarchaeota) and 7 bacterial lineages (Chloroflexi, Deltaproteobacteria, Gammaproteobacteria) in the background compared to 22 archaeal and 31 bacterial clusters in the hydrothermal samples, suggesting a greater biodiversity in the more extreme environment.

**The effect of environmental parameters on community assembly.** To better understand the factors that drive community assembly, we investigated the occurrence of major phylogenetic clusters across sites. First, we confirmed that the genomes accurately reflected the community as a whole based on the abundance of ribosomal protein S3 across sites (Supplementary Figure 4). Next, we used the genomes to estimate the occurrence of different phylogenetic groups across all samples (Supplementary Figure 5, Supplementary Data 7). Several bacterial lineages, such as Planctomycetes or Deltaproteobacteria, were more frequently detected in background sediments than in hydrothermal sediments. In contrast, archaea were increasingly detected within the deeper, hotter hydrothermal samples, but not in cool surface sediments on the periphery of hydrothermal hot spots. Dominant lineages in the hot samples were Thaumarchaeota and Archaeoglobales as well as Acetotherm, and Omnitrophica. Two genotypes dominated hot sediments: B48_G6 (Methanosarcinales, ANME-1) and B16_G6 (Thermodesulfobacteria, ~88% AAI to Ca. Desulfofervidus auxilia) (Supplementary Data 3, Supplementary Data 7). While the hydrothermal sediments had an overall similar distribution of taxa across depth profiles, the oily sediment from 4488_9 harbored only few abundant taxa, including Thermoplasmata, Aerophobetes and Thermotoga (Supplementary Figures 4, 5). Core 4488_9 differs from other hydrothermal samples in its high hydrocarbon content, quick downcore depletion of sulfate, and steep thermal gradients (Supplementary Data 1, Supplementary Methods). In combination these factors appear to reduce the microbial diversity, especially of the archaeal community. Altogether, the hydrothermal activity gives rise to a unique community that shows a marked enrichment in archaea that can represent up to 50% of recovered genomes (Supplementary Data 7). This enrichment appears to be largely driven by the rich substrate availability, by hydrothermal circulation and by inmixing of the electron acceptor sulfate (Supplementary Methods). However, a greater sampling size would be needed to disentangle the relative contribution of individual factors on community assembly such as temperature, methane or hydrocarbon availability.

**Carbon cycling.** Given that these genomes yielded such a large number of unique microbial lineages, we inferred their potential physiological capabilities by assigning metabolic functions to proteins in each individual genome. First, we investigated the ability of the community to degrade and metabolize complex carbohydrates and peptides deposited in sediments by searching genomes for the presence of carbohydrate-active enzymes.
(CAZYmes), peptidases and pathways for carbon metabolism. In total, we detected ~30,000 and ~11,000 potential CAZYmes and peptidases, respectively (Fig. 3, Supplementary Figure 6, Supplementary Data 8, 9). Generally, bacteria encoded for a broader repertoire of CAZYmes compared to archaea; for example GH13 (α-amylase), GH23 (lytic transglycosylase) or GH74 (xyloglucanase) were more common in bacteria (Fig. 3, Supplementary Data 8). Most CAZymes were assigned to Thermoproteales (n = 20) and Asgard archaea (n = 16) as well as Verrucomicrobia (n = 38) and Bacteroidetes (n = 30). Peptidases were more equally distributed across both domains and abundant in Asgard archaea (n = 34) and Thermococci (n = 24) as well as Aminicenantes (n = 54) and Acidobacteria (n = 51; Supplementary Figure 6, Supplementary Data 9). Approximately 2–3% of CAZYmes and peptidases are potentially secreted, suggesting that complex substrates are degraded outside of the cell and later taken up for degradation. Potentially secreted enzymes include CE8 (pectin methylesterase), and GH13 (α-amylase) as well as M28 (amino-peptidases) and S08 (subtilisin-like peptidases). A subset of CAZymes, such as GH23, may be involved in cell wall maintenance; however, the presence of sugar and peptide transporters as well as downstream metabolic pathways in most genomes suggest that other CAZymes might be involved in energy metabolism (see below).

Common pathways for the degradation of substrates produced by the activity of CAZymes and peptidases include glycolysis (glucokinase (glk), phosphofructokinase (pfk), pyruvate kinase (pyk)), gluconeogenesis (fructose-1,6-bisphosphatase (fbp), phosphoenolpyruvate carboxykinase (pckA)) and fermentation (Fig. 4 and Supplementary Data 10). In several cases, archaeal genomes encoded for more key genes of gluconeogenesis compared to glycolysis, which could imply that some archaea prefer peptides as an energy source; this finding is consistent with the occurrence of a high number of peptidases in their genomes (Supplementary Figure 6). Compared to archaea, bacteria contained a greater metabolic repertoire and might use both glycolysis and gluconeogenesis. Most genomes encoded for the potential to metabolize pyruvate produced during glycolysis to acetyl-CoA and further into fermentation pathways, producing formate, ethanol or acetate (Fig. 4). GB archaea were mainly capable of acetate formation using the ADP-forming acetyl-CoA synthetase (acdA), while bacteria encoded for phosphate acetyltransferase (pta) and acetate kinase (ackA) for acetate production; formate C-acetyltransferase (pyfD) and formate dehydrogenase (fdhG) for

![Fig. 2 Maximum likelihood phylogenetic tree of GB genomes based on 37 concatenated protein-coding genes. Grey: Reference Genomes. Blue: Genomes assembled from cold background sediments. Red: Genomes recovered from hot, hydrothermal sediments. The full tree can be found in Supplementary Figure 2 and the tree file is available in Supplementary Data 5](image-url)
formate production; and aldehyde dehydrogenase (aldh) and alcohol dehydrogenase (adhl) for ethanol production.

Not only is the GB microbiome able to process the deposited organic carbon pool by fermentation, but we also detected pathways for carbon fixation. The most common route of carbon fixation was the Wood-Ljungdahl pathway in both archaea and bacteria, while the Calvin-Benson-Bassham (CBB) and rTCA cycles were restricted mostly to Proteobacteria (Fig. 4, Supplementary Data 10). Although the Group III Ribulose-1,5-bisphosphate carboxylase-oxygenase (Rubisco, key marker gene of the CBB cycle) was detected in most archaea, this subgroup is implied in a nucleotide salvage pathway and not necessarily used for carbon fixation (Supplementary Data 10)\(^3\). A Group I/II Rubisco, feeding CO\(_2\) into the CBB cycle, was only detected in some Gammaproteobacteria (orders Chromatiales and Thiocarbionales). Additionally, marker genes for the rTCA cycle, including ATP-citrate-lyase (adlAB), pyruvate ferredoxin oxidoreductase (porABCD) and 2-oxoacid ferredoxin oxidoreductase (oorABCD), were mainly detected in Epsilonproteobacteria (order Campylobacterales; Supplementary Data 10). While several of the 3-hydroxypropionate or related cycles were present in a subset of genomes, a full pathway appeared to be absent (Supplementary Data 10). Conversely, the Wood-Ljungdahl pathway was present in several clusters, including Archaeoglobales and Methanosarcinales as well as Chloroflexi and Deltaproteobacteria (Fig. 4, Supplementary Data 10). Interestingly, we also detected genes from this pathway in candidate phyla, including Hydrothermarchaeota and Latescibacteria, which might either oxidize acetate or perform acetogenesis.

**Alkyl-coenzyme M reductase linked hydrocarbon cycling.** We detected the methyl-Coenzyme M reductase (mcra), a key enzyme for methanogenesis and AOM, in Syntrophoarchae, Methanomicrobia, and a deep-branching Thermoproteales lineage (designated DeepCrenGroup1; Fig. 4, Supplementary Data 3). To our knowledge this is the first report of mcrABG genes in the Crenarchaeota. The only bacteria able to utilize methane encoded for the particulate methane monooxygenase (pmoA), which was restricted to Gammaproteobacteria (orders Cellvibrionales and Methyllococcales; Supplementary Data 10). A closer phylogenetic analysis of McrA might even suggest a broader substrate usage potentially not restricted to methane (Fig. 5, Supplementary Data 11). McrA from most ANME-1, ANME-2c and DeepCrenGroup1 clustered with known methane oxidizers, while the McrA from one Syntrophoarchaeum (B49_G1) clustered with butane-oxidizers (Fig. 5). McrA from GoM-Arc1 branched between those two clusters, which is consistent with earlier work that suggested that GoM-Arc1 might utilize a different alkane, perhaps ethane, which can reach relatively high concentrations of 40-100 \(\mu\)M in GB\(^{11,20}\). However, further experimental evidence, preferably from enrichment cultures, is needed to confirm the substrate usage of these McrA proteins.

Surprisingly, ANME-1 bin B39_G2 contains two McrA proteins (on two different contigs, both mate-paired to other contigs from that bin) that are phylogenetically related to those from Ca. Syntrophoarchaeum spp. (Fig. 5). Similarly to Ca. Syntrophoarchaeum spp. B39_G2 contains genes with homology to those that encode for the butyryl-CoA oxidation pathway, such as acyl-CoA dehydrogenase and enoyl-CoA dehydratase (Supplementary Data 10, Supplementary Figure 7). This pathway appears to be involved in butane oxidation in Ca. Syntrophoarchaeum butanivorans\(^{16}\), making this the first example of an ANME-1 archaean potentially able to use short-chain alkanes. The detection of these unique methyl coenzyme-M reductase genes and pathways suggests that ANME-1 archaean are not limited to methane utilization and potentially able to oxidize alkanes anaerobically.

**Lipid and hydrocarbon utilization.** Pathways for lipid degradation were widespread in bacteria and less common in archaea, where they were mainly detected in Archaeoglobales, Bathyarchaeota and Geothermarchaeota (Fig. 4, Supplementary Data 10, Supplementary Figures 1 and 8). Additional CAZyme markers linked to lipid utilization were identified in many phylogenetic clusters (Fig. 3).}

---

**Fig. 3** Number of carbohydrate-active enzymes (CAZymes) encoded in GB genomes. Percentage of carbohydrate esterases (CE), glycoside hydrolases (GH) and polysaccharide lyases (PL) encoded in GB genomes summarized for each phylogenetic cluster. Brackets: Total number of genomes encoded in each phylogenetic cluster. Asterisk: CAZyme with potential secretion signal (see also Supplementary Data 8).
Data 10). The acyl-CoA dehydrogenase (acd) represents a key gene catalyzing the first step in beta-oxidation and accommodates a broad substrate range. GB ACDs fell alongside described glutaryl-CoA dehydrogenases, small/medium- and long-chain acyl-CoA dehydrogenases, potential butyryl-CoA dehydrogenases and isovaleryl-CoA dehydrogenases (Supplementary Figure 7, Supplementary Data 12). Only ~50% of archaean lineages encoded for acd, which was found scattered across taxa, for example only ~30% of Verstreetaarchaeota encoded for acd. This gene was common in Archaeoglobales, Asgard archaea and Geothermarchaeota, all of which encoded for other beta-oxidation genes, such as enoyl-CoA hydratase (EC 4.2.1.17) or 3-hydroxyacyl-CoA dehydrase (EC 1.1.1.35; Fig. 4, Supplementary Data 10). In contrast, 33 out of 37 bacterial lineages encoded for acd. However, only a subset of those lineages - including Aquificae, Chloroflexi or Deltaproteobacteria - encoded for further beta-oxidation genes. In these cases, enzymes, such as the glutaryl-CoA dehydrogenase, might be involved in amino acid catabolism or in benzoyl-CoA degradation.

Hydrocarbons are another abundant source for energy and biomass generation in GB. While we did not detect genes for aerobic hydrocarbon degradation, we found indications that GB genomes might anaerobically degrade hydrocarbons using glycol radical enzymes (GREs, Supplementary Figure 8, Supplementary Data 13). GREs use a radical-based chemistry to carry out challenging metabolic reactions under anaerobic conditions and are involved in a multitude of pathways, such as fermentation, DNA synthesis or hydrocarbon degradation. Compared to ACDs, GREs had a sparser distribution and were found in only 6 out of 24 archaean and 21 out of 37 bacterial lineages. GREs were common in Deltaproteobacteria (n = 32), Bacteroidetes (n = 23) or Asgard archaea (n = 15). Several GREs encoded for enzymes involved in anaerobic hydrocarbon degradation, such as benzylsuccinate synthase (bssA) in Deltaproteobacteria (B38_G6, B7_G9), alkylsuccinate synthase (assA) in Deltaproteobacteria (B2_G1, B111_G9) or hydroxynaphthalene carboxybatainase in Bathyarchaeota (B26_G17) and Chloroflexi (B43_G15). Some GREs grouped neither with the previously mentioned enzymes.
Respiratory processes. Next, we investigated the GB microbial communities for their involvement in respiratory processes. Overall, more bacterial and archaenal genomes contained genes that encode anaerobic rather than aerobic respiratory pathways, consistent with rapidly depleted oxygen levels within the first few millimeters of the sediment (Fig. 4, Supplementary Data 10).

Cytochrome c oxidases occurred in ~10% of genomes, but were mainly limited to Bacteroidetes, Epsilon- and Gammaproteobacteria, and Verrucomicrobia. Conversely, genes for hydrogen, nitrogen, sulfur and potentially arsenate and selenate cycling were more widespread. We detected [FeFe]-hydrogenases in ~10% of genomes and these mostly belonged to Group A, which can be involved in fermentative hydrogen evolution. Approximately 70% of genomes encoded for [NiFe]-hydrogenases belonging to Group 1 (a–e and h; membrane-bound hydrogen-uptake hydrogenases involved in hydrogenotrophic respiration), Group 3 (a–d; cytosolic bidirectional hydrogenases) and Group 4 (b,d,e and g; membrane-bound, hydrogen-evolving hydrogenases; Supplementary Data 10). The most common [NiFe]-hydrogenase was found in ~25% of genomes, and belongs to Group 3b that is involved in NADPH oxidation coupled to hydrogen evolution. Genes involved in the nitrogen and sulfur cycle were mostly restricted to bacteria, whereas archaenal nitrogen cycling genes were limited to nifH in Methanomicrobia (Fig. 4, Supplementary Data 10). Genes for dissimilatory nitrate reduction to ammonium (DNRA) were present in few Bacteroidetes (i.e. B27_G6, B58_G6), Epsilonproteobacteria (B6_G4, B37_G6) and several Gammaproteobacteria (Methylococcaceae, Thiotrichales). More commonly, we detected DNRA nor with the pyruvate formate lyase or other characterized GREs, suggesting that those might utilize different substrates, such as carbohydrates or peptides.

Fig. 5 Maximum likelihood phylogenetic tree of the methyl-Coenzyme M reductase (McrA) protein detected in GB genomes. Bold labels: McrA detected in GB genomes (see also Supplementary Data 10). Black circle: Bootstrap support ≥ 70 (number of bootstraps determined using the extended majority-rule consensus tree criterion). RaxML was run as raxmlHPC-PTHREADS-AVX -f a -m PROTGAMMAAUTO -N autoMRE. The tree file is available in Supplementary Data 11.
genes distributed separately over several genomes. A complete denitrification pathway (napA/napGH, nirK/nirS, norBC, nosZ) was present in a few genomes, including one Bacteroidetes (B2_G4), some Epsilonproteobacteria (i.e. B135_G9) and several Gammaproteobacteria genomes (Halieaceae, Thiotrichales); individual denitrification genes were found scattered across different taxonomic lineages. Genes involved in anaerobic ammonium oxidation (anammox) were not found, consistent with low nitrate and nitrite concentrations in GB sediments. Genes for the dissimilatory reduction of sulfate to sulfide (sat, aprAB and dsrAB) were found in few archaea (i.e. Archaeoglobales) and several bacteria including Deltaproteobacteria, Gammaproteobacteria and Zixibacteria. The sulfur-oxidation (SOX) system (soxAX, soxYZ, soxB, soxC/D) showed a restricted phylogenetic distribution and was only located in Epsilonproteobacteria and Gammaproteobacteria. While on average ~10% of all genomes contained genes for sulfur and nitrogen cycling, complete pathways for these processes were present in only few genomes.

**Redundancy and interconnectivity among GB microbes.** To assess whether hydrothermal sediments not only host a greater phylogenetic but also metabolic diversity than background samples (Fig. 2), we next investigated the spatial distribution of core metabolic genes across all sites and taxa. Regardless of their origin, most genomes encoded genes for general carbon cycling (CAZymes, peptidases, gluconeogenesis, glycolysis), fermentation and lipid oxidation (Fig. 6 and Supplementary Data 10). Respiratory genes were restricted to cooler, shallower samples but present in both background and hydrothermal sediment cores. For example, denitrification genes, SOX genes or the cytochrome c oxidase were found only in the shallower, colder sediments (temperature ~5 °C) and were present in ~20-30% of genomes. In contrast, these genes were represented in only ~0-4% of genomes in deeper, hotter samples (temperature range of 10 °C-60 °C). Exceptions were genes for sulfate/sulfite reduction, such as dsrAB, that were still found in ~8% of genomes in deeper, hotter sediments. Compared to background samples, genes involved in C1 metabolism and hydrogenases were more frequently found in hydrothermal sediments. In background sediments only one Bathyarchaeotal genome contained carbon fixation-related genes (cdhAB), while genes for methane cycling (mcrA) were undetectable. Hydrogenases belonging to Group 4 g, which represent membrane-bound hydrogenases that generate a proton-motive force for energy generation, were absent from the background but present in ~25-30% of genomes across all hydrothermal samples (Fig. 6 and Supplementary Data 10). These findings suggest that methane and hydrogen might be important drivers of metabolic processes in GB hydrothermal sediments.

With few exceptions most metabolic genes were encoded in several taxonomically distinct lineages. For example, C1-related genes (with the exception of mcrA) and genes related to beta-oxidation, hydrogen, nitrogen, sulfur and oxygen cycling were found in ~10 different phylogenetic lineages; fermentation genes were present in most phylogenetic clusters of both the archaean and bacterial community. While the studied genomic dataset from the cold and hydrothermal samples were not represented by an equal number of genomes (average of ~9 and ~60 genomes per habitat type, respectively), we still find that those genomes represent the community well in terms of phylogenetic diversity (Supplementary Figure 4). Additionally, when searching for a subset of these core metabolic genes in binned and unbinned contigs from the complete assembly (only considering contigs >2,000 bp), we observed a similar trend (Supplementary Data 14). For example, fermentation genes were abundant across all sites, denitrification genes were more common in cold and shallow samples and mcrA was completely absent from the background samples. Overall, these findings suggest that the GB genomes are representative of the community as a whole, and that they reflect key metabolic differences between the microbial communities present in hydrothermal and background samples.

**Discussion**

In this study, we employed the largest genomic sampling of GB sediments to date to investigate the interplay of community composition and functional diversity. Compared to earlier work on Guaymas Basin sediments, the higher sampling number and inclusion of background samples allowed to better describe the enhanced diversity present in these sediments and shed light on the drivers of community assembly. In contrast to previous studies showing that sulfidic- and methane-rich seep sediments host a lower microbial diversity compared to non-seep marine sediments, we demonstrate that GB hydrothermal sediments contain a diverse community that is enriched in archaea compared to a less diverse, bacterial-dominated community found in nearby cold sediments. Therefore, the more extreme conditions in hydrothermal sediments, which include steep thermal and geochemical gradients, appear not to inhibit microbial diversity. Due to difficulties in isolating sufficient amounts of DNA from deeper, hotter samples, we cannot exclude that diversity may decline in those sediments. Earlier work reported a decrease in cell numbers with increasing depth that did not necessarily correlate with a decrease in OTU numbers, potentially explaining our difficulties in isolating sufficient amounts of DNA but supporting our assumption that steep temperature gradients do not necessarily inhibit microbial diversity. Especially samples from core 4569.9 experience a highly variable, fluctuating thermal regime over time, where even surficial layers can vary from 20 °C to 70 °C, as determined by multi-day continuous thermal logging (Supplementary Figure 1). In response to such conditions, microbes must either adapt, have a wide thermal optimum, as shown for some ANME-1 archaea, or be able to recolonize the sediment after a temperature sweep from a surficial reservoir. Here, we propose that the diverse communities inhabiting hydrothermal sediments could serve as a flexible seed bank for the deeper, hotter sediments as well as for highly fluctuating environmental gradients in shallow sediments.

The differences we observed in community composition across sites were not always translated into obvious changes in functional capacities of those communities. For example, we detected abundant genes for carbon cycling and fermentation across all sites, while other metabolic processes such as respiration, were limited to shallow sediments but present in both background and hydrothermal sediments. Respiratory processes were often partitioned among the community and only few genomes were encoding for full pathways. Metabolic handoffs have been observed in other microbial communities and could allow a flexible interchange of metabolites between changing populations. Another metabolic feature that could allow for greater ecosystem stability could be metabolic plasticity, i.e. switching metabolic processes in response to changes in environmental conditions. We found indications for such plasticity in several bacterial genomes, especially within the Delta- and Gammaproteobacteria that might couple the reduction of sulfur with the oxidation of carbon, lipids or hydrocarbons. While we cannot determine which processes are active, enhanced genotypic diversity might provide an additional adaptation strategy to variable environmental conditions.

The only functional categories that were consistently enriched across all hydrothermal sites and almost absent in background sediments were group 4g hydrogenases and pathways for...
**Fig. 6** Metabolic profile across different GB sediment sites, depth profiles and temperature regimes. Shown is the number of core metabolic genes relative to the total number of genomes (in %) per site, depth and temperature regime. Temperatures are averages for the 2 or 3 cm thick sediment layers from which DNA was isolated. Background samples: Cold GB samples without hydrothermal activity. Vent1-3: Hydrothermal sediment sampling locations, see also Fig. 1. ID at the bottom: number codes designating every Alvin dive and sediment core (see also Supplementary Data 1 for further explanation). A complete list of metabolic genes can be found in Supplementary Data 10. Number in circles: Number of phylogenetic clusters that encode for individual core metabolic genes at each site.

| Site        | Background | Vent1 (out) | Vent1 (intermediate) | Vent1 (center) | Vent2 | Vent3 |
|-------------|------------|-------------|----------------------|----------------|-------|-------|
| Depth (cm)  | 0–3 cm     | 21–24 cm    | 0–3 cm               | 0–3 cm         | 0–3 cm| 4–6 cm|
| Temperature | 3°C        | 6°C         | 12–15 cm             | 21°C           | 10°C  | 60°C |

| Metabolic Pathway | Core Genes | No. of Genes/Size of Circle | No. of Phylogenetic Clusters |
|-------------------|------------|-----------------------------|-------------------------------|
| Fermentation      | korA, porA, fdoG, ald, adh, ackA, acs | 25, 50, 75, 100 | x |
| C1                |                                                                                     |                            |                |
| HCs               |                                                                                     |                            |                |
| Hydrogenases      | FeFe NiFe G1, NiFe G3b NiFe G3c NiFe G3d NiFe G4, napA, narG, nirB, nrfH, nirK, nirS, norB, nosZ, nifH, hao, hpc, octR |                      |                |
| Nitrogen          |                                                                                     |                            |                |
| Sulfur            |                                                                                     |                            |                |
| C2                |                                                                                     |                            |                |

| Gene          | 4567_28 | 4569_4 | 4569_2 | 4569_9 | 4571_4 | 4488_9 |
|---------------|---------|--------|--------|--------|--------|--------|
|噬菌体            | 4567_28 | 4569_4 | 4569_2 | 4569_9 | 4571_4 | 4488_9 |
methanogenesis and methane oxidation. Group 4g hydrogenases are not well characterized but are generally described to be membrane-bound hydrogenases that allow for energy-generation by establishing ion gradients over the membrane. These complexes are often found in thermophiles, such as *Pyrococcus furiosus*, and could potentially provide a selective advantage in hydrothermal sediments over other energy-generating systems. While trace concentrations of biogenic methane are present in hydrothermal sediments, which might reduce complex stability, novel lineages, including a new deep-branching Crenarchaeota, novel carbon and energy sources. Most functional properties are shared widely among different phylogenetic lineages across different sampling sites. The combination of dynamic seep and hydrothermal conditions in Guaymas Basin enhances microbial diversity compared to surrounding non-hydrothermal sediments.

Guaymas Basin is a hotspot for microbial biodiversity and an ideal study site to investigate the functional diversity of hydrothermally influenced seafloor sediments. Here we establish that these hydrothermal sediments contain a large number of archaeal and bacterial lineages, including several uncultivated phylum-level lineages that have not been described from other habitats. Intriguingly, hydrothermal GB sediments hosted a greater diversity compared to surrounding non-hydrothermal sediments, contrasting previous work on methane seep communities. These differences are likely linked to the unique environment in GB sediments characterized by convective mixing of fluids resulting in variable thermal regimes, and admixture of hydrothermal carbon and energy sources. Most functional properties are shared widely among different phylogenetic lineages across different sampling sites. The combination of dynamic seep and hydrothermal conditions in Guaymas Basin enhances microbial diversity, and sustains a distinctive microbial community, whose functional complexity and redundancy reflects the intricate and dynamic geochemical and thermal landscape of this habitat.

**Methods**

**Sampling.** Guaymas Basin sediment samples were collected from the Gulf of California (27°N0.388, 111°W24.560) at a depth of approximately 2000 m below the water surface. Sediment cores were collected during four Alvin dives (4488, 4569, 4567, and 4571) in 2008 and 2009 (Supplementary Data 1). Sample site photos were compiled from the Alvin frame grabber site [http://4dgeo.whoi.edu/alvin](http://4dgeo.whoi.edu/alvin). Intact sediments were collected during Alvin dives using polycarbonate cores (45-60 cm in length, 6.25 cm interior diameter), subsampled into cm layers under N₂ gas in the ship’s laboratory and immediately frozen at ~80°C. Eleven sediment subsamples for DNA isolation from different depth profiles yielded sufficient genomic DNA for metagenomic sequencing (Supplementary Data 1). Higher temperature samples were tested as well but did not yield sufficient DNA for metagenomic sequencing. Metadata for all dives, including details on the geographic (i.e. methane concentrations and dissolved organic carbon concentrations and δ¹³C values, sulfate and sulfide concentrations) and thermal profiles of the sampling sites, are available to compare microbial community composition across sediment cores (Supplementary Data 1, Supplementary Methods). Additional images and descriptions of the sampling locations are published in a survey of different Guaymas Basin habitats.

**Metagenomic sequencing and assembly.** Total DNA from ≥10 g of sediment from each of the eleven samples (see above) was extracted using the MoBio PowerMax soil kit using the manufacturer’s instructions. DNA concentrations were measured using a Qubit™ 3.0 Fluorometer and a final concentration of 10 ng/µl of each sample (using a total amount of 100 ng) was used to prepare libraries for paired-end Illumina (HiSeq®-2500 I1B) sequencing. Illumina library preparation and sequencing was performed at the Joint Genome Institute (JGI). Sequencing was performed on an Illumina HiSeq 2500 machine using the paired end 2 × 125 bp run-type mode. All runs combined provided a total of ~280 gigabases of sequencing data (Supplementary Data 2) Quality control and sequence assembly was performed by JGI. Briefly, sequences were trimmed and screened for low quality sequences using bbtools ([https://jgi doe gov/data-and-tools/bbtools/](https://jgi doe gov/data-and-tools/bbtools/)) and assembled using megapit v1.0.6 using the following options: –k-list 23,43,63,83,103,123. Summary statistics for the number of generated reads and the quality of the metagenomic assembly is provided in Supplementary Data 2. For further binning, only scaffolds ≥2000 bp were included.

**Metagenomic binning.** Metagenomic binning was performed on individual assemblies using the binning tools ESOM, Anvio and Metabat. ESOM binning was performed by calculating tetranucleotide frequencies of scaffolds with a minimum length of 2000 bp using the K-batch algorithm for training after running the perl...
search followed by the ‘phyloseq align’ mode. The concatenated protein alignments of 37 elite marker genes (concat.updated.1.fasta) were combined for all genomes of SV (version 2017.1) using TreeFamR. A phylogenetic tree was generated using a maximum likelihood-based approach using RAXML (version 8.2.10, called as: raxmlHPC-PTHREADS-AVX -f a -m PROTGAMMAAUTO -N autoMRE)60. The tree was visualized using the Interactive Tree Of Life (ITOL) webtool61. For better visualization, the initial tree was reduced to only include reference genomes and included a 224 genomes from cultured representatives and 330 genomes from uncultured genomes (including metagenome-assembled genomes, enrichment cultures, co-cultures and single-cell assembled genomes). All of these genomes were identified and organized using anvi-gen-contigs-database. These two files were further used as input for anvi-profile. Profiled genomes for the eleven different assemblies were combined using anvi-merge and the resulting bins summarized using anvi-summarize (-C CONCOT). If not mentioned otherwise, the scripts were used with default settings. Finally, binning was performed using metabat (v1)61. As described for Anvi’s the used input files consisted of the scaffold files (>2000 bp) and the mapping files to recover bins both by sequence composition and abundance across samples. First, each of the mapping files were summarized using jgi_summarize_bam_contig_depths and then metabat was run using the following settings: --minProb 75 --minContig 2000 --minContigByCov 2000. Results from the three different in the gplts packages could be combined, with the notable exception of the sample from 4567_28, from which the recovered MAGs only recruited ~18% of reads for an undetermined reason.

To determine the average abundance of major taxonomic groups (referred to as clusters), which were determined by the phylogenetic analysis described below, contigs were first assigned to their phylogenetic cluster (see description for the phylogenetic analysis below) and then summarized using the dpdp function from the phylr package62. These clusters do not represent a specific taxonomic rank but were chosen to account for both phylogenetic diversity (i.e., Crenarchaeota are usually represented at order rank or lower if possible) as well as available genomes (the different clusters represent a spectrum of taxonomic ranks and are ranked together based on their phylogenetic diversity). The counts were represented by only few genomes. The counts recruited by each taxonomic cluster, which were determined by the phylogenetic analysis described below, were then summarized using the dpdp function from the phylr package. The phylr package was used to calculate the phylogenetic identity of the genomes using the phyloseq package63.

Relative abundance. To determine the relative abundance of each genome across the eleven sequenced sediment samples, we mapped the contigs from all binned genomes to the representative genomes (fasta --write_bins 1). The binning process was enhanced by incorporating reference genomes (fasta --loyal 51). The binning process was enhanced by incorporating reference genomes (fasta --loyal 51). The binning process was enhanced by incorporating reference genomes (fasta --loyal 51). The binning process was enhanced by incorporating reference genomes (fasta --loyal 51).

To determine the average abundance of major taxonomic groups (referred to as clusters), which were determined by the phylogenetic analysis described below, contigs were first assigned to their phylogenetic cluster (see description for the phylogenetic analysis below) and then summarized using the dpdp function from the phylr package62. These clusters do not represent a specific taxonomic rank but were chosen to account for both phylogenetic diversity (i.e., Crenarchaeota are usually represented at order rank or lower if possible) as well as available genomes (the different clusters represent a spectrum of taxonomic ranks and are ranked together based on their phylogenetic diversity). The counts were represented by only few genomes. The counts recruited by each taxonomic cluster, which were determined by the phylogenetic analysis described below, were then summarized using the dpdp function from the phylr package. The phylr package was used to calculate the phylogenetic identity of the genomes using the phyloseq package63.

Relative abundance. To determine the relative abundance of each genome across the eleven sequenced sediment samples, we mapped the contigs from all binned genomes to the representative genomes (fasta --write_bins 1). The binning process was enhanced by incorporating reference genomes (fasta --loyal 51). The binning process was enhanced by incorporating reference genomes (fasta --loyal 51). The binning process was enhanced by incorporating reference genomes (fasta --loyal 51). The binning process was enhanced by incorporating reference genomes (fasta --loyal 51).

To determine the average abundance of major taxonomic groups (referred to as clusters), which were determined by the phylogenetic analysis described below, contigs were first assigned to their phylogenetic cluster (see description for the phylogenetic analysis below) and then summarized using the dpdp function from the phylr package. The phylr package was used to calculate the phylogenetic identity of the genomes using the phyloseq package63.

Relative abundance. To determine the relative abundance of each genome across the eleven sequenced sediment samples, we mapped the contigs from all binned genomes to the representative genomes (fasta --write_bins 1). The binning process was enhanced by incorporating reference genomes (fasta --loyal 51). The binning process was enhanced by incorporating reference genomes (fasta --loyal 51). The binning process was enhanced by incorporating reference genomes (fasta --loyal 51). The binning process was enhanced by incorporating reference genomes (fasta --loyal 51).

To determine the average abundance of major taxonomic groups (referred to as clusters), which were determined by the phylogenetic analysis described below, contigs were first assigned to their phylogenetic cluster (see description for the phylogenetic analysis below) and then summarized using the dpdp function from the phylr package. The phylr package was used to calculate the phylogenetic identity of the genomes using the phyloseq package63.
CAZymes database searches are summarized in Supplementary Data 8 and 9. In the case of protein-coding genes hitting to multiple genes in the before-mentioned databases, the best hit was chosen based on their e-value and bit-score using blast_best.pl (http://aiblab.research.pdx.edu/aquifiles/scripts/).

Genes assigned to core metabolic pathways are summarized in Supplementary Data 10. Hits for key metabolic marker genes found in major taxonomic clusters (Fig. 3) were verified across different databases (KAAS, PFAM, and TIGRFAMs) and cross-checked with results from close reference genomes that fell within the same phylogenetic group as the genome of interest to reduce the chance of contamination. Genes not found in close reference genomes were further validated with blastp using the NCBI webserver tool. If a hit could not be confirmed or if the top phylogenetic hit for whole contig was not consistent with the phylogenetic assignment of the genome, it was removed from the genome.

Data availability
All sequence data and sample information are available at NCBI under BioProject ID PRJNA362212. Accession numbers for individual genomes can be found in Supplementary Data 3. Additionally, the raw data is provided in IMG/MER and the IMG Genome IDs for the individual metagenomes are provided in Supplementary Data 1.

Received: 13 June 2018 Accepted: 31 October 2018
Published online: 27 November 2018

References
1. Hug, L. A. et al. A new view of the tree of life. Nat. Microbiol. 1, 16048 (2016).
2. Parkes, R. J. et al. A review of prokaryotic populations and processes in sub-seafloor sediments, including biosphere/gosphere interactions. Mar. Geol. 352, 409–425 (2014).
3. Lorzupone, C. A. & Knight, R. Global patterns in bacterial diversity. Proc. Natl Acad. Sci. 104, 11436–11440 (2007).
4. Orcutt, B. N., Sylvan, J. B., Knab, N. J. & Edwards, K. J. Microbial ecology of the dark ocean above, at, and below the sea. MMBR 75, 361–422 (2011).
5. Eisenhauer, N., Scheu, S. & Jousset, A. Bacterial diversity stabilizes community productivity. PLoS ONE 7, e34517 (2012).
6. Cordero, O. X. & Polz, M. F. Explaining microbial genomic diversity in light of evolutionary ecology. Nat. Rev. Microbiol. 12, 263–273 (2014).
7. Calvert, S. E. Origin of diatom-rich, varved sediments from the Gulf of California. J. Geol. 74, 546–565 (1966).
8. Einsele, G. et al. Intrusion of basalitic sills into highly porous sediments, and resulting hydrothermal activity. Nature 283, 441–445 (1980).
9. De la Lanza-Espino, G. & Soto, L. A. Sedimentary geochemistry of hydrothermal vents in Guaymas Basin, Gulf of California, Mexico. Appl. Geochem. 14, 499–510 (1999).
10. Simonet, B. R. T. & Lonsdale, P. F. Hydrothermal petroleum in mineralized mounds at the seafloor of Guaymas Basin. Nature 295, 198–202 (1982).
11. Dowell, F. et al. Microbial communities in methane- and short chain alkane-utilizing phylogenetically diverse orange and white seeps vs. hydrothermal vents sediments. Front. Microbiol. 7, 17 (2016).
12. Pearson, A., Seewald, J. S. & Eglinton, T. I. Bacterial incorporation of relict methane- and short chain alkane-utilizing phylogenetically diverse orange and white seeps vs. hydrothermal vents sediments. Deep Sea Res. Part Oceanogr. Res. Pap. 67, 21–31 (2012).
13. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MAG) of bacteria and archaea. Nat. Biotechnol. 35, 725–731 (2017).
14. Adam, P. S., Borrel, G., Brochier-Armanet, C. & Gribaldo, S. The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. ISME J. 11, 2407–2425 (2017).
15. Isenbarger, T. A., Finney, M., Rios-Velázquez, C., Handelsman, J. & Ruvkun, G. Miniprimer PCR, a new lens for viewing the microbial world. Appl. Environ. Microbiol. 74, 840–849 (2008).
16. Wrighton, K. C. et al. RubisCO of a nucleoside pathway known from Archaea is found in diverse uncultivated phyla in bacteria. ISME J. 10, 2702–2714 (2016).
17. Ghia, S. & Thorpe, C. Acyl-CoA dehydrogenases. Eur. J. Biochem. 271, 494–508 (2004).
18. Castelle, C. J., Brown, C. T., Thomas, B. C., Williams, K. H. & Banfield, J. F. Unusual respiratory capacity and nitrogen metabolism in a Pseudohalogena (OD1) of the Candidate Phyla Radiation. Sci. Rep. 7, 40101 (2017).
19. Wischgoll, S. et al. Decarboxylating and nondecarboxylating glutaryl-Coenzyme A dehydrogenases in the aromatic Metabolism of obligately anaerobic bacteria. J. Bacteriol. 191, 4401–4409 (2009).
20. Shuler, K. A. & Broderick, J. B. Glycol radical activating enzymes: Structure, mechanism, and substrate interactions. Arch. Biochem. Biophys. 546, 64–71 (2015).
21. Caccin, S. & Balhvs, E. P. Microbial conversion of choline to trimethylamine requires a glycol radical enzyme. Proc. Natl Acad. Sci. 109, 21307–21312 (2012).
22. Greening, C. et al. Genomic and metagenomic surveys of hydrogenase distribution indicate H2 is a widely utilised energy source for microbial growth and survival. ISME J. 10, 761–777 (2016).
23. Schulte, C. A. et al. Filamentous giant Beggioaaceae from Guaymas Basin are capable of both denitrification and dissimilatory nitrate reduction to ammonium (DNRA). Appl. Environ. Microbiol. AEM.02860-17. https://doi.org/10.1128/AEM.02860-17 (2018).
24. Ruff, S. E. et al. Global dispersion and local diversification of the methane seep microbiome. Proc. Natl Acad. Sci. 112, 4015–4020 (2015).
25. Lloyd, K. G. et al. Spatial structure and activity of sedimentary microbial communities underlying a Beggioa spp. mat in a Gulf of Mexico hydrocarbon seep. PLoS ONE 5, e6738 (2010).
26. Teske, A. et al. A molecular and physiological study of a diverse collection of hydrothermal vent Thermococcus and Pyrococcus isolates. Extremophiles 13, 905–915 (2009).
27. Gibbons, S. M. et al. Evidence for a persistent microbial seed bank throughout the global ocean. Proc. Natl Acad. Sci. USA 110, 4651–4655 (2013).
28. Anantharaman, K. et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat. Commun. 7, 13219 (2016).
29. Dombrowski, N. et al. Reconstructing metabolic pathways of hydrocarbon-degrading bacteria from the Deepwater Horizon oil spill. Nat. Microbiol. 1, 16057 (2016).
30. McTernan, P. M. et al. Intact functional fourteen-subunit respiratory membrane-bound [NiFe]-hydrogenase complex of the hyperthermophilic sulfur bacterium Pyrococcus furiosus. J. Biol. Chem. 289, 19364–19372 (2014).
31. Dombrowski, N. et al. Genomic insights into potential interdependencies in microbial hydrogenocarot and nutrient cycling in hydrothermal sediments. Microbiome 5, 106 (2017).
49. Loreau, M. Does functional redundancy exist? *Oikos* **104**, 606–611 (2004).
50. Louca, S. et al. Function and functional redundancy in microbial systems. *Nat. Ecol. Evol.* **2**, 936–943 (2018).
51. Hamilton, J. J. et al. Metabolic network analysis and metatranscriptomics reveal auxotrophies and nutrient sources of the cosmopolitan freshwater microbial lineage acl. *mSystems* **2**, e0092–17 (2017).
52. Zengler, K. & Zaramela, L. S. The social network of microorganisms — how auxotrophies shape complex communities. *Nat. Rev. Microbiol.* **16**, 383–390 (2018).
53. Embree, M., Liu, J. K., Al-Bassam, M. M. & Zengler, K. Networks of energetic and metabolic interactions define dynamics in microbial communities. *Proc. Natl Acad. Sci. USA* **115**, 10603–10608 (2018).
54. Wickham, H. The split-apply-combine strategy for data analysis. *J. Stat. Softw.* **40**, 1–22 (2011).
55. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-sufficient de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
56. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automatic alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 2972–2973 (2009).
57. Raghavan, M. & Schliep, K. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PloS One* **7**, e46056 (2012).
58. Darling, A. E. et al. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PloS One* **7**, e46056 (2012).
59. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 2972–2973 (2009).
60. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
61. Letunic, I. & Bork, P. Interactive tree of life (iTOl) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
62. Felsenstein, J. PHYLIP - Phylogeny inference package (version 3.2). *Mol. Biol. Evol.* **1**, 178–184 (1981).
63. Huelsenbeck, J. P. & Ronquist, F. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
64. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: a program to evolve phylogenies. *Bioinformatics* **19**, 733–734 (2003).
65. Larget, B. & Simon, D. A empirical Bayes approach to inferring evolutionary trees. *Comp. Appl. Biosci.* **10**, 305–312 (1994).
66. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
67. Happy, B. et al. SplitsTree4: a program for the analysis of phylogenetic and other evolutionary data. *Bioinformatics* **15**, 415–417 (1999).
68. Letunic, I. & Bork, P. Interactive tree of life (iTOl) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
69. Perna, N. T. & Kocher, T. L. GenBank: molecular sequences database. *Nucleic Acids Res.* **22**, 26–30 (1994).
70. Lartillot, N. & Philippe, H. Bayesian estimation of phylogeny and molecular evolution parameters using the Celest program. *Mol. Biol. Evol.* **20**, 87 (2003).
71. Rambaut, A. & Grassly, N. C. BEAST: Bayesian evolutionary analysis by sampling trees. *Mol. Biol. Evol.* **21**, 1167–1170 (2004).
72. Huelsenbeck, J. P. & Ronquist, F. MrBayes 3.2: evolution programs for the Bayesian analysis of phylogenetic trees. *Bioinformatics* **21**, 9–13 (2005).
73. 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).
74. Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–D141 (2004).
75. Haff, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).
76. Sonderegger, D., Pedersen, C. N. S. & Greening, C. HydDB: A web tool for hydrogenase classification and analysis. *Sci. Rep.* **6**, 24212 (2016).
77. Yin, Y. et al. dBiCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **40**, W445–W451 (2012).

**Acknowledgements**
We thank Kiley W. Seitz for detailed comments on the manuscript. 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. This work was funded by a Sloan Foundation Ocean Sciences fellowship (FG-2016-6301) and National Science Foundation DEB: Systematics and Biodiversity Sciences (grant number 1753661) provided to B.J.B. A.P.T. and Guaymas Basin fieldwork was supported by U.S. National Science Foundation grants OCE-0647633 and OCE-1357238.

**Author contributions**
B.J.B., A.P.T. and N.D. conceived, designed the study, and were involved in writing the manuscript. N.D. processed the data, reconstructed the genomes and performed the analyses.

**Additional information**
Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-07418-0.

**Competing interests:** The authors declare no competing interests.

**Reprints and permission** information is available online at http://npj.nature.com/reprintsandpermissions/.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.