The sulfur cycle connects microbiomes and biogeochemistry in deep-sea hydrothermal plumes

Zhichao Zhou[1], Patricia Q. Tran[1,2], Alyssa M. Adams[1], Kristopher Kieft[1,3], John A. Breier[4], Rupesh K. Sinha[5], Kottekkatu P. Krishnan[5], P. John Kurian[5], Caroline S. Fortunato[6], Cody S. Sheik[7], Julie A. Huber[8], Meng Li[9,10], Gregory J. Dick[11,12], Karthik Anantharaman[1,*]

[1] Department of Bacteriology, University of Wisconsin–Madison, Madison, WI, 53706, USA

[2] Department of Integrative Biology, University of Wisconsin–Madison, Madison, WI, 53706, USA

[3] Microbiology Doctoral Training Program, University of Wisconsin–Madison, Madison, WI, 53706, USA

[4] School of Earth, Environmental, and Marine Sciences, The University of Texas Rio Grande Valley, TX, 78539, USA

[5] National Center for Polar and Ocean Research, Ministry of Earth Sciences, Head land Sada, Vasco da Gama, Goa, 403804, India

[6] Department of Biology, Widener University, Chester, PA, 19013, USA

[7] Department of Biology and Large Lakes Observatory, University of Minnesota Duluth, Duluth, MN, 55812, USA

[8] Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA, 02543, USA

[9] Archaeal Biology Center, Institute for Advanced Study, Shenzhen University, Shenzhen, 518060, China

[10] Shenzhen Key Laboratory of Marine Microbiome Engineering, Institute for Advanced Study, Shenzhen University, Shenzhen 518060, China

[11] Department of Earth and Environmental Sciences, University of Michigan, Ann Arbor, MI, 48109, USA

[12] Cooperative Institute for Great Lakes Research, University of Michigan, Ann Arbor, MI, 48109, USA

*Correspondence and requests for materials should be addressed to K.A. (email: karthik@bact.wisc.edu)

Abstract

In globally distributed deep-sea hydrothermal vent plumes, microbiomes are shaped by the redox energy landscapes created by reduced hydrothermal vent fluids mixing with oxidized seawater. Plumes can disperse over thousands of kilometers and are complex. Their characteristics are determined by geochemical sources from hydrothermal vents, e.g., hydrothermal inputs, nutrients, and trace metals. However, the impacts of plume biogeochemistry on the oceans are poorly constrained due to a lack of integrated understanding of microbiomes, population genetics, and geochemistry. Here, we use microbial genomes to understand links between biogeography, evolution, and metabolic connectivity, and elucidate their impacts on biogeochemical cycling in the deep sea. Using data from 37 diverse plumes from 8 ocean basins, we show that sulfur metabolism defines the core microbiome of plumes and drives metabolic connectivity. Amongst all microbial metabolisms, sulfur transformations had the highest MW-score, a measure of metabolic connectivity in microbial communities. Our findings provide the ecological and evolutionary basis of change in sulfur-driven microbial communities and their population genetics in adaptation to changing geochemical gradients in the oceans.

Main

Hydrothermal vents are abundant and widely distributed across the deep oceans. The mixing of hot hydrothermally-derived fluids rich in reduced elements, compounds, and gasses, with cold seawater forms hydrothermal plumes[1, 2]. Usually, plumes rise up to hundreds of meters from the seafloor and can disperse over hundreds to thousands of kilometers through the pelagic oceans[3]. Surrounding microbes migrate into the plume and thrive on substantial reductants as the energy sources, making plumes 'hotspots' of microbial activity and geochemical transformations[1, 2]. Plumes constitute a relatively closed ecosystem that depends on chemical energy-based primary production and is mostly removed from receiving inputs of energy from the outside[4, 5]. Thus, plumes serve as an ideal natural bioreactor to study the processes and links between microbiome and biogeochemistry and the underlying ecological and evolutionary basis of microbial adaptation to contrasting conditions between energy-rich plumes and the energy-starved deep-sea[2].

The most abundant energy substrates for microorganisms in hydrothermal plumes include reduced sulfur compounds, hydrogen, ammonia, methane, and iron[2]. Amongst these, sulfur is a major energy substrate for diverse microorganisms in plumes across the globe[2, 6, 7, 8]. Sulfur transformations in plumes are dominated by oxidation of reduced sulfur species, primarily hydrogen sulfide and elemental sulfur. The metabolic pathways include oxidation of sulfide to elemental sulfur (*fcc*, *sqr*), oxidation of sulfur to sulfite (*dsr*, *sor*, and *sdo*), disproportionation of thiosulfate (*phs*) to hydrogen sulfide and sulfite, disproportionation of thiosulfate to elemental sulfur and sulfate (*sox*), thiosulfate oxidation to sulfate (*sox, tst,* and *glpE*), and sulfite oxidation to sulfate (*sat, apr*)[7, 9, 10, 11]. Complete oxidation of sulfur would involve oxidation of hydrogen sulfide all the way to sulfate. However, recent observations in other ecosystems indicate that individual microbes rarely possess a full set of the complete sulfide/sulfur oxidation pathway[10, 12], instead individual steps are distributed across different community members. This likely suggests that sulfur oxidation is a microbial community-driven process that is dependent on metabolic interactions, and asks for revisiting sulfur metabolism and biogeochemistry based on a holistic perspective of the entire community.
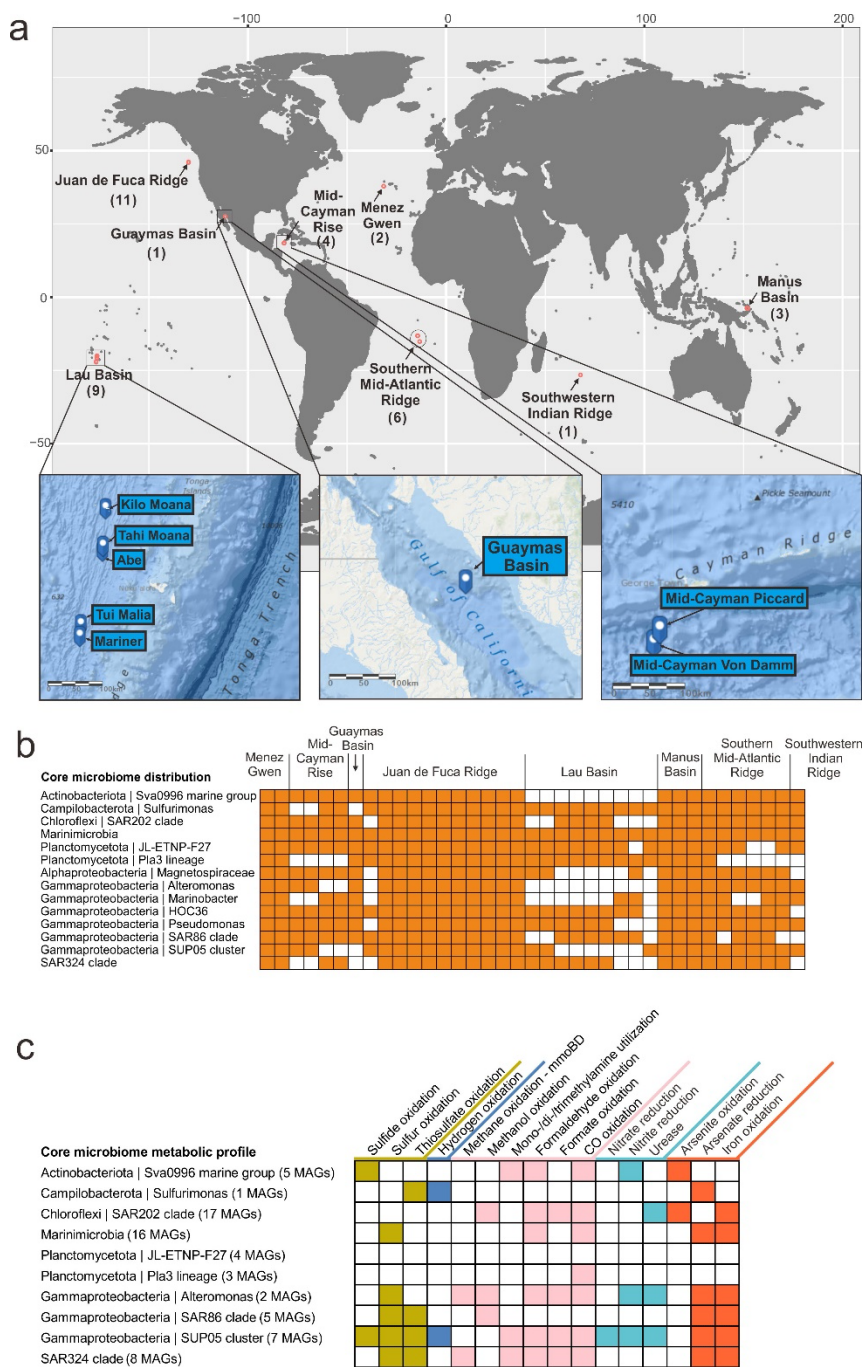
Recent microbiome-based ecological studies have focused on elucidating a genome-centric view of ecology and biogeochemistry[7, 10, 12, 13, 14, 15]. This approach has expanded our understanding of microbial diversity associated with specific energy metabolisms, including sulfur transformations in hydrothermal plumes, the deep sea, and beyond[7, 14, 16, 17, 18, 19]. However, the dynamics and microdiversity of the plume microbiome, and relevant biogeochemical impacts remain relatively underexplored[20, 21, 22, 23, 24]. Understanding how environmental constraints and selection shape the microdiversity and the genetic structure of plume microbial populations after migration from background seawater can provide fundamental insights into adaptation mechanisms. These insights can also inform future predictions of microbial responses to the changing oceans.

Here, we characterized the ecological and evolutionary bases of the assembly of the plume microbiome, and their strategies for sulfur cycling-based energy metabolisms. First, we studied globally distributed hydrothermal plume datasets to define a core plume microbiome. We followed this up with synthesis of genome-resolved metagenomics, metatranscriptomics, and geochemistry from three hydrothermal

90 vent sites (Guaymas Basin, Mid-Cayman Rise, and Lau Basin) to unravel community structure and
91 functional links to biogeochemistry, metabolic connectivity within plume and deep-sea communities,
92 and microdiversity in abundant microbial populations. We demonstrate that plume microbiomes have
93 a distinctive community composition and function, that is adapted towards energy conservation,
94 metabolic interactions, and stress response.

Results

96 We used publicly available microbiome data from hydrothermal vent plumes across the globe to (1)
97 define the core plume microbiome, (2) investigate plume microbiome structure, function, and activity,
98 and (3) identify links between plume microbiomes and geochemistry. To investigate the core
99 microbiome, we studied publicly available 16S rRNA gene datasets of hydrothermal plumes ($n = 37$)
100 and background deep-sea samples ($n = 14$) from eight ocean basins across the globe. To study the
101 microbiome structure, function, and activity, we reconstructed metagenome-assembled genomes
102 (MAGs) from three hydrothermal vent sites (containing both plume and background samples from
103 Guaymas Basin, Mid-Cayman Rise, and Lau Basin). We also mapped paired metatranscriptomes from
104 the same sites for some samples (Fig. 1, Fig. S1, and Supplementary Data 1). To study links between
105 biogeochemistry and the microbiome, we analyzed paired geochemical data from the above three
106 hydrothermal vent sites. To provide clarity on the plume and background samples, and DNA/cDNA
107 libraries used in this study, we provided a schematic diagram describing the locations of all samples in
108 the context of a hydrothermal vent system (Fig. S1).

Fig. 1 | Sampling sites, distribution, and metabolic profile of the core plume microbiome. a Sampling site maps of hydrothermal plume samples from which the 16S rRNA gene datasets were sourced. Numbers in brackets indicate dataset quantities. Three hydrothermal sites that have metagenome and metatranscriptome datasets in this study were specifically represented by inset maps. Ocean maps were remodified from ArcGIS online map (containing layers of "World Ocean Base" and "World Ocean Reference"; https://www.arcgis.com/). **b** Membership and distribution of the core plume microbiome. Heatmap shows the presence/absence of core plume microbial groups (tracing back to known taxonomic ranks from the genus-level taxa) in 37 hydrothermal plume 16S rRNA gene datasets across the world. **c** Metabolic profile of the core plume microbiome. From this study, MAGs that have 16S rRNA genes affiliated to the core plume microbiome were used as representatives (numbers labeled in brackets). This subpanel shows the presence or absence of metabolic potential associated with sulfur, carbon, nitrogen, hydrogen, and metal biogeochemical transformations.

109

110

Defining the core hydrothermal plume microbiome

To identify and study the core hydrothermal plume microbiome, we used 16S rRNA gene datasets from 51 hydrothermal plume and background deep-sea samples spread across eight ocean basins (Supplementary Data 2). Biogeographic patterns were delineated by Unifrac metrics of distance and PCoA-based ordination. Sample location had a stronger influence on biogeographic patterns than sample characteristics (plume/background) (Fig. S2, S3). Unweighted Unifrac PCoA plots indicated that paired plume/background deep-sea samples within the same site were closely correlated (Fig. S3). As revealed previously[2, 25, 26], this supports the understanding that the hydrothermal plume environment has its main constitutional microorganisms derived from surrounding seawaters, with

120   dispersal limitation having little effects locally.
121
122   We then identified genus-level taxa significantly distributed in plumes with high prevalence and
123   relative abundance. The core plume microbiome consists of 14 microbial groups (Fig. 1a, b) as
124   revealed from the 37 plume datasets with a cutoff of being distributed in at least two third of all plume
125   datasets and having at least 1% relative abundance on average. By choosing MAGs reconstructed from
126   this study that were affiliated to the same taxa, we characterized metabolic profiles for the core plume
127   microbiome which demonstrated highly versatile metabolic potential for utilizing various plume
128   substrates[2], including HS^-, S^0, H_2, CH_4, methyl-/C_1 carbohydrates, arsenite, and iron (Fig. 1c). Most
129   plume microorganisms are of seawater origin, consistent with prior reports[26] (Supplementary Table 1).
130   We also observed a small number of seafloor/subsurface dwelling and endosymbiotic microorganisms
131   that might be entrained in plumes[2, 27] (Supplementary Table 1). Collectively, our data suggest that
132   sulfur and other reduced organic/inorganic compounds significantly shape the global core plume
133   microbiome that are originally derived from the surrounding seawater.
134

Distinctive plume geochemistries influence energy landscapes and promote microbial growth

137   Previous thermodynamic modeling analyses have reflected energy landscapes for various
138   hydrothermal ecosystems[4, 7, 10, 16] by representing free energy yields for reactions of various energy
139   sources for microbial metabolism in hydrothermal fluids. Some of them have demonstrated the
140   consistency of thermodynamic modeling and omics-based biogeochemical estimation in individual
141   ecosytems[7, 10, 16]. Here based on geochemical parameters and predicted functions from reconstructed
142   MAGs (Fig. S4, S5, and Supplementary Data 3), we conducted an across-site comparison of
143   thermodynamic modeling and omics-based biogeochemical estimations to reflect the influences of
144   distinctive plume geochemical characteristics on plume microbes. We also conducted growth rate
145   analyses to identify whether microbial energy contributors are promoted with higher growth rates in
146   responding to differing geochemical conditions across plumes. To address these, we first reconstructed
147   plume energy landscapes through thermodynamic modeling (Fig. 2a).
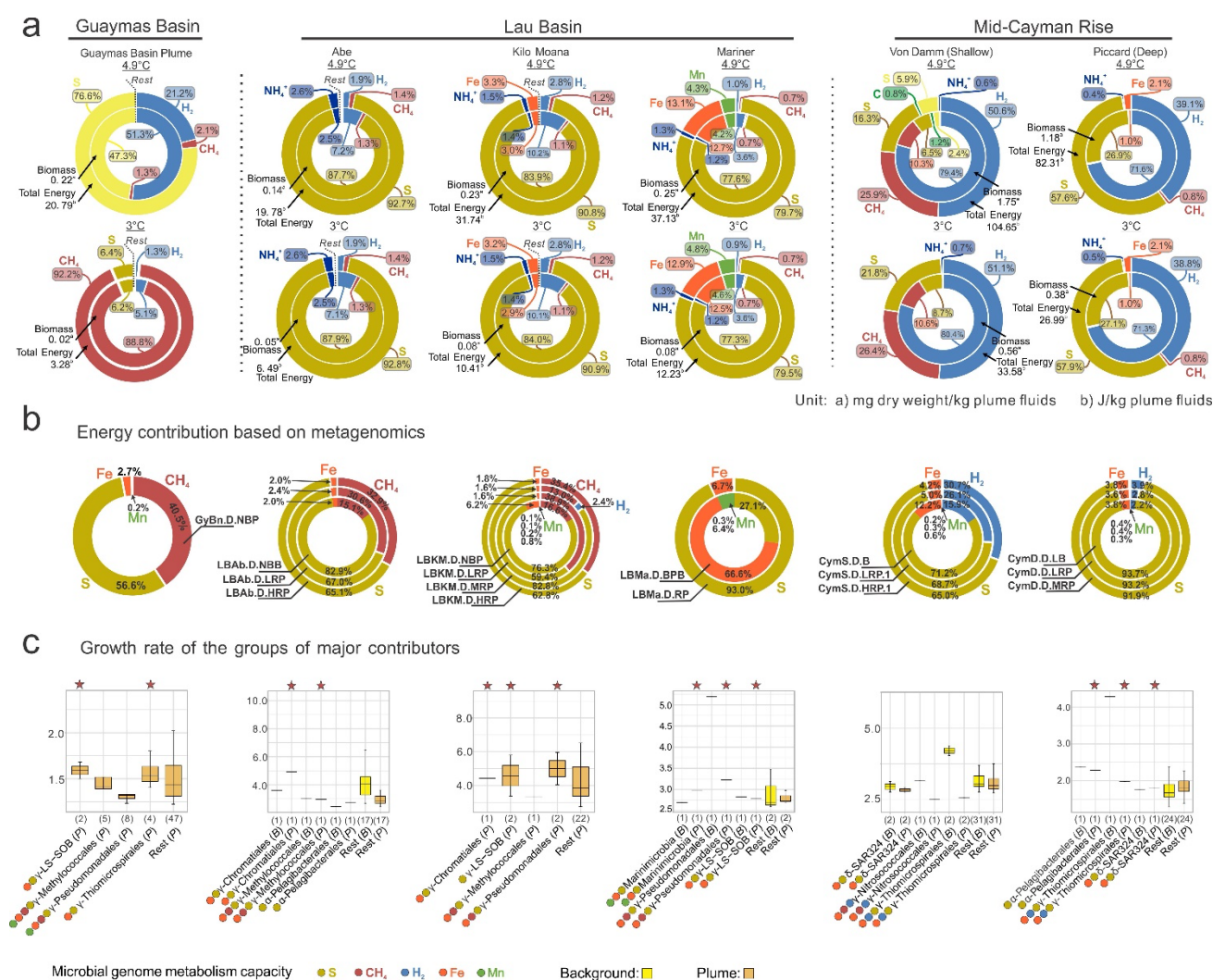148

149   Distinctive geochemical characteristics support the predicted energy landscapes when compared
150   among sites. Methane was the highest in end-member fluids from Guaymas Basin (63.4 mmol/kg)[7],
151   which supported the dominance of methane oxidation in the Guaymas Basin plume in the
152   thermodynamic model (Fig. 2a), and significant contributions of methane oxidation in metagenomics
153   datasets were also found (~40.5%) (Fig. 2b). Meanwhile, Lau Basin hydrothermal fluids had high Mn
154   and Fe concentrations (Mn: 3.9-6.3 mmol/kg, Fe: 3.8-13.1 mmol/kg)[28, 29] in the Mariner hydrothermal
155   field compared to other samples. This manifested in Fe and Mn oxidation contributing the highest
156   fractions (Mn: ~4-5%, Fe: 13%) in thermodynamic modeling (Fig. 2a) and the highest fractions (Mn:
157   0.3-6.4%, Fe: 6.7-66.6%) in omics-based estimations of Mariner among all sites (Fig. 2b). Similarly
158   in Mid-Cayman Rise, high hydrogen concentrations in the vent fluids were associated with high
159   contribution of hydrogen oxidation in the model, and in omics-based estimations (Fig. 2a, 2b,
160   Supplementary Table 2). Overall, reduced sulfur is the major energy source as reflected in both
161   thermodynamic modeling and omics-based biogeochemical estimations in all three sites. However,

162　individual plume geochemical conditions vary with diverse minor energy sources, such as iron,
163　methane, and hydrogen, leading to different energy landscapes which are mediated by microbes.
164
165　To study whether abundant organisms conducting biogeochemical transformations in each site were
166　also growing actively, we predicted microbial growth rates from metagenomic data using iRep[30]. iRep
167　can use a combination of cumulative GC skews and abundance of metagenomic reads to calculate the
168　difference in read abundance at the origin and terminus of a genome which is a proxy for the replication
169　or growth rate of organism[30, 31, 32]. The results suggest potential associations between growth rates and
170　geochemically-influenced energy landscapes for individual sites (Fig. 2c). A consistent pattern of the
171　abundant microorganisms in plumes having a higher predicted growth rate was also observed in certain
172　sites. For instance, LS-SOB and Thiomicrospirales both had the capacities for sulfur and iron oxidation,
173　and were predicted to have a higher growth rate than other microorganisms in Guaymas Basin plume
174　(Fig. 2c). Similarly, Methylococcales and Chromatiales were the major contributors to iron, methane,
175　and sulfur oxidation in Lau Basin (Abe plume) and their growth rates were higher than other organisms
176　(Fig. 2c). Collectively, we found a consistent pattern demonstrating that the abundant microorganisms
177　also have higher predicted growth rates potentially due to their ability to respond to varying
178　geochemistry in hydrothermal plumes.

Fig. 2 | Thermodynamic estimation of available free energies and biomass yields from electron donors, metagenomics-based contribution of electron donors to energy, and growth rates of major microbial contributors. a Thermodynamic estimation diagram of available free energy and biomass. For each hydrothermal environment, the contribution fraction of each electron donor species was labeled accordingly in the rings. The total available free energies and biomass were labeled accordingly to individual plumes. Two temperatures (3°C and 4.9°C) were picked to represent *in situ* temperatures in the upper and lower plume. Light yellow represents aerobic sulfur oxidation, dark yellow represents anaerobic sulfur oxidation. **b** Metagenomics-based estimation of energy contribution. Energy contribution for each electron donor was calculated based on metagenomic abundance of each reaction of electron donors and free energy yield of each reaction. The contribution ratio of electron donor species was calculated for individual environments respectively. **c** Growth rate of major microbial contributors in each hydrothermal environment. The y-axis for each barplot indicates the replication rate. The microbial groups starting with "α-", "γ-", and "δ-" represent Alphaproteobacteria, Gammaproteobacteria, and Deltaproteobacteria, respectively. Plume microbial groups were colored by dark yellow, background microbial groups were colored by light yellow and they were also all labeled with "(*P*)" or "(*B*)", respectively. Numbers in brackets indicate MAG numbers in each microbial group. Star-labeled plume microbial groups had higher growth rates than the 'Rest' plume microbial groups.

Across- and within-site comparisons for plumes show consistent links between geochemistry, function, and taxonomy

MAGs reconstructed from Guaymas Basin, Mid-Cayman Rise, and Lau Basin hydrothermal vents and corresponding omics-based profiling enabled taxonomic and functional comparisons among the three sites (Fig. S4, S5, and Supplementary Data 3). Across-site analyses of functional traits in MAGs indicate that different functions were significantly enriched in different plumes, e.g., arsenate reduction and long-chain alkane (C6+) degradation in Lau Basin, CO and methanol oxidation in Mid-Cayman

188 Rise, and toluene and benzene degradation in Guaymas Basin (Fig. 1c, Fig. S7b). In parallel, the
189 distribution and abundance of some microbial groups were also significantly enriched in the same
190 samples accordingly (Fig. S7a) and underlay the functional differentiation, e.g., arsenate reduction in
191 Lau Basin background deep-sea was attributed to members of Bacteroidetes and Thiomicrospirales
192 while that same function in Lau Basin plumes was attributed to only Thiomicrospirales. CO oxidation
193 in Mid-Cayman plumes was attributed to Chloroflexi, and toluene and benzene degradation in
194 Guaymas Basin plume attributable to Methylococcales and Pseudomonadales (Supplementary Data 5).
195 These observations are consistent with hydrothermal vent fluid geochemistry, e.g. Lau Basin
196 hydrothermal vents have high arsenic end-member concentrations[33] (ranging from 2.1-11 μmol/kg)
197 and Guaymas Basin fluids contain aromatic hydrocarbons (primarily benzene and toluene)[34].

199 As for within-site comparisons, the data indicate that the top three contributing taxa for major functions
200 (including eight categories, carbon fixation, denitrification, sulfur cycling, hydrogen oxidation,
201 methane oxidation, aerobic oxidation, iron oxidation, and manganese oxidation) are largely shared
202 between plume and background deep seawater in Mid-Cayman Rise and Lau Basin, indicating
203 functional consistency which was linked to taxonomy (Supplementary Data 5). Nevertheless, taxa
204 abundance differed between plume and background, as reflected by both DNA and cDNA datasets
205 associated with important functions (Supplementary Data 5, 6). Based on the results from energy
206 landscape and MAG-based comparisons, our results suggest the adaptation of the plume microbiome,
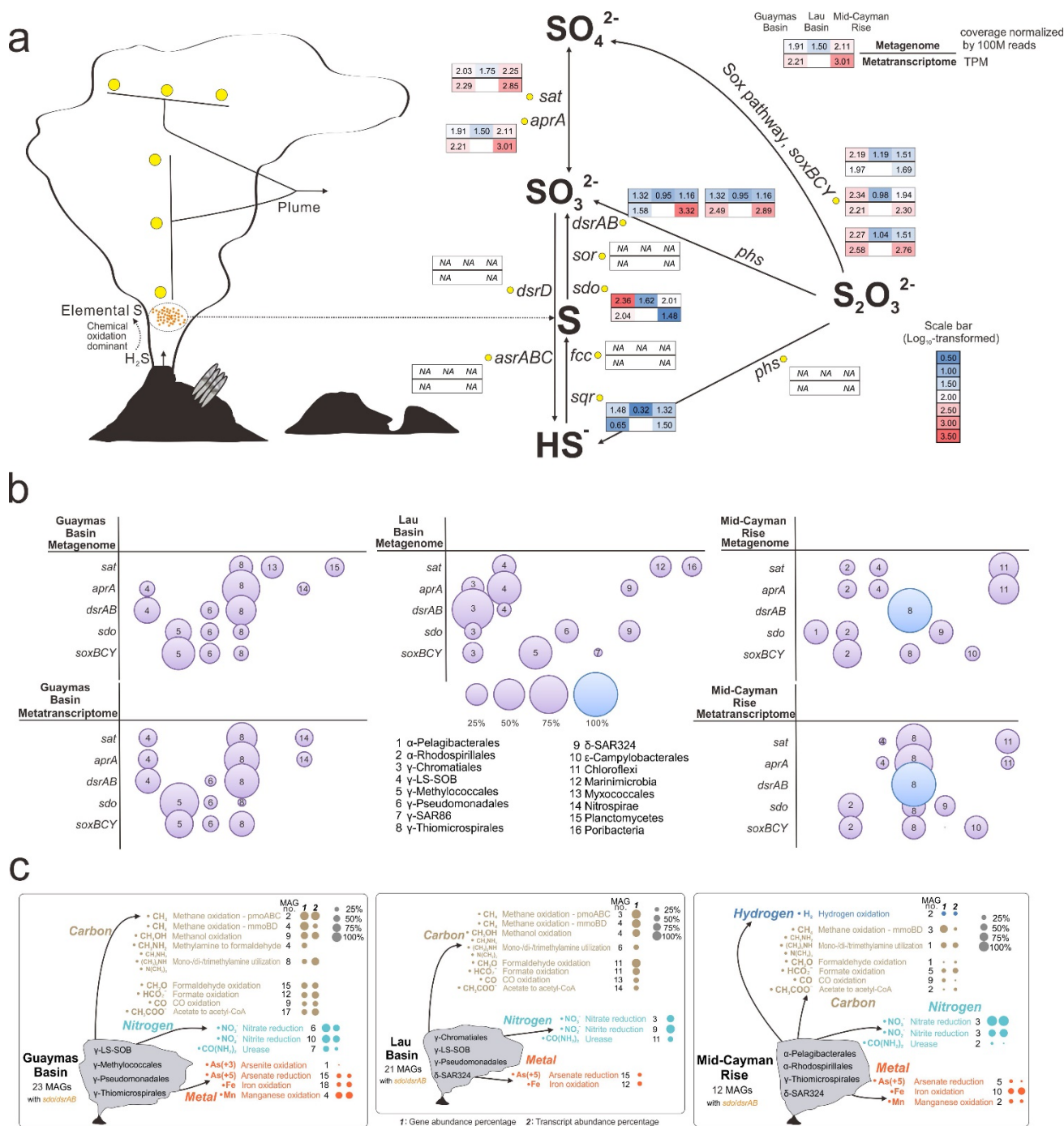207 and demonstrate the consistency of links between taxonomy, function, and geochemistry.

Sulfur cycling drives metabolic interactions in hydrothermal plumes

210 Building on our findings from both thermodynamic modeling and omics-based biogeochemical
211 estimations which indicated the importance of sulfur-based metabolisms, we studied microbial
212 metabolic interactions associated with sulfur cycling in all plumes. We recently developed a metric,
213 metabolic weight score (MW-score)[35] to measure the contribution of metabolic/biogeochemical steps,
214 and their metabolic connectivity in a microbial community. More frequently shared functions and their
215 higher abundances in a microbial community lead to higher MW-scores[35]. Both metagenomics and
216 metatranscriptomic data showed elemental sulfur oxidation to be the key reaction in the sulfur cycle
217 (Fig. 3a). In each community, sulfur oxidation had the highest MW-score (Fig. 4b, Fig. S10). Major
218 contributors (*dsrAB* and *sdo* containing MAGs) to sulfur oxidation varied in different hydrothermal
219 vent sites (Fig. 3b), indicating core sulfur oxidizers can have distinct distributions locally. Metabolic
220 overlaps existed as some sulfur oxidizers had additional metabolic potential associated with utilizing
221 various small carbon substrates and hydrogen, reducing nitrate/nitrite, and oxidizing
222 iron/manganese/arsenite[36] (Fig. 3c). Additionally, numerous connections of sulfur oxidation with other
223 electron-transferring reactions were observed in the functional network (Fig. 4b, c, d, and Fig. S10).
224 Previously, sulfur-oxidizing bacteria belonging to SUP05 (Thiomicrospirales in GTDB R83 or PS1 in
225 GTDB R202) and SAR324 lineages were identified to have metabolic plasticity involving the ability
226 to conduct hydrogen oxidation and nitrate reduction[7, 37] (in case of SUP05) and alkane/methane/carbon
227 monoxide oxidation[17, 38] (in case of SAR324) in plume and deep-sea environments, suggesting that
228 plume microorganisms are optimized to mediate energy transformations upon available electron
229 donors and acceptors. Here, our study indicates sulfur oxidizers are the primary group associated with

230   energy scavenging from plume substrates. Sulfur oxidizers have metabolic plasticity to connect sulfur
231   metabolism with other elemental transformations, are adapted to plume environments, and contribute
232   significantly to biogeochemical cycles in the deep sea.
233

234   While sulfur oxidation connects other metabolic reactions in the overall functional network and has
235   significant energy yields, its role on the overall network complexity remains elusive. Specifically, we
236   investigated the impact of sulfur metabolism on overall plume microbial metabolism. To address this,
237   we built networks based on reactions and the percent energy yields, and investigated reaction influence
238   on network complexity[39, 40, 41] (Fig. 4a, Fig. S11). The network of reactions works as a whole
239   mechanism where each reaction is one part[40] and high ΔC reactions are key features of the networks.
240   Most of these ΔC (complexity change) values are positive except for two points (Fig. 4a, Fig. S11).
241   This indicates that all but two of these reaction nodes drive the system away from randomness and
242   significantly contribute to the complexity of the network as a whole[40]. Meanwhile, in general, it seems
243   that most reactions that are closer to smaller ΔC have higher percent energy yields associated with
244   their reactions (Fig. 4a, Fig. S11). This phenomenon suggests that reaction nodes that result in higher
245   changes of percent energy yields are not necessarily contributing to the reaction network's complexity
246   the most. Overall, this suggests that while sulfur oxidation tends to have higher energy yields, other
247   reactions are also important components in plumes, and together cohesively contribute to the energy
248   landscape.

249
250

Fig. 3 | Sulfur metabolism and metabolic plasticity of sulfur oxidizers. a Details of sulfur metabolism pathways in the hydrothermal plume. The gene abundance (coverage normalized by 100M reads) and transcript expression level (TPM) for each step were calculated based on plume metagenomic and metatranscriptomic read mapping results. Log_{10}-transformed values of gene abundance and transcript expression level were labeled accordingly in the diagram. **b** Major contributors to sulfur metabolizing genes. For each sulfur metabolizing gene, microbial groups that occupied > 10% of the total gene abundance (by metagenome) or transcript expression (by metatranscriptome) values were labeled in the diagram. For some genes with only three or less than three contributors, all contributors were labeled. **c** Metabolic plasticity of sulfur oxidizers. For each hydrothermal vent site, three parameters were given to show the metabolic plasticity of sulfur oxidizers in conducting each electron transferring reaction related to carbon, nitrogen, hydrogen, and metal biogeochemical cyclings: the number of sulfur-oxidizing gene containing MAGs, gene abundance percentage, and transcript abundance percentage.
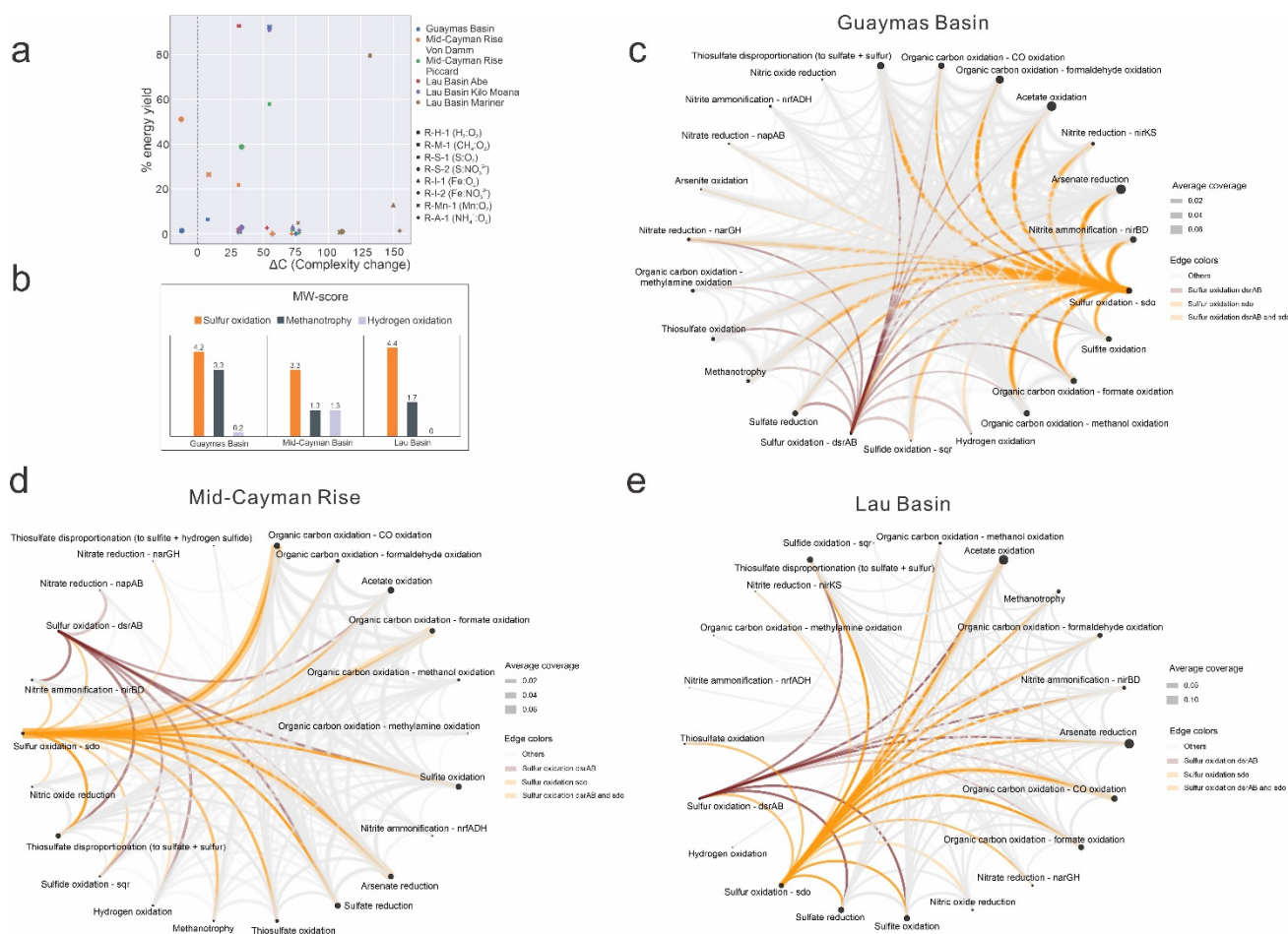
Fig. 4 | Network complexity, MW-scores (metabolic weight scores), and functional network diagrams of the three hydrothermal vent sites. a Network complexity diagram representing each reaction's influence on the complexity of the network. In the figure, different colors represent different hydrothermal environments, different symbol shapes represent different reactions. The substrates (including electron donors and acceptors) were listed for each reaction in the legend. The x-axis is the change in complexity (ΔC) of the whole network for a node (a reaction here) and the y-axis is the percent energy yield of that reaction in the whole community. This network complexity diagram was based on thermodynamic estimation results at $3°C$. **b** MW-scores of three major energy contributing reactions. **c** Functional network diagram of Guaymas Basin. **d** Functional network diagram of Mid-Cayman Rise. **e** Functional network diagram of Lau Basin. A group of metabolic cycling steps that are important in reflecting the plume substrate metabolisms were selected from METABOLIC-C regular MW-score results to make these functional network diagrams (**c, d, e**), respectively. In each functional network diagram, the size of a node is proportional to gene coverage associated with the metabolic/biogeochemical cycling step. The thickness of the edge represents the average gene coverage values of the two connected metabolic/biogeochemical cycling steps. Edges related to two reactions of sulfur oxidation were colored accordingly in each diagram.

Low diversity, short migration history, and gene-specific sweeps in plume populations

Metagenomes provide full repertoires of genomic variation and facilitate interpreting fine-scale evolutionary mechanisms[42, 43, 44]. Here, we used *Tara* Ocean metagenomic datasets[45] from the mesopelagic oceans to compare metagenomes from hydrothermal plume environments to the wider pelagic oceans and study the population genetic diversity of each MAG. We discovered that a large portion of MAGs exhibited a similar tendency of normalized single nucleotide variation (SNV) counts, nonsynonymous/synonymous substitution ratio of SNV (N/S SNV), and genome-wide mean r^2 (Fig. 5a and Supplementary Data 11). In hydrothermal plumes, their SNV count is lower than *Tara* Ocean samples, N/S SNV ratio is higher than *Tara* Ocean samples, and mean r^2 is higher than *Tara* Ocean samples. This suggests that in the plume: (1) Less SNVs are present, and population diversity is lower;

264 (2) The population is younger with a short migration history. The higher N/S SNV ratio indicates that
265 younger populations are less subjected to purifying (negative) selection to remove deleterious
266 mutations; (3) The population is less subjected to recombination. The higher mean r^2 reflects higher
267 SNV linkage frequency at the genome-wide scale, indicating a lower recombination rate among
268 population members.

269

270 Next, we investigated potential signals of genome/gene sweeps using fine-scale evolutionary
271 parameters. Consensus base frequency (frequency of reads supporting the consensus base), N SNV,
272 and S SNV all showed no significant differences between plumes and the pelagic ocean
273 (Supplementary Data 11). This indicates that these populations are unlikely to have undergone
274 selective genome sweeps and clonal expansion during migration. We calculated fixation index F_{ST}[46]
275 based on gene allele frequencies between these two environments (Fig. 5b and Supplementary Data
276 12) to investigate environmental selection. High F_{ST} genes are potential loci where selective pressures
277 act on and indicate niche-specific adaptation. Further stringent criteria require the lower gene
278 nucleotide diversity and higher N/S SNV ratio (Fig. 5b and Supplementary Data 12). Decreases of
279 nucleotide diversity indicate gene-specific selective sweeps in the hydrothermal plume environment,
280 and higher N to S SNV ratios suggest that these genes underwent a recent selection. Amongst 260
281 identified high F_{ST} genes using our stringent criteria, many of them involved transporters, aerobic
282 oxidation, and stress responses (Fig. 5b and Supplementary Data 12). Transporters were associated
283 with diverse substrates, e.g., metals (Co, Fe, and Mg), amino acids, Na^+/H^+, anions
284 (nitrate/sulfonate/bicarbonate), carbohydrates (ribose/xylose/arabinose/galactoside), and aliphatic
285 polyamines (spermidine/putrescine); meanwhile, these transporters are associated with many
286 transporter families (Supplementary Data 12), including ABC superfamily, tripartite ATP-independent
287 periplasmic (TRAP) family, tripartite tricarboxylate transporter (TTT) family, etc.

288

289 Given the observed importance of sulfur metabolism in plumes, we focused on the 238 identified sulfur
290 metabolism genes. 23 of these genes had signals of being fixed after migration with F_{ST} values higher
291 than the genome average (Fig. 5c and Supplementary Data 13). These genes were associated with
292 sulfur oxidation, thiosulfate oxidation, and sulfite oxidation/sulfate reduction (*sat*, *aprA*, *sdo*, oxidative
293 *dsrAB*, and *soxBC*) (Supplementary Data 13). This provides evidence that though not reaching the
294 level of gene-specific selection sweeps, these genes were still being selected across the genome.
295 Overall, this suggests a genetic adaptation to a sulfur-dominated environment after migration. An
296 underlying evolutionary paradigm can be outlined from our population-level microdiversity analyses
297 (Fig. 5c). As microbes enter the hydrothermal plume, some groups are selected for, and thrive due to
298 substrates provided locally. This stimulates the growth of certain populations; meanwhile, constraints
299 in the plume environment also induce selection effects and reduce the diversity of the population
300 majority. Higher N/S SNV indicates they are young populations and are growing in the plume,
301 consistent with the higher growth rates of major energy contributors. Gene-specific sweeps indicate
302 local adaptation to the plume environment, and partially change population genetic structures after
303 migration. Plume microbial populations are still in the early stage of evolution; as time goes on,
304 mutations will progressively accumulate and deleterious SNVs will be gradually purged.
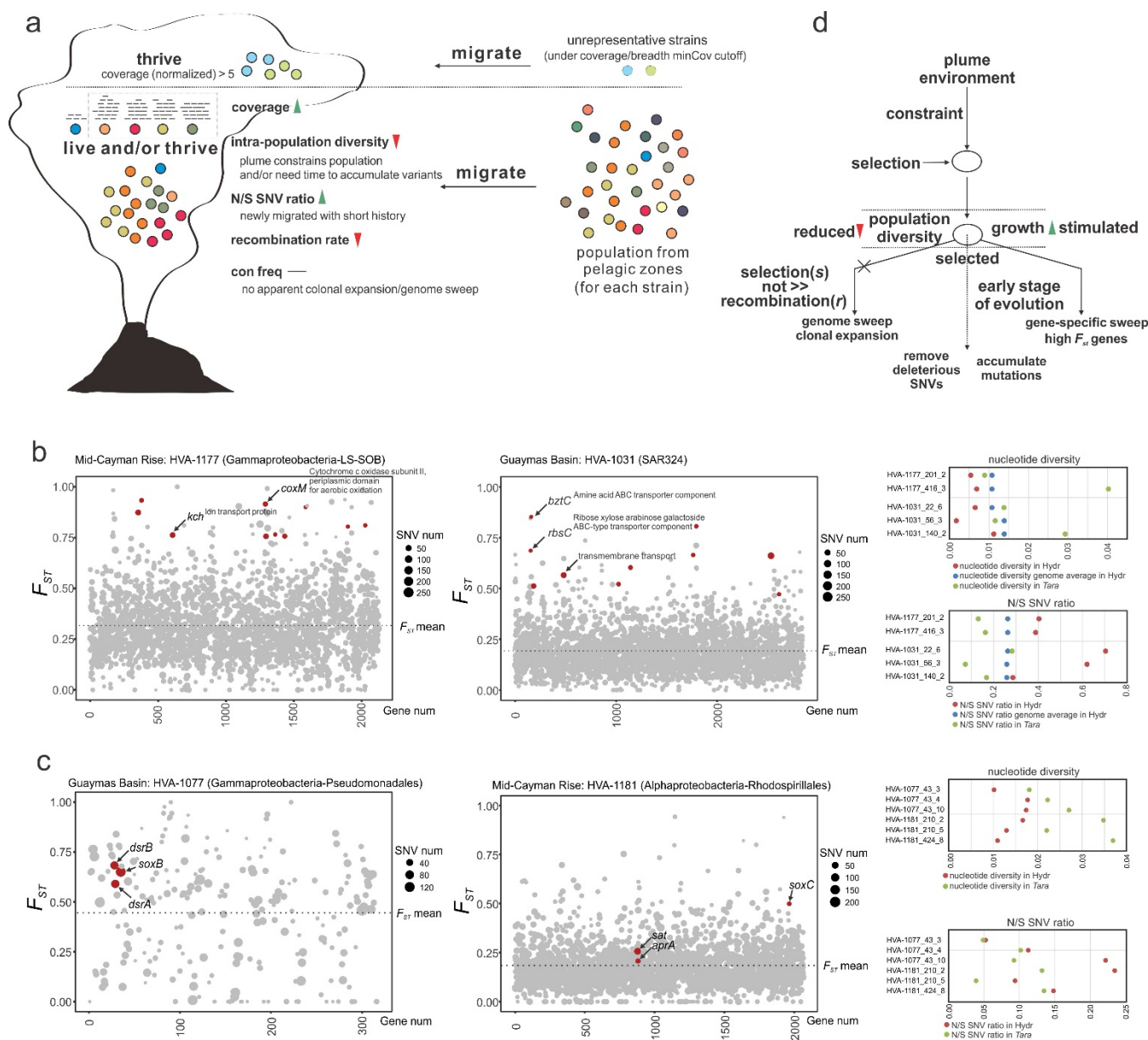
305

Fig. 5 | Evolutionary mechanism of plume microbial populations during migration. a Schematic diagram showing the changing trend of microdiversity parameters during migration. Individual solid dots with various colors represent microbial populations. Two scenarios were depicted in this panel: unrepresentative strains and strains that have detectable read mapping results in both environments. **b** Two representative charts showing F_{ST} distribution in MAGs that contain high F_{ST} genes. In each chart, the x-axis represents gene numbers (only genes with detectable F_{ST}; negative values were removed). Dot sizes were proportional to SNV numbers in individual genes, and F_{ST} genome-wide mean was depicted in each chart with dash lines. Red-colored dots represent high F_{ST} genes that also passed the requirements of F_{ST}, nucleotide diversity, N/S SNV ratios, and coverages (see methods). The nucleotide diversity and N/S SNV ratio distribution for high F_{ST} genes and genome-wide mean of all genes in different environments were depicted in the chart on the right side. Details of high F_{ST} genes and related parameters in individual genomes (all hits, also including these two representative genomes) were listed in Supplementary Data 12. **c** Two representative charts showing F_{ST} distribution in MAGs that contain sulfur metabolizing genes with signals of being fixed. In each chart, the x-axis represents gene numbers (only genes with detectable F_{ST}; negative values were removed). Dot sizes were proportional to SNV numbers in individual genes, and F_{ST} genome-wide mean was depicted in each chart with dash lines. Red-colored dots represent sulfur metabolizing genes that passed the requirements of F_{ST}, nucleotide diversity, N/S SNV ratios, and coverages (see methods). The nucleotide diversity and N/S SNV ratio distribution for sulfur metabolizing genes in different environments were depicted in the chart on the right side. Details of sulfur metabolizing genes with signals of being fixed and related parameters in individual genomes (all hits, also including these two representative genomes) were listed in Supplementary Data 13. **d** Frame diagram showing the underlying evolutionary processes during migration. Circles represent microbial populations. Dash line arrows indicate the direction of the next evolutionary step.

306

Discussion

In this study, we observed that distinctive plume geochemistry influences the energy landscape across three different hydrothermal vent sites[4, 29]. Sulfur oxidation is the major energy-yielding reaction, while different sites are also represented by different energy landscapes influenced by differing vent geochemistry. For instance, other important energy sources like methane and hydrogen also have important roles in the energy landscape of hydrothermal plumes. The existence of a core plume microbiome indicates that a general biogeochemical feature – energy and substrate supply – within hydrothermal plumes supports the growth of these globally dispersed cosmopolitan microorganisms. As a consequence, the core plume microbiome is likely a result of the sulfur oxidation-based energy landscapes shared among many hydrothermal plumes around the globe. The increased taxa abundance and higher growth rates of major energy contributing taxa supports the interpretation that microbiomes act in response to geochemically-influenced energy landscapes with some taxa fueled by plume substrates. These analyses support the theory of an ocean seed bank origin of the hydrothermal plume microbiome[47]. Plume geochemistry defines the energy availability, serving as a key control on the microbiome distribution and abundance[2, 9]. The consistent taxonomy-function-geochemistry links demonstrated by us suggest that omics-based profiling that reflects the full genetic repertoire of plume microorganisms can be a powerful tool to unravel the relationship between environments and microbiomes.

Characterization of sulfur metabolism in plumes reveals that though all plumes have sulfur oxidation as the reaction with the highest MW-score, and sulfur-oxidizing genes were highly expressed, the major populations contributing to these processes (*dsrAB* and *sdo* containing MAGs) vary in different hydrothermal vent sites. This indicates the variable composition of core sulfur oxidizers in individual environments which suggests the endemicity of microbial community structure. Core sulfur oxidizers can be derived from the pelagic ocean through stochastic processes that can be influenced by dormancy capacity to provide resilient seed microbes, ocean currents to overcome dispersal limitations, and adaptive strategies to nutrient and temperature fluctuations[2]. Core members of the plume microbiome derived in this manner likely thrive under favorable geochemical conditions[48]. For example, Pseudomonadales, Thiomicrospirales, and SAR324 are members of the core plume microbiome, but are also known to be abundant cosmopolitan bacteria in the pelagic oceans. These microorganisms can be distributed as seed banks in the global oceans, triggered by plume sulfur substrates, and subsequently become active sulfur oxidizers in hydrothermal plumes[9, 48]. Sulfur oxidizers within the community have metabolic plasticity to connect other energy transformation activities, e.g., small carbon substrate utilization, nitrate/nitrite reduction, and iron/manganese/arsenite oxidation, etc. This indicates that sulfur and other energy sources can be simultaneously utilized for energy conservation by sulfur oxidizers even in various plume environments with different energy landscapes. At the same time, as described in the network complexity analysis, though sulfur oxidation dominates in energy generation, other reactions are also important components in the metabolic network connected to sulfur oxidation, and cohesively contribute to the energy landscape.

Finally, the microdiversity patterns observed in plume microorganisms depict a scheme of populations selected by environmental constraints. Low population diversity and high N/S SNV ratio indicate that

349 microbes are selected by plume conditions and actively grow after a short migration history. Evidence
350 shows that gene-specific sweeps within certain plume populations are related to nutrient uptake,
351 aerobic oxidation, and stress responses, and some sulfur metabolizing genes are also selected during
352 the environmental change. These traits help microbial cells to be more adaptable and resilient in sulfur
353 oxidation-dominated hydrothermal plume conditions. Collectively, the plume microbiome has a
354 distinctive composition, function, and genetic structure focused on allowing organisms to better adapt
355 to hydrothermal plume conditions. Population alteration in plumes compared to the background deep
356 sea involves both reshaping community-level structure and fine-scale strain-level genetic adjustments
357 that includes advantageous metabolisms being fixed. These nuanced microdiversity changes can lead
358 to fundamental changes in population fitness towards niche adaptation. Overall, the plume microbial
359 community is associated with energy conservation, metabolic distribution, and cell stress response
360 which likely facilitates more efficient adaptation of the plume microbiome in mediating
361 biogeochemical cycles. The connected relationship between microbiome and biogeochemistry we
362 demonstrate reflects the overall ecological and evolutionary basis of microbial strategies for thriving
363 in geochemically-rich energy landscapes.
364

Data availability

366 The MAG genomic sequences are deposited into the NCBI Genome database under the BioProject ID
367 of PRJNA488180. The genome annotation results from this study are publicly available at
368 https://doi.org/10.5281/zenodo.5034800 (all plume MAG annotations are deposited to this location).
369

Code availability

371 The Perl and R codes for parsing, calculating, and visualizing in this study are publicly available at
372 https://github.com/AnantharamanLab/Hydrothermal_plume_omics_Zhou_et_al._2021.
373

Author contributions

384 Z.Z. and K.A. conceived the project. K.A. designed the framework of this project. Z.Z. led the research
385 and conducted the data analysis. Z.Z. and K.A. wrote the manuscript. J.A.B. and G.J.D. sampled in the
386 cruises and J.A.B. contributed to the thermodynamic modeling. K.K., P.Q.T, and A.M.A. helped on
387 data analysis and/or visualization. All authors (Z.Z., P.Q.T., A.M.A., K.K., J.A.B., R.K.S., K.P.K.,

388   P.J.K., C.S.F., C.S.S., J.A.H., M.L., G.J.D., and K.A.) reviewed the results, revised, and approved the
389   manuscript.

Competing interests

391   The authors declare that they have no competing interests.

Additional information

393   Supplementary Information is available for this paper on XXX website.

394

395   **Reprints and permissions information** is available at www.nature.com/reprints
396   **Correspondence and requests for materials** should be addressed to K.A.

397

Methods

Sample information and omics sequencing

400   Hydrothermal plume and surrounding background samples were collected from the corresponding
401   cruises: R/V New Horizon sampling in Guaymas Basin, Gulf of California (July 2004), R/V Atlantis
402   and R/V Falkor sampling in Mid-Cayman Rise, Caribbean Sea (Jan 2012 and June 2013), two
403   consecutive cruises on the R/V Thomas G Thompson sampling in Eastern Lau Spreading Center
404   (ELSC), Lau Basin, western Pacific Ocean (May-July 2009), R/V MGS Sagar sampling in Central
405   Indian Ridge and Southwest Indian Ridge (Jan-Mar 2017), and R/V Thomas G Thompson sampling in
406   Axial Seamount, Juan de Fuca Ridge, northeastern Pacific Ocean (Aug 2015). In brief, Guaymas Basin
407   plume and background samples were collected by 10 L CTD-Rosette bottles and N_2-pressure filtered
408   on board for microbial specimen collection by 0.2 μm pore size, 142 mm polycarbonate membranes[11].
409   The samples were preserved immediately in RNAlater. Mid-Cayman hydrothermal plume and
410   surrounding background samples were collected by Suspended Particulate Rosette (SUPR) filtration
411   device[49] mounted to the remotely operated vehicle Jason II. SUPR collected water with the amount of
412   10-60 L from different sampling locations, and these samples were *in-situ* filtered for microbial
413   specimens by 0.2 μm pore size SUPOR polyethersulfone membranes and preserved in RNAlater
414   flooded conical vials and frozen at -80°C. For Lau Basin samples, SUPR-collected samples were *in-*
415   *situ* filtered by SUPOR polyethersulfone membranes with 0.8 μm and 0.2/0.8 μm pore size for
416   geochemical analysis and microbial specimen collection, respectively[26]. Samples were preserved in
417   RNAlater flooded conical vials and frozen at -80°C. Central Indian Ridge and Southwest Indian Ridge
418   samples were collected by 10 L CTD-Rosette bottles and filtered by 0.2 μm pore size, 47 mm SUPOR
419   polyethersulfone membranes, and preserved in RNAlater flooded conical vials and frozen at -80°C.
420   For Axial Seamount samples, both plume and background samples were collected by a Seabird
421   SBE911 CTD and 10L Niskin bottles[50]. Samples of 3 L were then transferred into cubitainers, filtered
422   through 0.22 μm Sterivex filters, and preserved for downstream analysis[50].

423

424   Details for sample collection, preservation, geochemical analysis, and
425   metagenomic/metatranscriptomic sequencing refer to the previous publications[22, 50, 51]. Detailed cruises
426   and sampling information refer to Supplementary Data 1. The geological map and schematic diagram
427   represent the details of sampling locations (Fig. 1a, Fig. S1). The metagenomic DNA and

428   metatranscriptomic cDNA were extracted and synthesized from corresponding samples and processed
429   for Illumina HiSeq 2000/2500 sequencing as described previously[11, 14, 18, 25, 50]. The distribution of
430   acquired metagenomes (DNAs, labeled as "D") and metatranscriptomes (cDNA, labeled as "C") was
431   represented in Fig. S1b (only for samples with detailed location and physicochemical characterization;
432   distribution of other samples refers to Supplementary Data 1). The raw reads (both DNA/cDNA reads)
433   were dereplicated by SeqTools v4.28 (https://www.sanger.ac.uk/tool/seqtools/) and processed by
434   Sickle v1.33 (https://github.com/najoshi/sickle) to trim reads of low quality with default settings.
435   Command   "reformat.sh"   in   BBTools   (last   modified   on   Feb   11,   2019;
436   https://www.sourceforge.net/projects/bbmap/) was used to calculate fastq sequence and nucleotide
437   numbers.
438

439   **Core hydrothermal plume microbiome analysis**
440   In total, 51 hydrothermal plume and background 16S rRNA gene datasets were used for analyzing the
441   microbiome of hydrothermal plume, within which 24 datasets were obtained in this study, containing
442   datasets from samples of Mid-Cayman Rise, Guaymas Basin, Lau Basin, Central and Southwestern
443   Indian Ridge, and Axial Seamount plume (Supplementary Data 2). For hydrothermal plume and
444   background samples with only metagenome datasets, 16S rRNA gene sequences were parsed out from
445   metagenomes and these sequences were weighted according to their coverages. Simulated 16S rRNA
446   gene datasets were used in subsequent analysis. The original datasets of paired-end reads were merged
447   into combined 16S tags by FLASH v1.2.11[52] with default settings. The bioinformatic analyses,
448   including pre-analysis quality control, 16S chimera checking, open-reference OTU picking, taxonomy
449   assigning, OTU table file 'biom' generating and rarefying, OTU representative sequences filtering and
450   aligning, alignment filtering, and phylogenetic tree reconstructing, were performed according to the
451   instructions of QIIME v1.9.1[53], respectively. The 16S rRNA reference database was based on
452   "SILVA_132_QIIME_release"[54]. The resulted 'biom' (OTU table file), 'tre' (phylogenetic tree), and
453   "map" (sample characterization map) files were imported into R (using R package '*phyloseq*') for
454   downstream analysis and visualization. Taxa summary and principal coordinates analysis (PCoA) were
455   conducted accordingly to delineate the community structure and biogeographic pattern of
456   hydrothermal plume and background seawater microbiome. Genus-level taxa summary table was used
457   to find core hydrothermal plume microbiome from 37 hydrothermal plume datasets by filtering genera
458   that exist in > 67% plume datasets and have > 1% relative abundance on average. Core plume
459   microbiome metabolic profiles were conducted by choosing MAGs (see the following sections of
460   obtaining these MAGs) from this study that contain 16S rRNA genes affiliated to the core plume
461   microbial genera. Metabolic profiling for these MAGs was based on the result from "MAGs genomic
462   property and annotation".
463

464   **Assembling and metagenomic binning**
465   QC-processed reads were assembled *de novo* by MEGAHIT v1.1.2[55] with settings as "--k-min 45 --k-
466   max 95 --k-step 10". Hydrothermal plume and background metagenomes from the same hydrothermal
467   site were assembled together. QC-processed reads were re-mapped to assemblies by Bowtie 2 v2.2.8[56]
468   with default settings. For each hydrothermal site, hydrothermal plume and background reads were
469   mapped to corresponding assemblies separately; bam files by plume and background samples for

470 individual assemblies were used for downstream binning, Subsequently, the assemblies were subjected
471 to a MetaBAT v0.32.4[57] based binning with 12 combinations of parameters. Afterward, DAS Tool
472 v1.0[58] was applied to screen MetaBAT MAGs, resulting in high quality and completeness MAGs. This
473 MetaBAT/DAS Tool method enables a comprehensive "slice-layer profiling" for searching potential
474 MAGs with a better outcome (in-house tested). CheckM v1.0.7[59] was used to assess MAG quality and
475 phylogeny. Outlier scaffolds with abnormal coverage, tetranucleotide signals, and GC pattern within
476 potential high contamination MAGs (by CheckM) and erroneous SSU sequences within MAGs were
477 screened out and decontaminated by RefineM v0.0.20[60] with default settings. Afterwards, further
478 MAG refinement for decontaminating certain MAGs was manually inspected based on VizBin[61].
479 MAGs are picked using a threshold of < 10% contamination (namely genome redundancy) and > 50%
480 completeness.
481

MAG genomic property and annotation
483 Genome phylogeny was determined by RefineM and GTDB-Tk v 0.2.1[62] (GTDB database, release 83).
484 Additionally, phylogenies of those genomes that could not be assigned to a meaningful microbial group
485 were inferred from ribosomal protein (RP) trees using the phylogenetic reconstruction method
486 described below. Genomic properties, including genome coverage, genome and 16S rRNA taxonomy,
487 tRNAs, genome completeness, and scaffold parameters, were parsed from results that were calculated
488 by CheckM and tRNAscan-SE 2.0[63]. Relative genome coverages were normalized by setting each
489 metagenomic dataset size as 100M paired-end reads. MAG ORFs were parsed out by the Prokka
490 annotation pipeline v1.12[64] with default settings. For ORF annotation, GhostKOALA v2.0[27], KAAS
491 v2.1[26,] and eggNOG-mapper v4.5.1[28] were applied to thoroughly annotate ORFs to KOs. For eggNOG-
492 mapper annotation, we used its first KO hit as annotation result; if there was only COG annotation, we
493 translated it into KO using 'ko2cog.xl' provided by KEGG database. When combining three software
494 annotations, we use resulted KO from the first software as the final annotation; if there is no annotation
495 from the first software, then we will move to the next software accordingly. Annotation by NCBI nr
496 database (Mar 6, 2017 updated) was conducted with default settings and for each annotation the first
497 meaningful hit (hit not assigned as 'hypothetical protein') was extracted. Genomic-specific metabolic
498 traits were searched against TIGRfam, Pfam, Kofam, and custom HMM profiles using hmmscan[65] and
499 custom protein database using DIAMOND BLASTP[66]. For searching against custom HMM databases,
500 noise cutoff values are determined according to previous settings[12], respectively. For DIAMOND
501 BLASTP searching, a stringent criterion as "-e 1e-20 --query-cover 65 --id 65" was applied.
502 Carbohydrate active enzymes (CAZymes) were searched against dbCAN2 with default settings[67];
503 Peptidases were searched against MEROPS 'pepunit' database with stringent DIAMOND BLASTP
504 settings as "-e 1e-10 --subject-cover 80 --id 50"[68].
505

Phylogenetic tree reconstruction
507 The syntenic block of universal 16 ribosomal proteins (RPs) (L2-L6, L14-L16, L18, L22, L24, S3, S8,
508 S10, S17, and S19) were used for inferring RP phylogenetic tree, after hmmscan-based[65] searching for
509 RPs from all MAGs. The individual RP was pre-aligned with local custom RP database by MAFFT
510 v7.123b[69] and curated in Geneious Prime v2019.0.4[70] by manually masking out begin and end regions
511 with lots of gaps. Out of 206 MAGs, 177 containing > 4 RPs were used; the concatenated and curated

512 16RP-alignment (7741 aligned columns) was used for phylogenetic inference by IQTREE-based
513 maximum likelihood method (IQ-TREE multicore v1.6.3[71]) with settings of "-m MFP -bb 1000 -redo
514 -mset WAG,LG,JTT,Dayhoff -mrate E,I,G,I+G -mfreq FU -wbtl". The resulted phylogenetic tree was
515 rooted by archaea lineages and visualized by iTOL[72]. Functional traits were added accordingly to each
516 MAG on the tree. Bacterial and archaeal SSU sequences (> 300 bp and the longest from individual
517 MAG) parsed out by local pipeline (use CheckM ssuFinder[59] to pick and RefineM to filter erroneous
518 hits) were aligned in SINA aligner[73] with default settings. The 16S sequence taxonomy was checked
519 by BLASTn searching against SILVA_128_SSUParc_tax_silva database[54] and 16S sequences with
520 resulted taxonomy different from their MAG phylogeny (at the phylum level) were filtered due to the
521 high possibility of contamination. IQTREE-based[71] phylogenetic inference was conducted with
522 settings of "-st DNA -m MFP -bb 1000 -alrt 1000". The 16S rRNA gene tree based on the alignment
523 of 85 sequences with 50000 columns was rooted by archaea lineages, visualized by iTOL[72], and
524 manually curated.
525

Metagenomic and metatranscriptomic mapping
527 QC-passed metagenomic reads were mapped to MAGs separately (metagenomic datasets from
528 Guaymas Basin, Mid-Cayman Rise, and Lau Basin sites were mapped individually to the
529 corresponding MAGs) using Bowtie 2 v2.2.8 with default settings[56]. MetaBAT integrated
530 "jgi_summarize_bam_contig_depths" script and homemade Perl scripts were used to calculate MAG
531 coverage (normalized coverage with each metagenomic dataset size set as 100M paired-end reads).
532 QC-passed metatranscriptomic reads (use the same QC-process as described above with an additional
533 SortMeRNA v2.1[74] rRNA filtering step) were mapped to MAGs separately, with TPM (Transcripts Per
534 Kilobase Million) calculated for individual genes within each genome.
535

Statistical comparison on MAG and functional trait abundance
537 Metagenome/metatranscriptome-based MAG mapping results and functional annotations for all the
538 MAGs were summarized individually. Afterwards, significance tests on the differentiation pattern of
539 MAG (also MAG taxonomic group) and functional trait abundances across all the
540 metagenomic/metatranscriptomic samples were calculated by R package DESeq2[75]. Log2 Fold
541 Change value with adjusted P-value (by nbinomWaldTest) < 0.05 was considered as significant.
542 Relative abundances of MAG (also MAG taxonomic group) and functional traits were visualized by
543 R (using R package '*pheatmap*') with the relative abundance at row normalized by removing the mean
544 (centering) and dividing by the standard deviation (scaling). Sunburst figures were generated to depict
545 the relative abundance of MAGs based on metagenomic/metatranscriptomic mapping results, with the
546 significant Log2 Fold Change values labeled to individual MAGs that have differential abundances
547 between different hydrothermal ecological niches, e.g., plume and background.
548

549 To find taxa in microbial community that are responsible for enriched functions (functions that are
550 significantly enriched in each environment), major functions (including functions that are in the
551 categories of carbon fixation, denitrification, sulfur cycling, hydrogen oxidation, methane oxidation,
552 aerobic oxidation, iron oxidation, and manganese oxidation), and specific functions, custom Perl
553 scripts were written to get the corresponding microbial community contribution information (scripts

554   deposited in https://github.com/AnantharamanLab/Hydrothermal_plume_omics_Zhou_et_al._2021).
555   Functional trait results of all MAGs, MAG coverage within the community, and targeted function list
556   were used as inputs to conduct the calculation. For environments with metatranscriptomic reads, we
557   also used active MAG coverage (calculated by metatranscriptomic reads mapping result) as the input
558   to calculate microbial community contribution information based on metatranscriptomes.
559

560   **Bioenergetic and thermodynamic modeling**
561   Equilibrium thermodynamic reaction path modeling was used to predict chemical concentrations and
562   activity coefficients resulting from the mixing of seawater with end-member vent fluids
563   (Supplementary Table 2). Our thermodynamic modeling builds on the specific plume model
564   implementation described in Breier et al[76]. The estimated temperature of bottom seawater is according
565   to the previous report[10]. The original chemical data is derived from Reeves et al[77] and Anantharaman
566   et al[10]. For each hydrothermal vent system, we choose at least one representative end-member fluid
567   sample(s), respectively (1 for Guaymas Basin, 2 for Mid-Cayman Rise, and 3 for Lau Basin)
568   (Supplementary Table 2).
569

570   Bioenergetic and thermodynamic modeling procedures were conducted as described in Anantharaman
571   et al[7] and Li et al[18] (More details refer to Supplementary Information and Tables). Reaction path
572   modeling was performed with REACT, part of the Geochemist's Workbench package[78]. Conductive
573   cooling was neglected and mixture temperatures were a strict function of conservative end-member
574   mixing. Precipitated minerals were allowed to dissolve and their constituents to re-precipitate based
575   on thermodynamic equilibrium constraints. Thermodynamic data were predicted by SUPCRT95[79] for
576   the temperature range of 2°C to end-member vent fluid temperature and a pressure of 500 bar. The
577   estimated biomasses and free energies of individual environments were calculated and their relative
578   abundance change along the temperature range (2 - 121°C) was visualized by R. Two temperatures (3
579   and 4.9°C) were picked to conduct the biomass and free energy estimation for representing typical
580   plume temperatures in nature.
581

582   **Energy contribution and MAG growth rate calculation**
583   Based on metabolic prediction of each MAG and MAG gene coverage and expression level within
584   each environment, energy contribution for each electron donor was calculated based on gene
585   coverage/expression level and free energy of each catabolic reaction. The contribution ratio of electron
586   donor species was calculated for individual samples respectively. We also included influence of the
587   presence of electron acceptors to energy contribution calculation. To simplify the hydrothermal
588   condition, we only included two major electron acceptors (O_2 and NO_3^-) and used the ratio of these
589   two electron acceptors to infer energy contribution of electron donors at different oxidative conditions.
590

591   Microbial genome replication starts directionally from a single origin[31]. Based on metagenomic
592   mapping, at a single time-point the coverage ratio between the replicating origin and terminus of a
593   microbial genome can be used as a proxy to represent the replication rate/growth rate[30, 32]. Growth rate
594   for each MAG was calculated by iRep v1.10[30] with default settings. MAGs that are from the same
595   environments were pooled together as the input genomes. Sam files that were generated by

596   metagenomic mapping described above were used as the iRep input. Barcharts that reflect the growth
597   rate and significant difference test result (by *t*-test) of MAG taxonomic groups were generated using
598   R package '*ggplot2*' and '*PairedData*'.

599

Network complexity analysis

600
601   For each community, a bipartite network was built based on reaction/substrate relationships and the
602   percent energy yields for each reaction. Briefly, the plume chemical reaction table for each reaction
603   was stored; within the table, the substrate and product for a reaction were recorded[39]. Then, for each
604   community, reactions (represented as one set of nodes in the bipartite network) with different percent
605   energy yields were connected with substrates and products in the network (represented as the second
606   set of nodes) via directed edges between both sets of nodes. The energy yields are based on the result
607   from "Bioenergetic and thermodynamic modeling" and are represented on the network as node size
608   proportional to the percent energy yield. These networks were constructed using the Python package
609   '*networkx*'[80] (https://networkx.org/).

610

611   The network complexity change as a function of reaction energy yield was calculated by the following
612   steps[40]. For each plume community network, the complexity of the network's structure was measured.
613   A node was taken from the network; as a consequence, the change in complexity (ΔC) before and after
614   the node was taken was calculated accordingly. The ΔC was assigned to that node as a property
615   representing that node's contribution to the network's overall complexity. Then this node was placed
616   back and these steps were repeated for each reaction node[40].

617

618   In this study, complexity (C) was calculated by estimating the algorithmic complexity. Because
619   algorithmic complexity cannot be directly computed, we used an estimate known as the Block
620   Decomposition Method (BDM)[41]. The perturbation analysis to calculate each node's complexity
621   contribution (ΔC) is called Minimal Information Loss Selection, MILS[32]; in this study, successive edge
622   deletion was replaced as node deletion which also works with good performance[33]. This method has
623   been used to characterize complex properties of biological networks and is proven to be a good
624   measure among many other algorithms[40, 41]. For all reaction nodes in each community plume reaction
625   network, we conducted this measurement for each reaction node and came up with the scatterplots.

626

Community-level metabolic analysis

627
628   Resulted MAGs and plume metagenomic reads were used to conduct community-level metabolic
629   analysis using METABOLIC-C v4.0[35] with default settings. For Guaymas Basin, Mid-Cayman Rise,
630   and Lau Basin sites, all MAGs and plume metagenomic reads from each site were used separately.
631   From METABOLIC-C regular MW-score results, a group of metabolic cycling steps that are important
632   in reflecting the plume substrate metabolisms were specifically selected to make functional network
633   diagrams (using R script '*draw_functional_network.R*' from METABOLIC-C). For each site, MW-
634   score table and functional network diagram (based on both all and selected metabolic steps) were
635   generated, respectively.

636

Evolution analysis

Metagenomic reads from mesopelagic *Tara* Ocean metagenomic datasets (with > 800m depth)[45] were used as the regular ocean environment representatives to compare microdiversity characteristics with that of hydrothermal environments from this study. To simplify analyses, *Tara* Ocean reads from samples collected by filtration with various filter sizes at each station were pooled as one to represent all reads from that station. Both *Tara* Ocean reads and hydrothermal environment reads (including both background and plume environments; background and plume reads were also pooled together individually to simplify analyses and satisfy coverage requirement of each MAG) from this study were first mapped to hydrothermal environment MAGs recovered from individual sites by Bowtie 2[56] with default settings. After mapping, reads within resulted bam files were filtered according to the following rules to calculate downstream microdiversity parameters: (1) minimum percent identity of read pairs to reference > 95%; (2) maximum insert size between two reads < 3× median insert size and minimum insert size > 50bp (so only paired reads are retained). Filter steps were either conducted by inStrain v1.4.1[42] or inStrain_lite v0.4.0[81] (for generating bam files) with the same rules. Software inStrain was further employed to calculate microdiversity parameters for each MAG in individual sites from this study. Subsequently, interested parameters[42] were picked and parsed accordingly from resulted folders, including 'coverage' (average coverage depth of all scaffolds of one genome), 'breadth minCov' (percentage of bases in the scaffold that have at least 'min_cov' coverage), 'SNV count / (breadth minCov × length)' (total number of SNVs called on one genome normalized by genome length and breadth minCov), 'N/S SNV ratio' (nonsynonymous to synonymous SNV ratio of one genome), 'r2_mean' (r^2 mean between linked SNVs), 'con freq mean' (mean value of fraction of reads supporting the consensus base within one genome), 'con freq mean for N SNV' (mean value of con freq on all nonsynonymous SNV sites), and 'con freq mean for S SNV' (mean value of con freq on all synonymous SNV sites). MAGs that have breadth_minCov value < 50% or do not pass the 'min_cov' requirement by inStrain were removed from microdiversity analysis in each site.

In order to identify gene-specific selective sweep in hydrothermal environment, we further pooled reads together into two categories, one contains hydrothermal environment datasets (including both background and plume environment datasets) and the other contains *Tara* Ocean samples (all *Tara* Ocean sample datasets are pooled together). After reads mapping and filtering as described above, F_{ST} (fixation index) between hydrothermal and *Tara* Ocean environments was calculated using scikit-allel package[82] (Hudson method[83]) within inStrain_lite to identify genes with skewed allele frequencies across the whole genome. Subsequently, high F_{ST} genes from each MAG within each hydrothermal vent site were identified if they have F_{ST} value > F_{ST} mean (genome-wide F_{ST} average) + 2.5 × F_{ST} std (genome-wide F_{ST} standard deviation) and the lowest gene coverage in either hydrothermal and *Tara* Ocean environment samples should be higher than 5×. Meanwhile, for each genome the number of genes with empty F_{ST} value should not be more than half of all genes, otherwise high F_{ST} genes will not be taken into account for this genome. We set gene coverage in both environments to be at least 5× due to the fact that reduction of gene coverage (or loss of coverage in some genome regions) can also lead to low nucleotide diversity. Furthermore, to confirm that these genes are specifically selected in hydrothermal environment, additional requirements were added: (1) gene nucleotide diversity in hydrothermal environment < nucleotide diversity genome average in hydrothermal environment; (2)

679   gene N/S SNV ratio in hydrothermal environment > N/S SNV ratio genome average in hydrothermal
680   environment; (3) gene nucleotide diversity in hydrothermal environment < gene nucleotide diversity
681   in *Tara* Ocean samples; (4) gene N/S SNV ratio in hydrothermal environment > gene N/S SNV ratio
682   in *Tara* Ocean samples.

683

684   To find sulfur metabolizing genes that have signals of being fixed after migration, a relatively less
685   stringent set of criteria were used to screen gene F_{ST} values compared to high F_{ST} gene identification
686   method in the above paragraph. For each sulfur metabolizing gene (including genes of *sat*, *aprA*, *sdo*,
687   oxidative *dsrAB*, and *soxBCY*) containing MAGs, the identified genes should meet the following
688   criteria: (1) F_{ST} value > F_{ST} mean (genome-wide F_{ST} average) and both F_{ST} and F_{ST} mean should be
689   positive values; (2) gene nucleotide diversity in hydrothermal environment < gene nucleotide diversity
690   in *Tara* Ocean samples; (3) gene N/S SNV ratio in hydrothermal environment > gene N/S SNV ratio
691   in *Tara* Ocean samples; (4) gene coverages in hydrothermal environments and *Tara* Ocean samples
692   both > 5×. Sulfur metabolizing genes that meet all the four criteria were indicated to have positive
693   gene fixation signals though the selective power across the genome did not reach the level of gene-
694   specific selective sweeps as indicated by the above method.

Figure Captions

Fig. 1 | Sampling sites, distribution, and metabolic profile of the core plume microbiome. a Sampling site maps of hydrothermal plume samples from which the 16S rRNA gene datasets were sourced. Numbers in brackets indicate dataset quantities. Three hydrothermal sites that have metagenome and metatranscriptome datasets in this study were specifically represented by inset maps. Ocean maps were remodified from ArcGIS online map (containing layers of "World Ocean Base" and "World Ocean Reference"; https://www.arcgis.com/). **b** Membership and distribution of the core plume microbiome. Heatmap shows the presence/absence of core plume microbial groups (tracing back to known taxonomic ranks from the genus-level taxa) in 37 hydrothermal plume 16S rRNA gene datasets across the world. **c** Metabolic profile of the core plume microbiome. From this study, MAGs that have 16S rRNA genes affiliated to the core plume microbiome were used as representatives (numbers labeled in brackets). This subpanel shows the presence or absence of metabolic potential associated with sulfur, carbon, nitrogen, hydrogen, and metal biogeochemical transformations.

Fig. 2 | Thermodynamic estimation of available free energies and biomass yields from electron donors, metagenomics-based contribution of electron donors to energy, and growth rates of major microbial contributors. a Thermodynamic estimation diagram of available free energy and biomass. For each hydrothermal environment, the contribution fraction of each electron donor species was labeled accordingly in the rings. The total available free energies and biomass were labeled accordingly to individual plumes. Two temperatures (3°C and 4.9°C) were picked to represent *in situ* temperatures in the upper and lower plume. Light yellow represents aerobic sulfur oxidation, dark yellow represents anaerobic sulfur oxidation. **b** Metagenomics-based estimation of energy contribution. Energy contribution for each electron donor was calculated based on metagenomic abundance of each reaction of electron donors and free energy yield of each reaction. The contribution ratio of electron donor species was calculated for individual environments respectively. **c** Growth rate of major microbial contributors in each hydrothermal environment. The y-axis for each barplot indicates the replication rate. The microbial groups starting with "α-", "γ-", and "δ-" represent Alphaproteobacteria, Gammaproteobacteria, and Deltaproteobacteria, respectively. Plume microbial groups were colored by dark yellow, background microbial groups were colored by light yellow and they were also all labeled with "(*P*)" or "(*B*)", respectively. Numbers in brackets indicate MAG numbers in each microbial group. Star-labeled plume microbial groups had higher growth rates than the 'Rest' plume microbial groups.

Fig. 3 | Sulfur metabolism and metabolic plasticity of sulfur oxidizers. a Details of sulfur metabolism pathways in the hydrothermal plume. The gene abundance (coverage normalized by 100M reads) and transcript expression level (TPM) for each step were calculated based on plume metagenomic and metatranscriptomic read mapping results. Log$_{10}$-transformed values of gene abundance and transcript expression level were labeled accordingly in the diagram. **b** Major contributors to sulfur metabolizing genes. For each sulfur metabolizing gene, microbial groups that occupied > 10% of the total gene abundance (by metagenome) or transcript expression (by metatranscriptome) values were labeled in the diagram. For some genes with only three or less than three contributors, all contributors were labeled. **c** Metabolic plasticity of sulfur oxidizers. For each hydrothermal vent site, three parameters were given to show the metabolic plasticity of sulfur oxidizers in conducting each electron transferring reaction related to carbon, nitrogen, hydrogen, and metal

738 biogeochemical cyclings: the number of sulfur-oxidizing gene containing MAGs, gene abundance
739 percentage, and transcript abundance percentage.
740
741 **Fig. 4 | Network complexity, MW-scores (metabolic weight scores), and functional network**
742 **diagrams of the three hydrothermal vent sites. a** Network complexity diagram representing each
743 reaction's influence on the complexity of the network. In the figure, different colors represent different
744 hydrothermal environments, different symbol shapes represent different reactions. The substrates
745 (including electron donors and acceptors) were listed for each reaction in the legend. The x-axis is the
746 change in complexity (ΔC) of the whole network for a node (a reaction here) and the y-axis is the
747 percent energy yield of that reaction in the whole community. This network complexity diagram was
748 based on thermodynamic estimation results at 3°C. **b** MW-scores of three major energy contributing
749 reactions. **c** Functional network diagram of Guaymas Basin. **d** Functional network diagram of Mid-
750 Cayman Rise. **e** Functional network diagram of Lau Basin. A group of metabolic cycling steps that
751 are important in reflecting the plume substrate metabolisms were selected from METABOLIC-C
752 regular MW-score results to make these functional network diagrams (**c, d, e**), respectively. In each
753 functional network diagram, the size of a node is proportional to gene coverage associated with the
754 metabolic/biogeochemical cycling step. The thickness of the edge represents the average gene
755 coverage values of the two connected metabolic/biogeochemical cycling steps. Edges related to two
756 reactions of sulfur oxidation were colored accordingly in each diagram.
757
758 **Fig. 5 | Evolutionary mechanism of plume microbial populations during migration. a** Schematic
759 diagram showing the changing trend of microdiversity parameters during migration. Individual solid
760 dots with various colors represent microbial populations. Two scenarios were depicted in this panel:
761 unrepresentative strains and strains that have detectable read mapping results in both environments. **b**
762 Two representative charts showing F_{ST} distribution in MAGs that contain high F_{ST} genes. In each chart,
763 the x-axis represents gene numbers (only genes with detectable F_{ST}; negative values were removed).
764 Dot sizes were proportional to SNV numbers in individual genes, and F_{ST} genome-wide mean was
765 depicted in each chart with dash lines. Red-colored dots represent high F_{ST} genes that also passed the
766 requirements of F_{ST}, nucleotide diversity, N/S SNV ratios, and coverages (see methods). The
767 nucleotide diversity and N/S SNV ratio distribution for high F_{ST} genes and genome-wide mean of all
768 genes in different environments were depicted in the chart on the right side. Details of high F_{ST} genes
769 and related parameters in individual genomes (all hits, also including these two representative genomes)
770 were listed in Supplementary Data 12. **c** Two representative charts showing F_{ST} distribution in MAGs
771 that contain sulfur metabolizing genes with signals of being fixed. In each chart, the x-axis represents
772 gene numbers (only genes with detectable F_{ST}; negative values were removed). Dot sizes were
773 proportional to SNV numbers in individual genes, and F_{ST} genome-wide mean was depicted in each
774 chart with dash lines. Red-colored dots represent sulfur metabolizing genes that passed the
775 requirements of F_{ST}, nucleotide diversity, N/S SNV ratios, and coverages (see methods). The
776 nucleotide diversity and N/S SNV ratio distribution for sulfur metabolizing genes in different
777 environments were depicted in the chart on the right side. Details of sulfur metabolizing genes with
778 signals of being fixed and related parameters in individual genomes (all hits, also including these two
779 representative genomes) were listed in Supplementary Data 13. **d** Frame diagram showing the
780 underlying evolutionary processes during migration. Circles represent microbial populations. Dash
781 line arrows indicate the direction of the next evolutionary step.

782

783

References

1. Dick G, Anantharaman K, Baker B, Li M, Reed D, Sheik C. The microbiology of deep-sea hydrothermal vent plumes: ecological and biogeographic linkages to seafloor and water column habitats. *Front Microbio* **4**, 1-16 (2013).

2. Dick GJ. The microbiomes of deep-sea hydrothermal vents: distributed globally, shaped locally. *Nat Rev Microbiol* **17**, 271-283 (2019).

3. German C, Von Damm K. Hydrothermal processes. *Treatise on geochemistry* **6**, 625 (2003).

4. McCollom TM. Geochemical constraints on primary productivity in submarine hydrothermal vent plumes. *Deep Sea Research Part I: Oceanographic Research Papers* **47**, 85-101 (2000).

5. Winn CD, Karl DM, Massoth GJ. Microorganisms in deep-sea hydrothermal plumes. *Nature* **320**, 744-746 (1986).

6. Zhang Y, Zhao Z, Chen C-TA, Tang K, Su J, Jiao N. Sulfur Metabolizing Microbes Dominate Microbial Communities in Andesite-Hosted Shallow-Sea Hydrothermal Systems. *PLoS One* **7**, e44593 (2012).

7. Anantharaman K, Breier JA, Sheik CS, Dick GJ. Evidence for hydrogen oxidation and metabolic plasticity in widespread deep-sea sulfur-oxidizing bacteria. *Proc Natl Acad Sci U S A* **110**, 330 (2013).

8. Mattes TE*, et al.* Sulfur oxidizers dominate carbon fixation at a biogeochemical hot spot in the dark ocean. *ISME J* **7**, 2349-2360 (2013).

9. Zhou Z*, et al.* Gammaproteobacteria mediating utilization of methyl-, sulfur- and petroleum organic compounds in deep ocean hydrothermal plumes. *ISME J* **14**, 3136-3148 (2020).

10. Anantharaman K, Breier JA, Dick GJ. Metagenomic resolution of microbial functions in deep-sea hydrothermal plumes across the Eastern Lau Spreading Center. *ISME J* **10**, 225 (2015).

11. Lesniewski RA, Jain S, Anantharaman K, Schloss PD, Dick GJ. The metatranscriptome of a deep-sea hydrothermal plume is dominated by water column methanotrophs and lithotrophs. *ISME J* **6**, 2257 (2012).

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

13. Baker BJ, Lazar CS, Teske AP, Dick GJ. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. *Microbiome* **3**, 14 (2015).

14. Li M, Baker BJ, Anantharaman K, Jain S, Breier JA, Dick GJ. Genomic and transcriptomic evidence for scavenging

825   of diverse organic compounds by widespread deep-sea archaea. *Nat Commun* **6**, 8933 (2015).

826

827 15. Dombrowski N, Teske AP, Baker BJ. Expansive microbial metabolic versatility and biodiversity in dynamic

828   Guaymas Basin hydrothermal sediments. *Nat Commun* **9**, 4999 (2018).

829

830 16. Anantharaman K, Duhaime MB, Breier JA, Wendt K, Toner BM, Dick GJ. Sulfur Oxidation Genes in Diverse Deep-

831   Sea Viruses. *Science* **344**, 757-760 (2014).

832

833 17. Sheik CS, Jain S, Dick GJ. Metabolic flexibility of enigmatic SAR324 revealed through metagenomics and

834   metatranscriptomics. *Environ Microbiol* **16**, 304-317 (2013).

835

836 18. Li M, Toner BM, Baker BJ, Breier JA, Sheik CS, Dick GJ. Microbial iron uptake as a mechanism for dispersing iron

837   from deep-sea hydrothermal vents. *Nat Commun* **5**, 3192 (2014).

838

839 19. Li M, Jain S, Dick GJ. Genomic and Transcriptomic Resolution of Organic Matter Utilization Among Deep-Sea

840   Bacteria in Guaymas Basin Hydrothermal Plumes. *Front Microbio* **7**, 1-13 (2016).

841

842 20. Sjöqvist C, Delgado LF, Alneberg J, Andersson AF. Ecologically coherent population structure of uncultivated

843   bacterioplankton. *ISME J* **15**, 3034-3049 (2021).

844

845 21. Larkin AA*, et al.* Subtle biogeochemical regimes in the Indian Ocean revealed by spatial and diel frequency of

846   Prochlorococcus haplotypes. *Limnol Oceanogr* **65**, S220-S232 (2020).

847

848 22. Anderson RE*, et al.* Genomic variation in microbial populations inhabiting the marine subseafloor at deep-sea

849   hydrothermal vents. *Nat Commun* **8**, 1114 (2017).

850

851 23. Galambos D, Anderson RE, Reveillaud J, Huber JA. Genome-resolved metagenomics and metatranscriptomics

852   reveal niche differentiation in functionally redundant microbial communities at deep-sea hydrothermal vents.

853   *Environ Microbiol*,   (2019).

854

855 24. Meier DV*, et al.* Niche partitioning of diverse sulfur-oxidizing bacteria at hydrothermal vents. *ISME J* **11**, 1545

856   (2017).

857

858 25. Dick GJ, Tebo BM. Microbial diversity and biogeochemistry of the Guaymas Basin deep-sea hydrothermal

859   plume. *Environ Microbiol* **12**, 1334-1347 (2010).

860

861 26. Sheik CS, Anantharaman K, Breier JA, Sylvan JB, Edwards KJ, Dick GJ. Spatially resolved sampling reveals

862   dynamic microbial communities in rising hydrothermal plumes across a back-arc basin. *ISME J* **9**, 1434 (2014).

863

864 27. Reed DC*, et al.* Predicting the response of the deep-ocean microbiome to geochemical perturbations by

865   hydrothermal vents. *ISME J* **9**, 1857-1869 (2015).

866

867 28. Flores GE*, et al.* Inter-field variability in the microbial communities of hydrothermal vent deposits from a back-

868   arc basin. *Geobiology* **10**, 333-346 (2012).

29. Mottl MJ*, et al.* Chemistry of hot springs along the Eastern Lau Spreading Center. *Geochim Cosmochim Acta* **75**, 1013-1038 (2011).

30. Brown CT, Olm MR, Thomas BC, Banfield JF. Measurement of bacterial replication rates in microbial communities. *Nat Biotechnol* **34**, 1256 (2016).

31. Prescott DM, Kuempel PL. Bidirectional Replication of the Chromosome in *Escherichia coli*. *Proc Natl Acad Sci U S A* **69**, 2842-2845 (1972).

32. Korem T*, et al.* Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* **349**, 1101-1106 (2015).

33. Breuer C, Pichler T. Arsenic in marine hydrothermal fluids. *Chem Geol* **348**, 2-14 (2013).

34. Simoneit BRT, Kawka OE, Brault M. Origin of gases and condensates in the Guaymas Basin hydrothermal system (Gulf of California). *Chem Geol* **71**, 169-182 (1988).

35. Zhou Z*, et al.* METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome* **10**, 33 (2022).

36. Fru EC*, et al.* Arsenic and high affinity phosphate uptake gene distribution in shallow submarine hydrothermal sediments. *Biogeochemistry* **141**, 41-62 (2018).

37. Mattes TE, Ingalls AE, Burke S, Morris RM. Metabolic flexibility of SUP05 under low DO growth conditions. *Environ Microbiol* **23**, 2823-2833 (2020).

38. Swan BK*, et al.* Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333**, 1296-1300 (2011).

39. Montañez R, Medina MA, Solé RV, Rodríguez-Caso C. When metabolism meets topology: Reconciling metabolite and reaction networks. *Bioessays* **32**, 246-256 (2010).

40. Zenil H, Kiani NA, Abrahão FS, Rueda-Toicen A, Zea AA, Tegnér J. Minimal Algorithmic Information Loss Methods for Dimension Reduction, Feature Selection and Network Sparsification. *arXiv*, arXiv:1802.05843 (2018).

41. Zenil H, Kiani NA, Tegnér J. A Review of Graph and Network Complexity from an Algorithmic Information Perspective. *Entropy* **20**, 551 (2018).

42. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat Biotechnol* **39**, 727-736 (2021).

913   43.   Simmons SL, DiBartolo G, Denef VJ, Goltsman DSA, Thelen MP, Banfield JF. Population Genomic Analysis of
914   Strain Variation in Leptospirillum Group II Bacteria Involved in Acid Mine Drainage Formation. *PLoS Biol* **6**, e177
915   (2008).
916

917   44.   Eppley JM, Tyson GW, Getz WM, Banfield JF. Genetic Exchange Across a Species Boundary in the Archaeal
918   Genus *Ferroplasma*. *Genetics* **177**, 407 (2007).
919

920   45.   Pesant S*, et al.* Open science resources for the discovery and analysis of Tara Oceans data. *Sci Data* **2**, 150023
921   (2015).
922

923   46.   Meirmans PG, Hedrick PW. Assessing population structure: F_{ST} and related measures. *Mol Ecol Resour* **11**, 5-18
924   (2011).
925

926   47.   Gonnella G*, et al.* Endemic hydrothermal vent species identified in the open ocean seed bank. *Nat Microbiol* **1**,
927   16086 (2016).
928

929   48.   Lennon JT, Jones SE. Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nat Rev
930   Microbiol* **9**, 119-130 (2011).
931

932   49.   Breier JA*, et al.* A suspended-particle rosette multi-sampler for discrete biogeochemical sampling in low-
933   particle-density waters. *Deep Sea Research Part I* **56**, 1579-1589 (2009).
934

935   50.   Fortunato CS, Larson B, Butterfield DA, Huber JA. Spatially distinct, temporally stable microbial populations
936   mediate biogeochemical cycling at and below the seafloor in hydrothermal vent fluids. *Environ Microbiol* **20**,
937   769-784 (2017).
938

939   51.   Reveillaud J*, et al.* Subseafloor microbial communities in hydrogen-rich vent fluids from hydrothermal systems
940   along the Mid-Cayman Rise. *Environ Microbiol* **18**, 1970-1987 (2016).
941

942   52.   Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies.
943   *Bioinformatics* **27**, 2957-2963 (2011).
944

945   53.   Caporaso JG*, et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**, 335-
946   336 (2010).
947

948   54.   Quast C*, et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.
949   *Nucleic Acids Res* **41**, D590-D596 (2013).
950

951   55.   Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex
952   metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics* **31**, 1674-1676 (2015).
953

954   56.   Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357 (2012).
955

956   57.   Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from

957   complex microbial communities. *PeerJ* **3**, e1165 (2015).

958

959   58.   Sieber CM*, et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring

960   strategy. *Nat Microbiol* **3**, 836-843 (2018).

961

962   59.   Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial

963   genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043-1055 (2015).

964

965   60.   Parks DH*, et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of

966   life. *Nat Microbiol* **2**, 1533-1542 (2017).

967

968   61.   Laczny CC*, et al.* VizBin-an application for reference-independent visualization and human-augmented binning

969   of metagenomic data. *Microbiome* **3**, 1 (2015).

970

971   62.   Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome

972   Taxonomy Database. *Bioinformatics* **36**, 1925-1927 (2020).

973

974   63.   Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes.

975   *Nucleic Acids Res* **44**, W54-W57 (2016).

976

977   64.   Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).

978

979   65.   Eddy SR. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195 (2011).

980

981   66.   Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59-60

982   (2015).

983

984   67.   Zhang H*, et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids*

985   *Res* **46**, W95-W101 (2018).

986

987   68.   Rawlings ND, Barrett AJ, Finn R. Twenty years of the MEROPS database of proteolytic enzymes, their substrates

988   and inhibitors. *Nucleic Acids Res* **44**, D343-D350 (2016).

989

990   69.   Katoh K, Standley DM. MAFFT: iterative refinement and additional methods. In: *Multiple Sequence Alignment*

991   *Methods*). Humana Press (2014).

992

993   70.   Kearse M*, et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization

994   and analysis of sequence data. *Bioinformatics* **28**, 1647-1649 (2012).

995

996   71.   Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for

997   estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274 (2014).

998

999   72.   Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation.

1000   *Bioinformatics* **23**, 127-128 (2007).

1001

1002   73.   Pruesse E, Peplies J, Gloeckner FO. SINA: Accurate high-throughput multiple sequence alignment of ribosomal
1003   RNA genes. *Bioinformatics* **28**, 1823-1829 (2012).

1004

1005   74.   Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic
1006   data. *Bioinformatics* **28**, 3211-3217 (2012).

1007

1008   75.   Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with
1009   DESeq2. *Genome Biol* **15**, 550 (2014).

1010

1011   76.   Breier J, *et al.* Sulfur, sulfides, oxides and organic matter aggregated in submarine hydrothermal plumes at
1012   9°50'N East Pacific Rise. *Geochim Cosmochim Acta* **88**, 216-236 (2012).

1013

1014   77.   Reeves EP, McDermott JM, Seewald JS. The origin of methanethiol in midocean ridge hydrothermal fluids. *Proc*
1015   *Natl Acad Sci U S A* **111**, 5474 (2014).

1016

1017   78.   Bethke CM. *Geochemical and biogeochemical reaction modeling*. Cambridge University Press (2007).

1018

1019   79.   Johnson JW, Oelkers EH, Helgeson HC. SUPCRT92: A software package for calculating the standard molal
1020   thermodynamic properties of minerals, gases, aqueous species, and reactions from 1 to 5000 bar and 0 to
1021   1000 °C *Computers & Geosciences* **18**, 899-947 (1992).

1022

1023   80.   Hagberg A, Swart P, S Chult D. Exploring network structure, dynamics, and function using NetworkX. In:
1024   *Proceedings of the 7th Python in Science conference*) (2008).

1025

1026   81.   Crits-Christoph A, Olm MR, Diamond S, Bouma-Gregson K, Banfield JF. Soil bacterial populations are shaped by
1027   recombination and gene-specific selection across a grassland meadow. *ISME J* **14**, 1834-1846 (2020).

1028

1029   82.   Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting F_{ST}: the impact of rare variants.
1030   *Genome Res* **23**, 1514-1521 (2013).

1031

1032   83.   Hudson RR, Slatkin M, Maddison WP. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**,
1033   583-589 (1992).

1034

1035