



# Selection Is a Significant Driver of Gene Gain and Loss in the Pangenome of the Bacterial Genus *Sulfurovum* in Geographically Distinct Deep-Sea Hydrothermal Vents

Alief Moulana,<sup>a,b</sup>  Rika E. Anderson,<sup>a</sup> Caroline S. Fortunato,<sup>c</sup>  Julie A. Huber<sup>d</sup>

<sup>a</sup>Biology Department, Carleton College, Northfield, Minnesota, USA

<sup>b</sup>Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, USA

<sup>c</sup>Department of Biology, Wilkes University, Wilkes-Barre, Pennsylvania, USA

<sup>d</sup>Marine Chemistry & Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts, USA

Alief Moulana and Rika Anderson contributed equally to this work. Author order was determined in order of increasing seniority.

**ABSTRACT** Microbial genomes have highly variable gene content, and the evolutionary history of microbial populations is shaped by gene gain and loss mediated by horizontal gene transfer and selection. To evaluate the influence of selection on gene content variation in hydrothermal vent microbial populations, we examined 22 metagenome-assembled genomes (MAGs) (70 to 97% complete) from the ubiquitous vent Epsilonbacteraeota genus *Sulfurovum* that were recovered from two deep-sea hydrothermal vent regions, Axial Seamount in the northeastern Pacific Ocean (13 MAGs) and the Mid-Cayman Rise in the Caribbean Sea (9 MAGs). Genes involved in housekeeping functions were highly conserved across *Sulfurovum* lineages. However, genes involved in environment-specific functions, and in particular phosphate regulation, were found mostly in *Sulfurovum* genomes from the Mid-Cayman Rise in the low-phosphate Atlantic Ocean environment, suggesting that nutrient limitation is an important selective pressure for these bacteria. Furthermore, genes that were rare within the pangenome were more likely to undergo positive selection than genes that were highly conserved in the pangenome, and they also appeared to have experienced gene-specific sweeps. Our results suggest that selection is a significant driver of gene gain and loss for dominant microbial lineages in hydrothermal vents and highlight the importance of factors like nutrient limitation in driving microbial adaptation and evolution.

**IMPORTANCE** Microbes can alter their gene content through the gain and loss of genes. However, there is some debate as to whether natural selection or neutral processes play a stronger role in molding the gene content of microbial genomes. In this study, we examined variation in gene content for the Epsilonbacteraeota genus *Sulfurovum* from deep-sea hydrothermal vents, which are dynamic habitats known for extensive horizontal gene transfer within microbial populations. Our results show that natural selection is a strong driver of *Sulfurovum* gene content and that nutrient limitation in particular has shaped the *Sulfurovum* genome, leading to differences in gene content between ocean basins. Our results also suggest that recently acquired genes undergo stronger selection than genes that were acquired in the more distant past. Overall, our results highlight the importance of natural selection in driving the evolution of microbial populations in these dynamic habitats.

**KEYWORDS** hydrothermal vents, metagenomics, pangenome

Microbial populations can adapt to their environment through the acquisition of genes via horizontal gene transfer (HGT), enabling the introduction of novel functions (1, 2). Because HGT facilitates genetic exchange across distinct phylogenetic


**Citation** Moulana A, Anderson RE, Fortunato CS, Huber JA. 2020. Selection is a significant driver of gene gain and loss in the pangenome of the bacterial genus *Sulfurovum* in geographically distinct deep-sea hydrothermal vents. mSystems 5:e00673-19. <https://doi.org/10.1128/mSystems.00673-19>.

**Editor** Holly Bik, University of Georgia

**Copyright** © 2020 Moulana et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Rika E. Anderson, [randerson@carleton.edu](mailto:randerson@carleton.edu).

This is C-DEBI contribution number 525.

 The *Sulfurovum* pangenome from deep-sea hydrothermal vents shows evidence of biogeographic differentiation and selective signatures

**Received** 16 October 2019

**Accepted** 30 March 2020

**Published** 14 April 2020

clades and even domains (3, 4), its common occurrence allows microorganisms to develop high variation in gene content within individual species or genera (5, 6). Although some gene content variation arises during duplication and gene loss events, HGT is hypothesized to be the primary mechanism of novel gene acquisition (1). This acquisition allows for some genes, called accessory genes, to be present in only a few strains, but not all strains, of a given species (7). In contrast, genes that belong to the core genome are mostly inherited and conserved across almost all strains within the species. The pangenome of a given microbial species or genus includes the sum of all core and accessory genes identified within that genus or species (7, 8).

The evolutionary processes that drive the accumulation of accessory genes in a microbial species in any system are still debated. Daubin and Ochman (10) show that compared to the core genome, the accessory genome of *Escherichia coli* has a significantly higher ratio of nonsynonymous (altering the amino acid sequence) to synonymous (not altering the amino acid sequence) nucleotide substitutions in an alignment of sequences. This difference suggests that the accessory genome must undergo much weaker, or neutral, selection compared to the highly conserved core genome. Consequently, it was suggested that neutral processes, such as genetic drift rather than natural selection, drive the accumulation of accessory genes. More recently, Andreani et al. (11) argued for the neutral evolution of pangenomes by showing a significant positive correlation between genome fluidity, a measure of dissimilarity in gene content within species, and effective population size. They argued that because larger effective populations had more pangenome diversity, this was consistent with the expectation that larger populations should have more genetic diversity due to neutral evolution (12). In contrast, other evidence suggests that selection, rather than neutral processes, is a strong factor in driving pangenome evolution. Microbes with similar ecology tend to share the same genes, even after controlling for phylogeny and geography, suggesting that novel genes acquired through HGT are maintained by selection pressures from the local ecosystem. For example, an analysis of 657 sequenced prokaryote genomes revealed that highly connected donors and recipients in transfers are of the same general habitat (13). Similarly, in the human microbiome, a network of 10,770 recently transferred genes was also shaped by ecological niche, and in particular, body site location (14). Further, acquired accessory genes have enabled bacteria to survive in extreme environments (15–17), suggesting that the acquired genes allow their hosts to exploit new habitats. Thus, it remains an open question whether the accumulation and retention of accessory genes are generally driven by natural selection or neutral processes.

In cases where selection is the primary driver for the accumulation of accessory genes, one would expect selection signatures to be present in these genes. A selective sweep, for instance, is known to occur in microbial populations when a genotype has a fitness advantage and sweeps through the whole population along with its associated genome (18). However, despite the support for this phenomenon in theoretical models (19, 20) and laboratory settings (21, 22), genome-wide selective sweeps have rarely been observed among environmental microbes (23, 24), and a high genetic diversity has been found in the core genome within species (25–27). Instead of sweeping at the genome-wide level, individual genes within a region could sweep independently of the rest of the genome, a phenomenon that has been observed, for instance, in the *gapA* and *pabB* loci of *E. coli* (28) and in single locus variation sites in the haloarchaeal genus *Halorubrum* (29, 30). Such patterns of gene-specific sweeps are hypothesized to emerge from high recombination rates, which can unlink a gene from the rest of the genome (31). Other hypotheses suggest that negative frequency-dependent selection constrains the adaptive-acquired genes so that they, and the recipient genome, do not sweep the entire population (24, 32). In this case, top-down control exerted by phages could selectively remove individuals that rise to high frequency because of greater fitness (33).

In order to identify the drivers of evolution in accessory genes in a microbial population in nature, we examined the evolutionary dynamics of genome variation in

deep-sea hydrothermal vent habitats, which are powered by rich chemical redox gradients that support diverse microbial communities (34–37). In hydrothermal vent systems, HGT is extensive and facilitates the accumulation of accessory genes in microbial pangenomes in the ecosystem (38–41). A high diversity of other microorganisms in close proximity to one another may allow for interspecies transfers to occur, and the high abundance of viruses at vents provides another mechanism for such horizontal gene transfer to take place (42, 43). Moreover, organisms have to adapt to rapidly changing conditions in the gradient-driven environments of deep-sea hydrothermal vents (43–46), making it an ideal environment to study the evolutionary dynamics of genome variation.

Despite the prevalence of HGT and the unique challenges presented to microorganisms native to the vent habitat, very few studies have examined gene flow and biogeographic structuring of microbial populations in hydrothermal systems. Work on macrofauna in deep-sea vent sites has indicated that the degree of gene flow between sites varies widely depending on the species and location (47), but few studies have focused on the metapopulation structuring of microbial populations in this environment. Work based on multilocus sequence typing (MLST) to characterize specific microbial strains has demonstrated wide dispersal for vent microbes in general, and a correlation between genetic and geographic distance (48). Little is known about biogeographic structuring of other microbial populations at hydrothermal vents, and no studies have investigated the degree to which gene flow molds microbial pangenomes in this habitat. Therefore, such studies are crucial for understanding how evolution molds microbial lineages across space and time at these globally distributed, highly productive deep-sea environments (49).

We specifically studied the structure of the *Sulfurovum* pangenome using metagenome-assembled genomes (MAGs) collected from low-temperature diffuse fluids from two vent fields, one in the Caribbean Sea and the other in the Pacific Ocean. The *Sulfurovum* genus is a diverse group of mesophilic, sulfur-oxidizing Epsilonbacteraeota that is ubiquitous and abundant in deep-sea hydrothermal vent fluids across the world's oceans (27, 50). Previous work has shown that sulfur-oxidizing bacteria in general demonstrate niche partitioning across fluid gradients in deep-sea hydrothermal vents and other sulfide-rich marine ecosystems, such as anoxic basins (36, 51–53). Work based on both 16S rRNA gene and metagenomic sequencing has revealed that the *Sulfurovum* clade in particular exhibits extensive genomic diversity, and it has been hypothesized that this is due to the steep geochemical gradients that are characteristic of the habitats where *Sulfurovum* has been found (27, 51, 54). Others have examined the nature of genomic variation in the *Sulfurovum* clade, revealing intraclade variation in the *sox* gene complex, oxygen- and nitrogen-cycling genes (51, 54). Although previous studies have demonstrated some genomic variation governing metabolic capabilities among Epsilonbacteraeota in general (55) and *Sulfurovum* in particular (51, 54), here we focused on biogeographic structuring and signatures of selection on the *Sulfurovum* pangenome (55). We investigated evolutionary dynamics on the *Sulfurovum* pangenome in two geographically distinct hydrothermal vent habitats to determine what, if any, selection pressures are present by assessing functional differences in genomes, computing selection strength, and searching for evidence of selective sweeps.

Our first study site is the Mid-Cayman Rise site in the Caribbean Sea, an ultraslow spreading ridge which harbors two geochemically distinct hydrothermal vent fields along the ~110-km-long ridge: Von Damm (~2,350 m), an ultramafic-hosted vent field, and Piccard (~4,950 m), a deep, mafic-hosted vent field (56). Although these vent fields are substantially different in terms of depth and geological context, both vent fields host fluids rich in hydrogen and hydrogen-utilizing microbes (50, 56, 57). The second site is Axial Seamount, an active submarine volcano on the Juan de Fuca Ridge in the northeastern Pacific Ocean, which is home to multiple vent fields within the caldera. Hydrogen sulfide oxidation is the dominant chemical energy source for microbial metabolism at Axial Seamount, and the availability of hydrogen and nitrate influences the metabolism

of these diverse microbial communities (34, 58–61). Here, using metagenomic techniques, we study pangenome evolution in two geographically and geochemically distinct hydrothermal vent systems and demonstrate the significance of selection in the maintenance of accessory genes in the pangenome.

## RESULTS

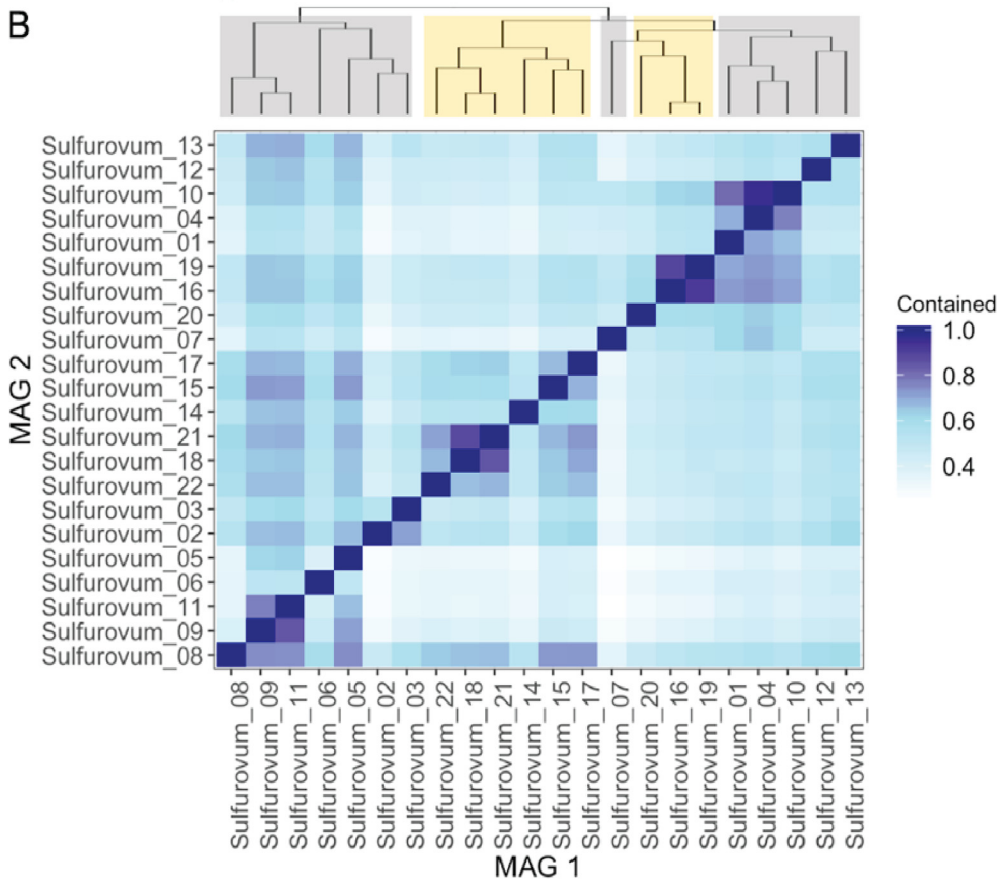
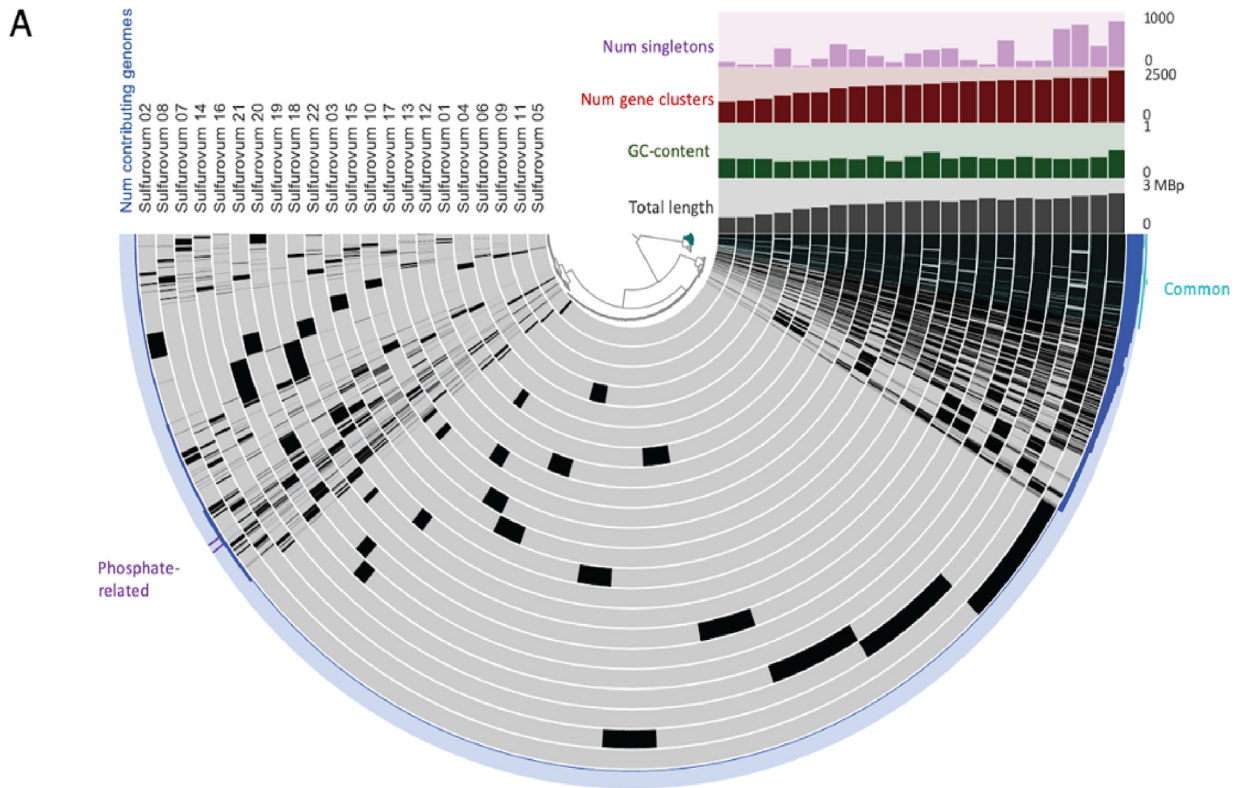
**Gene content variability among 22 *Sulfurovum* MAGs.** We collected a total of 159 MAGs from Axial Seamount and 77 MAGs from the Mid-Cayman Rise. The taxonomic identification of the MAGs recovered from the two vent fields were different (see Fig. S1A in the supplemental material), and the collection of MAGs recovered from Axial Seamount had a higher species richness than those from the Mid-Cayman Rise. While both vent fields were dominated by *Sulfurovum*, the taxonomic compositions then diverged. *Aquificales*, *Thiotrichales*, and *Sulfurovum* made up at least one-third of the microbial community recovered from the Mid-Cayman Rise samples. In contrast, despite the abundance of *Arcobacter* and *Sulfurovum* in the Axial samples, other taxa with much lower frequency were present at Axial Seamount. Due to the high abundance of *Sulfurovum* in both fields, we decided to use this taxon for our subsequent pangenomic analyses.

We examined 22 *Sulfurovum* MAGs (13 from Axial Seamount and 9 from the Mid-Cayman Rise) to create a pangenomic profile of *Sulfurovum* in the two vent fields (Fig. 1A). A phylogenomic tree of all *Sulfurovum* MAGs based on universal single-copy genes (Fig. S1B) indicates that MAGs from Axial Seamount and the Mid-Cayman Rise did not cluster entirely separately but that MAGs from the same location tended to cluster together within clades. We treated each MAG as a population of the same (or similar) *Sulfurovum* strains. The MAGs had similar GC content (mean, 36.44; maximum [max], 50.57; minimum [min], 29.79) and genome size (mean, 1.6 Mbp; max, 2.2 Mbp; min, 875 kbp) (see Table S2 in the supplemental material). The open reading frames (ORFs) in all *Sulfurovum* MAGs were clustered based on their amino acid sequence similarities (see Materials and Methods). For each gene cluster, we counted the number of genomes the gene cluster was found in and defined this count as gene frequency across MAGs.

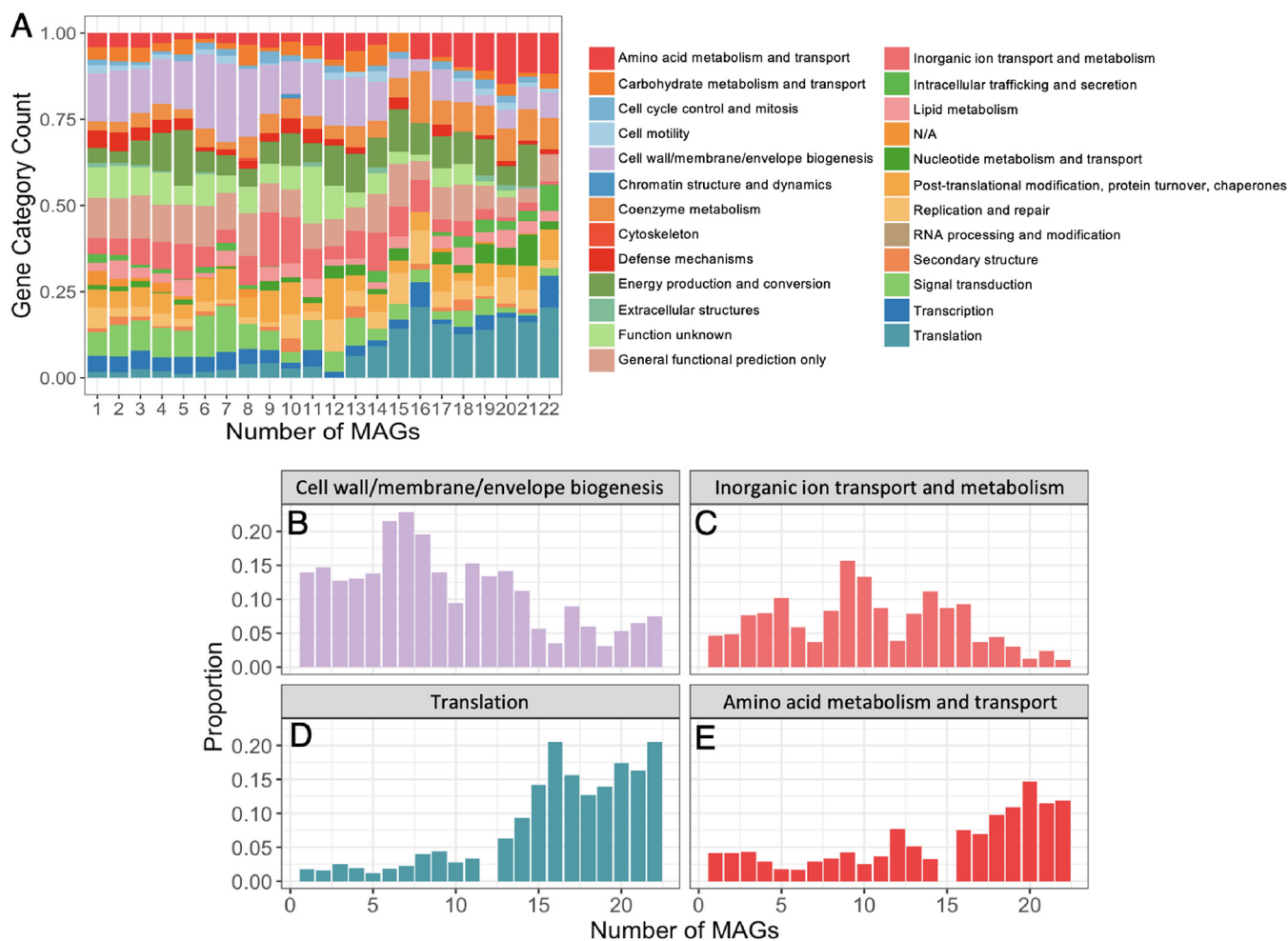
In total, 10,263 gene clusters were identified, formed by 39,120 unique ORFs. On average, ~4.6% of the genes in the genome were shared by all MAGs, and 83 gene clusters constituted the core genome (Fig. 1A). The addition of more genomes into the analysis would likely lead to a higher proportion of accessory genes because the *Sulfurovum* pangenome, as defined by these MAGs, is open (Fig. S2A). It is important to note that due to the incompleteness of the MAGs recovered (70 to 97%), the core genome is likely much larger, with core genes missed in our analysis. To examine this, we estimated the probability of an accessory gene being found in all MAGs as a core gene and determined that genes identified in 21 and 22 MAGs had, on average, a probability of 0.0604 and 0.2554, respectively, to be found in all *Sulfurovum* genomes (Fig. S2B). This probability declined for lower-frequency genes. Moreover, some genes found in all MAGs may in reality be missing in some of the MAGs due to redundancy, which ranged from <1 to 5% for the *Sulfurovum* MAGs. Thus, here we focus on the relative frequency at which a specific gene cluster is found across MAGs and do not adhere to a strict definition of what constitutes an accessory and core gene.

Pairwise comparison among the MAGs showed a wide range of gene content similarity (from 0.25 to 0.96; Fig. 1B). Based on this gene content similarity, MAGs from the same region (Axial Seamount or Mid-Cayman Rise) tended to cluster together, but not exclusively (Fig. 1B). Further, the asymmetry of this similarity matrix highlights the different genome sizes and completeness of the MAGs. For instance, although 95% of genes present in *Sulfurovum\_10* were contained in *Sulfurovum\_04*, only 76% of *Sulfurovum\_04* genes were in *Sulfurovum\_10* (Fig. 1B).

**Functional differences between high- and low-frequency genes in the *Sulfurovum* pangenome.** Gene annotation revealed functional differences between high- and low-frequency genes in the *Sulfurovum* pangenomes. To simplify, we put each



**FIG 1** Pangenomic structure of *Sulfurovum* populations in the two vent fields. (A) Anvi'o image of each *Sulfurovum* metagenome-assembled genome (MAG) represented by a gray ring in the circle, ordered from MAGs with the most (outermost) to the least (innermost) gene clusters. The (Continued on next page)

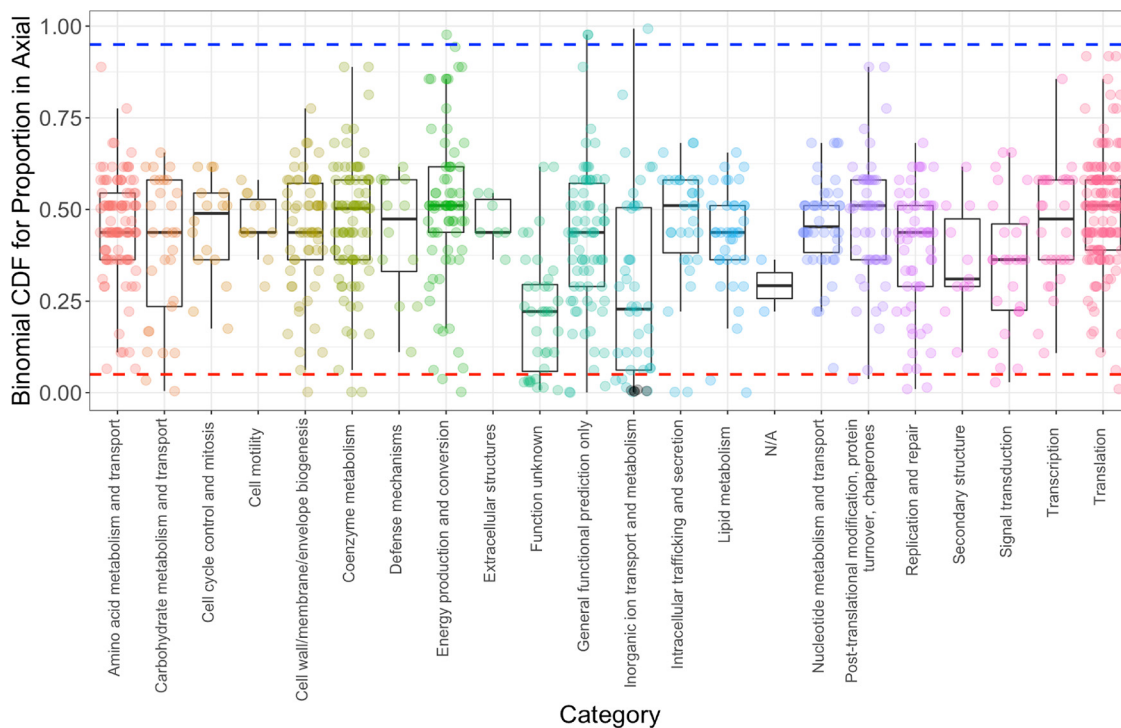


**FIG 2** ORF annotations across the *Sulfurovum* metagenome-assembled genomes (MAGs) as a function of gene cluster frequency. (A) Each bar represents a cluster of orthologous groups (COG) category for the gene cluster of interest in the *Sulfurovum* MAGs. Each color corresponds to each COG category. Gene clusters with unknown annotations are excluded. N/A, no COG category assigned. (B through E) The proportion of ORFs that are in each of four COG categories is shown as a function of the gene cluster frequency (number of MAGs that contains the gene cluster). The x axis represents the number of MAGs sharing a specific gene cluster. The y axis represents the proportion of gene clusters that fall within a specific COG category out of all gene clusters shared among that number of MAGs. These four COG categories were included because they exhibited the strongest differences in abundance between high-frequency and low-frequency genes.

gene annotation into its corresponding cluster of orthologous groups (COG) category. The distribution of these COG categories varies across gene cluster frequency in the *Sulfurovum* pangenome (Fig. 2A). For instance, translation (COG category J) and amino acid metabolism (COG category E) were significantly enriched in the highly conserved, higher-frequency genes compared to lower-frequency genes (Fig. 2D and E; Fig. S3;  $P < 0.0001$ ). Other categories that were enriched in higher-frequency genes included coenzyme metabolism (COG category H) and nucleotide metabolism and transport (COG category F) (Fig. S3). The translation functions enriched in the higher-frequency genes included ribosomal proteins, 16S rRNA, and tRNA synthetases (see Data Set S1A in the supplemental material). Genes involved in other housekeeping functions, espe-

**FIG 1** Legend (Continued)

black boxes in each genome represent open reading frames (ORFs) recovered in that MAG. Boxes that align create a gene cluster for that ORF. The outermost ring in blue represents the number of MAGs containing the corresponding ORF. Other properties are shown in the upper right quadrant, where “num singletons” represents the number of genes found in only one MAG and “num gene clusters” represents the number of total gene clusters found in that MAG. The most common gene clusters (green) and phosphate-related gene clusters (blue) are highlighted. (B) Proportion of gene clusters present in one *Sulfurovum* MAG compared to every other *Sulfurovum* MAG. MAGs are hierarchically clustered by these proportions. Gray branches show Axial Seamount lineages, whereas golden branches show Mid-Cayman Rise lineages.



**FIG 3** The cumulative distribution function (CDF) value for the proportion of genes present in Axial Seamount metagenome-assembled genomes (MAGs) relative to Mid-Cayman Rise MAGs across different clusters of orthologous groups (COG) categories. A higher value means that the open reading frame (ORF) is found in more Axial MAGs than expected, and vice versa. Some COG categories are not shown because they were not represented in enough MAGs (see Results). The blue dashed line represents a CDF value of 0.95, and a statistical significance cutoff for ORFs represented more in Axial MAGs ( $P < 0.05$ ). The red dashed line represents the cutoff for ORFs represented more in Mid-Cayman Rise MAGs. Phosphate-related genes are shown in black under “Inorganic ion transport and metabolism.”

cially those related to transcription, electron transport, and general cell cycle, were also common functions for high-frequency genes in the *Sulfurovum* pangenome.

In contrast, genes with functions relating to cell membrane/wall/envelope biogenesis (COG category M), signal transduction (COG category T), and inorganic ion transport and metabolism (COG category P) were significantly more common in lower-frequency genes relative to higher-frequency genes within the *Sulfurovum* pangenome (Fig. 2B and C; Fig. S3;  $P < 0.0001$ ;  $t$  test). A large amount of singleton genes in the cell membrane/wall/envelope biogenesis category (COG category M), for instance, coded for glycosyltransferase involved in cell wall biosynthesis (Data Set S1B). Other common functions in genes found in <15 MAGs included transposases, ABC-type transport systems, and surface antigens. In general, outer membrane proteins made up a significant proportion of the lower-frequency genes. Further, the genes related to inorganic ion transport and metabolism (COG category P) were uniquely enriched among the medium-frequency genes (found in 9 to 10 MAGs) (Fig. 2C). Genes related to carbohydrate metabolism and transport (COG category G) shared a similar pattern to genes in the P category (inorganic ion transport and metabolism).

We next sought to determine whether specific genes were enriched in *Sulfurovum* genomes at either Axial Seamount or the Mid-Cayman Rise. For each gene cluster, we tallied all MAGs in which the cluster was found and then calculated the number of those MAGs that were recovered from Axial Seamount. Then, we normalized the proportions of gene clusters found in Axial Seamount by calculating their respective binomial cumulative distribution function (CDF) values with an expected value of 13/22 (see Materials and Methods; Fig. 3). The CDF is effectively the probability that the actual number of Axial Seamount MAGs that a gene cluster is found in is higher than would be expected randomly. To avoid bias driven by lower-frequency genes, only gene

**TABLE 1** COG categories for inorganic transport and metabolism functions with the lowest representation in the Axial Seamount genomes<sup>a</sup>

Cluster ID	COG function(s)	Axial proportion
GC_00001517	Outer membrane receptor proteins, mostly Fe transport; outer membrane cobalamin receptor protein	0.00191752
GC_00000911	Truncated hemoglobin Yjbl	0.00449269
GC_00001253	ABC-type phosphate transport system, periplasmic component	0.00449269
GC_00001177	ABC-type phosphate transport system, ATPase component	0.00449269
GC_00001293	ABC-type phosphate transport system, permease component	0.00449269
GC_00001150	Phosphate uptake regulator	0.00449269
GC_00001267	ABC-type phosphate transport system, permease component	0.00984906
GC_00001114	Copper chaperone CopZ	0.01435327
GC_00001019	Arsenate reductase and related proteins, glutaredoxin family	0.03377563
GC_00000935	Arsenate reductase and related proteins, glutaredoxin family	0.03736214
GC_00001144	Adenylyl- and sulfurtransferase ThiI participates in tRNA 4-thiouridine and thiamine biosynthesis; rhodanese-related sulfurtransferase	0.06183014
GC_00001031	Cu/Ag efflux pump CusA	0.06183014
GC_00000853	Exopolyphosphatase/pppGpp-phosphohydrolase	0.06711717
GC_00000940	Divalent metal cation (Fe/Co/Zn/Cd) transporter	0.06711717
GC_00000827	Copper oxidase (laccase) domain	0.10814304

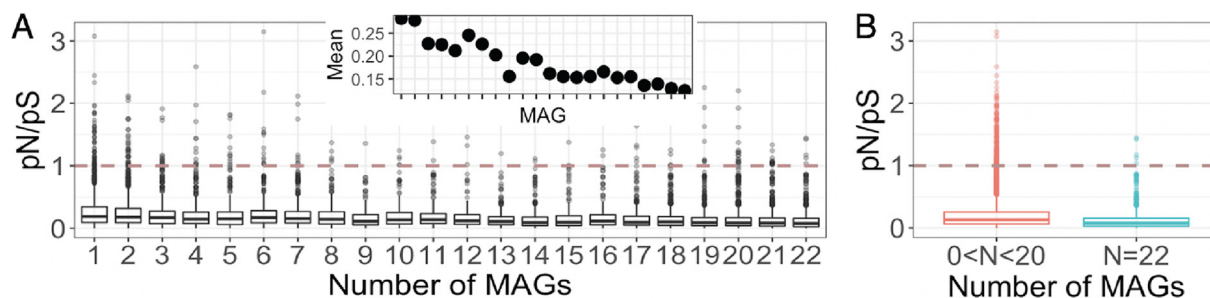
<sup>a</sup>Fifteen functions from the COG P category with the least representation are presented along with each of their cluster identifiers (IDs) in the pangenome and proportion value.

clusters that were found in at least 7 Mid-Cayman Rise MAGs or 12 Axial MAGS were included. This also ensured that the genes included in the analysis were conserved in at least one of the vent fields (i.e., these genes are at least medium-frequency genes). Although the proportions for most of the gene clusters were similar and concentrated around the expected value of 0.5, gene clusters related to inorganic ion transport and metabolism (COG category P) and unknown function (COG category S) had significantly higher representation at the Mid-Cayman Rise compared to Axial Seamount ( $P = 0.0005$  and  $P < 0.0001$ , respectively;  $t$  test, Fig. 3). Of the 15 P category genes that were enriched at the Mid-Cayman Rise, four of them were related to phosphate uptake and regulation (Table 1). Three of these functions make up the components of ABC-type phosphate transport system. Moreover, arsenate reductase genes and some heavy metal transporters were more highly represented in *Sulfurovum* MAGs from the Mid-Cayman Rise compared to Axial Seamount. In contrast, only three gene clusters had significantly higher representation in the Axial MAGs (CDF > 0.95), with two of them being DNA-binding beta-propeller fold protein YncE (Fig. 3), which is known to be involved in iron metabolism (62).

**Signatures of selection in the pangenome.** In order to determine whether the *Sulfurovum* pangenome showed evidence of natural selection, we estimated the  $pN/pS$  ratios of each of the 39,120 identified ORFs in the pangenome. The  $pN/pS$  ratio is defined as the ratio of the proportion of nonsynonymous polymorphisms to the proportion of synonymous polymorphisms given available sites. Excluding all nonpolymorphic ORFs and variants that do not pass the criteria (see Materials and Methods), 14,226 ORFs in the *Sulfurovum* pangenome had defined, finite  $pN/pS$  ratio values, with a mean of 0.1850. Most of the genes had  $pN/pS$  values of less than one, indicating purifying or stabilizing selection (Fig. 4A). In total, only 233 ORFs, 60 of which were singletons, had a  $pN/pS$  value over one with a maximum value of 3.1429, indicating positive selection. In contrast, 1,575 ORFs had a  $pN/pS$  value of 0, a strong negative selection signature. We observed different selective signatures across the *Sulfurovum* MAGs. Out of the 14,226 ORFs with a defined  $pN/pS$  value, only three of them were in *Sulfurovum\_01* and none in *Sulfurovum\_08*. Additionally, *Sulfurovum\_16* and *Sulfurovum\_19* had ORFs with higher  $pN/pS$  values compared to the rest of the MAGs (Fig. S4A).

Generally, genes that were more highly conserved across *Sulfurovum* MAGs had lower  $pN/pS$  values (Fig. 4A). For comparison, singletons had a mean  $pN/pS$  value of 0.2823, whereas genes found in all 22 MAGs had an average  $pN/pS$  value of 0.1244.





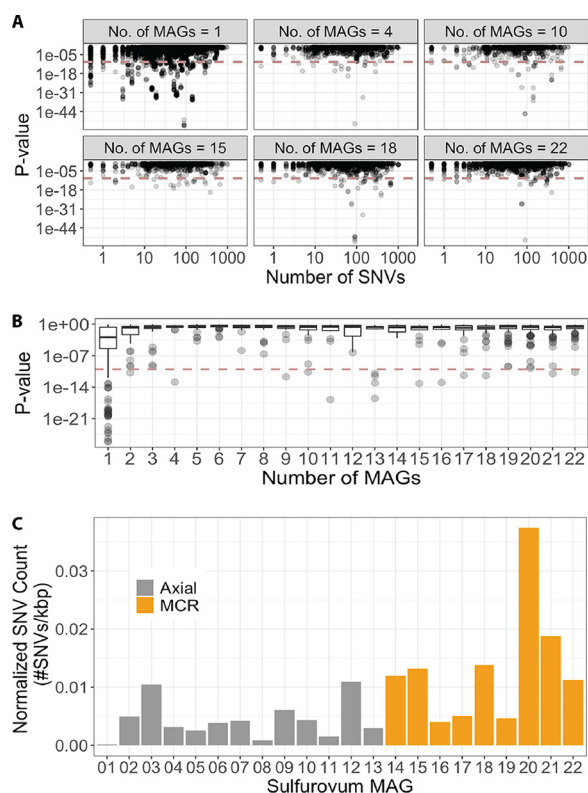
**FIG 4** (A) The estimated  $pN/pS$  ratio of each identified open reading frame (ORF) within each gene cluster as a function of gene cluster frequency in the pangenome. The x axis represents the number of metagenome-assembled genomes (MAGs) sharing a specific gene cluster. The inset in panel A shows the trend of the mean of  $pN/pS$  ratios from lowest to highest frequency. The x axis is the same as in the larger figure. The dashed red line represents the value of  $pN/pS = 1$ . A similar plot is shown in panel B, where all ORFs with certain frequencies are grouped together ( $N$ ), showing that ORFs within core gene clusters have lower  $pN/pS$  ratios.

These values significantly differed from the nonsingleton genes and genes not found in all MAGs, respectively (Fig. S4B; both  $P < 0.0001$ ; Wilcoxon two-sided test). In addition, the mean  $pN/pS$  values for genes found in fewer than 20 MAGs (i.e., “accessory” genes) were significantly higher than for genes found in 22 MAGs (i.e., “core” genes) (Fig. 4B;  $P < 0.0001$ ; Wilcoxon two-sided test). There was no correlation between the gene function category and the  $pN/pS$  ratios (Fig. S4C). We also observed that the  $pN/pS$  ratios for the phosphate uptake and regulation genes that were in higher abundance in Mid-Cayman Rise MAGs had an average  $pN/pS$  ratio of 0.0776 (Fig. S4D). Five of these 21 genes had a  $pN/pS$  ratio of 0.

To determine whether the acquisition of specific genes led to gene-specific sweeps, we searched for unusually low single nucleotide variation (SNV) density contigs containing a gene in the pangenome. We assumed that the number of SNVs throughout the region follows a Poisson distribution with an expected value derived from the genome-wide SNV density (see Materials and Methods). Lower  $P$  values signify that the number of SNVs observed in the gene are lower than expected, indicating strong selective pressure. In general, singletons exhibited lower  $P$  values and a higher proportion of SNV-free contigs (Fig. 5A; Fig. S5A) compared to other genes. In fact, the singleton group had the highest proportion of genes with a  $P$  value of less than  $1e-10$  (Fig. S5B), whereas the gene clusters found in 22 MAGs (i.e., high-frequency genes) had the lowest proportion. When considering only genes in SNV-free contigs (1,382 genes are in SNV-free contigs), most of the genes with lower  $P$  values were singletons, which generally belonged to contigs with significantly lower SNV density (Fig. 5B). MAG *Sulfurovum\_01* from Axial Seamount had a particularly low SNV density of only 0.1 SNV per 1 kbp (Fig. 5C). Further, *Sulfurovum\_08* and *Sulfurovum\_11* from Axial Seamount had fewer than 2 SNVs per 1 kbp.

## DISCUSSION

Efforts to study wild microbial pangenomes face inherent challenges due to the difficulties of culturing environmental microorganisms (63). This study uses metagenomic data to investigate pangenome evolution in deep-sea hydrothermal vents, an environment where horizontal gene transfer, the primary mechanism of accessory genome accumulation, is extensive, and where microbial populations inhabit dynamic, geologically and geographically distinct habitats (38–41). We used metagenome-assembled genomes as an analog for genomes of a microbial population to study the forces that drive the evolution of microbial communities at deep-sea hydrothermal vents. To simplify our analyses, we focused on the pangenome of *Sulfurovum*, one of the most abundant genera in the two vent fields studied and one of the most important sulfur- and hydrogen-oxidizing bacteria found at hydrothermal vents globally (27, 34, 35, 50, 51) as well as in other sulfur-rich habitats, including caves and benthic sediments (64, 65).



**FIG 5** Gene-specific sweep signatures based on single nucleotide variants (SNVs) and  $P$  values. (A) The  $P$  value for each contig is plotted relative to the number of actual SNVs in the contig. Each point represents a single contig. The plots are separated according to the number of metagenome-assembled genomes (MAGs) in which that specific gene cluster was found. (B)  $P$  values as a function of gene cluster frequency in the pangenome. The x axis represents the number of MAGs sharing a specific gene cluster. (C) The SNV density for each MAG at either Axial Seamount of Mid-Cayman Rise. The MAG numbering is indicated in Table S2 in the supplemental material.

**The *Sulfurovum* pangenome exhibits extensive variability and biogeographic structuring.** We observed extensive variation within the *Sulfurovum* pangenome, which is consistent with previous reports for this clade (51). Although some pairs of MAGs had high (>90%) overlap in gene content, most MAG pairs did not share a high proportion of gene content. In fact, genes found in all 22 MAGs made up less than 5% of the whole pangenome. This is possibly an artifact of the incomplete nature of MAGs, such that some genes that appear to be accessory actually belong to the core genome. However, the low number of universally conserved genes has also previously been observed in the pangenomes of *Escherichia coli* (13%) at the species level (66) and in the low-light *Prochlorococcus* clade (10%) at the genus level (67). Similarly, as more MAGs are recovered and more *Sulfurovum* genomes are sequenced, we expect the size of the core genome to continue to decrease in size (see Fig. S2A in the supplemental material), as is commonly observed in pangenomic studies (7).

To understand the evolutionary mechanisms that led to this observed variation, we identified the gene functions encoded by the *Sulfurovum* MAGs. The gene function analyses suggested that translation, coenzyme metabolism, amino acid metabolism, lipid metabolism, and cell cycle functions were all enriched among higher-frequency genes. Cells need fundamental housekeeping genes to function, necessitating the maintenance of such genes across genomes and ensuring a low proportion of these genes being lost. This maintenance of housekeeping cell cycle functions in a core genome is frequently observed in pangenome studies (63, 68). The maintenance of some housekeeping genes at medium or low frequency might be a result of the various translational and housekeeping complexes that are only partially conserved across

genomes. In addition, we observed a high proportion of signal transduction, membrane/envelope biogenesis, and carbohydrate metabolism functions in lower-frequency genes, while these functions were rare in higher-frequency genes. Moreover, ion transport and metabolism functions were especially enriched in medium-frequency genes. The presence of nutrient uptake and membrane-related genes in the variable genome has been documented previously in nonvent environments (68–70) as well as in the Epsilonbacteraeota genus *Lebetimonas* from hydrothermal vent systems (17). Cell membrane proteins represent the first contact between the cell and the environment and are also common binding sites for viruses, and thus, diversifying selection may operate on these genes in response to environmental stimuli. The nonrandom distribution of different functions according to gene frequency suggests that the accumulation of genes in the pangenome is likely maintained by a frequency-dependent selection mechanism that suppresses the frequency of some genes while allowing the spread and conservation of other genes (24, 25, 32).

Importantly, we also observed biogeographic structuring of the *Sulfurovum* pangenome. Biogeographic structuring of microbial communities has previously been observed at the community level in deep-sea hydrothermal vent systems, with divergent community structure persisting over time within individual vent sites (34, 58, 59). This most likely results from the seafloor plumbing at individual sites and the degree of fluid flux between sites, which may isolate communities and produce divergent population structure within hydrothermal systems, even at the scale of individual vents (71). However, biogeographic studies have not been conducted at the population or pangenomic scale for microbial populations in hydrothermal systems. Here, we found some evidence for biogeographic structuring in the *Sulfurovum* pangenome, with more migration occurring locally. Some Axial Seamount *Sulfurovum* MAGs tended to cluster together with regard to their gene content similarity, as did the Mid-Cayman Rise MAGs. However, the MAGs did not cluster strictly by geographic location (Fig. 1), and there was high local gene content diversity within each region. Although different environments have been shown to harbor closely related microbial populations with divergent gene content due to selection (68, 72, 73), this pattern suggests that the clustering is more likely to be the result of a relaxed migration barrier within the local vent field compared to the barrier between the ocean basins. This weak biogeographic structuring within vent fields suggests that gene flow is not restricted among distinct *Sulfurovum* populations within vent fields at the evolutionary scales that mold a pangenome. It is important to note that the 22 *Sulfurovum* MAGs analyzed here do not represent an exhaustive data set, and therefore, most likely, some amount of genomic variation is missed. Nevertheless, these data sets were sufficient to observe large-scale biogeographic patterns between and within ocean basins.

We propose that natural selection acting on the *Sulfurovum* pangenome leads to local adaptation of microbial populations to their respective vent environment. The genome similarity analyses revealed gene content differences between Axial Seamount and Mid-Cayman Rise *Sulfurovum* MAGs, in which some genes were more enriched in one vent field compared to the other. The clearest and most striking example of selection molding the *Sulfurovum* pangenome emerges from patterns regarding phosphate uptake. Phosphate regulation- and uptake-related genes were enriched in Mid-Cayman Rise *Sulfurovum* MAGs compared to Axial Seamount MAGs. This observed pattern most likely results from the lower phosphate content of the Atlantic Ocean, where the Mid-Cayman Rise is located, relative to that of the Pacific Ocean, where Axial Seamount is located (74–76). This result is consistent with a previous *Prochlorococcus* study in the surface ocean showing enrichment of phosphate-acquisition genes in *Prochlorococcus* strains in the Atlantic compared to the Pacific (73). Deep-water phosphate values are approximately 1 mm/kg in the Atlantic Ocean and 3 mm/kg in the Pacific Ocean (74), and end-member hydrothermal fluid has extremely low phosphate concentrations because phosphate is removed by hydrothermal processes (77). The diffuse fluids in which hydrothermal microbial communities are found represent a mixture of deep seawater and hydrothermal fluid, and our results indicate that the

differences in nutrient abundances that distinguish the Atlantic and Pacific Oceans are strong enough to mold the genomes of microbes inhabiting hydrothermal vent systems. Moreover, arsenate reductase genes were more highly represented in the Mid-Cayman Rise *Sulfurovum* MAGs. This may result from phosphate scavenging in a phosphate-limited environment, which can cause microorganisms to take up arsenate instead, requiring arsenate reduction into arsenite by arsenate reductase in order to ameliorate arsenate toxicity (73, 78). Because arsenic is known to be naturally enriched in deep-sea hydrothermal systems (79, 80), this arsenate reduction mechanism is particularly important in these habitats, especially when phosphate is limited. Our results extend the range of phosphate- and arsenate-driven selection pressures from the surface oceans (68, 72, 73) to the extreme deep-sea environment despite the multifold increase in phosphate content from the surface to the deep ocean (74, 75).

**Signatures of selection in the *Sulfurovum* pangenome.** If natural selection drives the accumulation of genes in the accessory genome, then evolution should act differently on the accessory genes compared to the core genome. Newly acquired genes might be expected to first undergo positive selection before purifying selection. For instance, Mid-Cayman Rise microbes are adapted to the phosphate-limited environment by incorporating phosphate uptake and regulation genes. These genes exhibited purifying selection with near-zero  $pN/pS$  ratio (Fig. S4D), suggesting that they are already adaptive. In contrast, other accessory genes can have a very different evolutionary scheme. When these genes are beneficial but not yet at the peak of their fitness landscape, positive selection would be expected to act on them. As a result, some accessory genes might have a low  $pN/pS$  ratio due to local adaptation or frequency-dependent adaptation, but some have a higher  $pN/pS$  ratio due to positive selection. In accordance with this, our results show that the  $pN/pS$  ratio spread among lower-frequency genes was higher than that of higher-frequency genes (Fig. 4 and Fig. S6). Moreover, lower-frequency genes in the *Sulfurovum* pangenome generally had a higher  $pN/pS$  ratio compared to the higher-frequency genes, suggesting that genes undergoing positive selection were more likely to be accessory genes than core genes. While it is difficult to determine absolutely from the  $pN/pS$  ratio whether a gene is undergoing positive or negative selection (81), previous work has found that accessory genes undergo more relaxed purifying or negative selection compared to core genes in bacterial genomes (82, 83) or stronger positive selection in the case of *Pseudomonas aeruginosa* populations (84). Thus, our results are consistent with previous work and suggest that selective pressures on the higher- and lower-frequency genes differ, where the lower-frequency genes are more likely to be under positive selection.

Moreover, the ecotype model proposed by Cohan et al. (85) suggests that an individual gaining an evolutionarily beneficial trait should have higher fitness than others in the population, causing it to sweep the population and thus purging genomic diversity. “Ecotype” here is defined as a group of ecologically similar bacteria in which genetic diversity is limited due to selection, drift, or both. However, evidence for genome-wide selective sweeps have mainly come from theoretical simulations and laboratory experiments, whereas naturally occurring sweeps have rarely been observed (23). In the two hydrothermal systems, we observed a significantly low SNV density in MAG *Sulfurovum*\_01 from one vent at Axial Seamount, possibly resulting from either a selective sweep or clonal expansion (Fig. 5C). Purging variants from the population via a sweep or rapid growth of a single clone would reduce nucleotide diversity in the population, resulting in low SNV density. Other MAGs, such as *Sulfurovum*\_08 and *Sulfurovum*\_11 from Axial Seamount, also harbored this low sequence diversity. These *Sulfurovum* MAGs had lower SNV density than the *Sulfurovum* MAGs previously reported from the Mid-Cayman Rise that may have been undergoing selective sweeps or clonal expansions (27). Without time-series data, we cannot distinguish whether this low diversity was caused by a recent bloom event or a selective sweep. However, previous work at Axial Seamount observed an increase in the abundance of *Sulfurovum* at Marker 113 in 2014 compared to other years (58), and *Sulfurovum*\_08 was recovered

from this sample, supporting the hypothesis that a clonal expansion was responsible for the reduced sequence diversity in this MAG.

When the recombination rate and gene exchange among individuals is high enough, specific genes could sweep throughout populations independently of the rest of the genome (29, 31). This gene-specific sweep could also be promoted by phage predation that causes negative frequency-dependent selection (86). Gene-specific sweeps are likely to be particularly important in hydrothermal vent systems due to the high rate of horizontal gene transfer within and among microbial populations in this system (43). This high HGT rate not only creates highly diverse pangenomes but also allows for selection to act on a gene-by-gene basis. We indeed observed that specific regions of the *Sulfurovum* MAGs had lower SNV density compared to the rest of the genome, especially for regions containing singleton genes (Fig. 5). Not only were singletons more likely to have lower SNV density compared to the rest of the genome, they were more likely to be SNV free. Although SNV-free regions could be pervasive, especially in low-SNV-density genomes such as *Sulfurovum\_01*, SNV-free singletons were contained in high-SNV-density genomes too. One possible reason for this observation is low coverage of singletons compared to other genes, but we found no evidence of such bioinformatic artifacts (Fig. S7). Assuming no other possible bioinformatic artifact that results in this observation, gene-specific sweeps might occur more frequently in singletons than in other genes. In this case, the evidence suggests that gene-specific sweeps mostly occur in genes that are newly acquired in the species.

Investigating patterns of selection in microbial genomes is crucial for understanding how microbial populations evolve and adapt to the environment over time. Here, we show evidence for natural selection operating on the *Sulfurovum* pangenome of hydrothermal vents and identify some of the drivers of that selection. Genes that are highly conserved in the *Sulfurovum* pangenome have different functional annotations than lower-frequency genes, suggesting that gene acquisition is not random. At the Mid-Cayman Rise, *Sulfurovum* populations appear to have adapted to the low-phosphate environment through the acquisition and maintenance of phosphate uptake and regulation-related genes. In addition, *pN/pS* ratios reveal that lower-frequency genes are either more susceptible to positive selection or more resistant to negative selection than higher-frequency genes. Finally, we observed some evidence for a genome-wide sweep in one of the *Sulfurovum* populations and gene-specific sweep events in singleton genes. Altogether, we have revealed patterns in the pangenome structure of *Sulfurovum* populations from two distinct hydrothermal vent regions and conclude that their accessory genome structure is molded by natural selection rather than neutral forces. These analyses of genomic variation provide important insights into the dynamics that drive diversity and mold the evolution of microbial populations.

## MATERIALS AND METHODS

**Data collection.** All of the data used for this study was collected from samples obtained from diffuse flow hydrothermal vent fluids emanating from seafloor rocks and sulfide deposits. All methods for sample collection, DNA extraction, and sequencing are described by Reveillaud et al. (50) and Fortunato et al. (58) for Axial Seamount and by Reveillaud et al. (50) and Anderson et al. (27) for the Mid-Cayman Rise. For all samples, diffuse flow hydrothermal fluid was filtered through 0.2- $\mu$ m filters *in situ* while monitoring temperature to capture the microbial communities. We stored filters at  $-80^{\circ}\text{C}$  onboard the ship and extracted DNA on shore as described in the above publications. All samples were sequenced at the Josephine Bay Paul Center at the Marine Biological Laboratory using the Illumina HiSeq or NextSeq sequencing platform. These metagenomes were previously described by Fortunato et al. (58), Reveillaud et al. (50), Anderson et al. (27), and Galambos et al. (54). The data are available from the European Nucleotide Archive Archive (ENA) under study accession numbers [PRJEB7866](#), [PRJEB12000](#), and [PRJEB19456](#) for 2013, 2014, and 2015, respectively, at Axial Seamount, and under study accession [PRJEB15541](#) for the Mid-Cayman Rise. All metagenomes used for this study are shown in Table S1 in the supplemental material.

**Metagenome assembly and mapping.** We conducted all metagenomic processing of *Sulfurovum* MAGs from the Mid-Cayman Rise samples as described by Anderson et al. (27). Briefly, we quality filtered all reads using the illumina-utils package (87), assembled with idba-ud v.1.1.2 (88), and reads were mapped to contigs using bowtie v1.2.2 (89). For the Axial Seamount metagenomes, we first quality filtered the reads using "iu-filter-quality-minoche" within the illumina-utils package (87) and then assembled metagenomic reads using idba-ud v1.1.3 (88) with default settings. For subsequent analyses,

we included only contigs of at least 1,000 bp in length to ensure robust contig clustering based on tetranucleotide frequency and coverage. We mapped the metagenomic reads of each sample to the assembled contigs using bowtie v1.2.2 (89) with default settings. We used anvi'o v4.1.0 (90) to organize the metagenomic contig samples into profiles using the anvi'o command "anvi-profile" with the flag "--profile-SCVs" in order to detect and analyze single nucleotide variants (SNVs) and single codon variants (SCVs).

**Metagenome binning and gene annotation.** We created metagenomic bins based on tetranucleotide composition and relative coverage of each contig across all samples using anvi'o. To estimate the completion and redundancy of metagenomic bins, anvi'o used PRODIGAL v2.6.3 (91) to identify open reading frames (ORFs) in our contigs, and HMMER v3.1b2 (92) to search for the presence of single-copy core genes in bins based on collections for bacteria (93) and archaea (94). Using these estimates, we designated 159 metagenomic bins as metagenome-assembled genomes (MAGs) using a threshold of <10% redundancy and >70% completion, where "redundancy" is determined by the identification of multiple copies of genes that are usually present in a single copy within microbial genomes. We determined the taxonomy of each MAG using PhyloSift (95), using the "phylosift all" flag. Bins that were characterized by PhyloSift as having multiple significant taxonomic hits were excluded from the analysis. We identified 13 *Sulfurovum* MAGs from the Axial Seamount samples and 8 *Sulfurovum* MAGs from the Mid-Cayman Rise samples.

**Tree construction.** The *Sulfurovum* phylogenomic tree was constructed by compiling aligned, concatenated single-copy universal genes created by PhyloSift (95) (using the concat.codon.updated.1.fasta output) for all *Sulfurovum* MAGs in addition to seven reference genomes (NCBI taxonomy identifiers [NCBI:txid] are shown in parentheses): *Sulfurimonas denitrificans* DSM1251 (NCBI:txid326298), *Sulfurovum* sp. strain PC08-66 (NCBI:txid1539063), *Sulfurovum* sp. strain F508-3 (NCBI:txid1539065), *Sulfurovum* sp. strain AS07-7 (NCBI:txid1539062), *Sulfurovum* sp. strain F506-10 (NCBI:txid1539064), *Sulfurovum* sp. strain SCGC AAA036-F05 (NCBI:txid1218800), and *Sulfurovum lithotrophicum* strain ATCC BAA-797 (NCBI:txid206403). The tree was constructed using RAXML v8.2.9 (96) with 100 rapid bootstraps and the GTRGAMMA model (general tree reversible model with gamma distribution) of rate heterogeneity, using *Sulfurimonas denitrificans* DSM1251 as the outgroup.

**Pangenome profiling.** We used a pangenomic workflow within anvi'o to build and perform analyses on the *Sulfurovum* pangenome (67). First, using anvi'o, we constructed a genome storage database that stores all reads, contigs, nucleotide variation, and annotation information from all *Sulfurovum* MAGs. Then, using the "anvi-pan-genome" command, we annotated the called ORFs in each MAG with Diamond v0.9.22.123 (97) to compare each ORF against NCBI's Conserved Domains Database to obtain the cluster of orthologous groups (COG) annotation for each ORF with a maximum E value of  $1e-05$ . Anvi'o then resolved gene clusters across all MAGs using a Markov Cluster Algorithm (MCL) (98) with default arguments (minbit value of 0.5 and inflation value of 2) using the minbit heuristic implemented from ITEP to eliminate weak amino acid matches so that only ORFs with strong amino acid similarities were clustered together (90, 97). The gene clusters were sorted based on their frequency across MAGs to create an across-MAG gene presence-absence matrix. We used the gene presence-absence matrix to conduct a pairwise comparison of the MAGs and draw their distance based on the number of genes present in a MAG that were not contained in another MAG, from which we created a hierarchical clustering dendrogram in R (99). In order to take into account the various levels of MAG completion, we estimated the probability for a gene cluster found in  $n$  MAGs to be found in all MAGs for  $n < 22$  by independently simulating the probability of some genes to be missing in a MAG (scripts and an explanation of the code provided at <https://github.com/carleton-spacehogs/pangenome-selection>).

**Gene annotation analyses.** For all analyses on gene functions, we used the annotations produced in the pangenome profiling step implemented in anvi'o. Although most gene clusters mapped to a unique annotation, a small proportion did not. In those cases, we assigned multiple annotations for those clusters. We then conducted analysis on these annotations and their COG categories in R (see the above URL for the GitHub page for the code and a description of the code). Moreover, we studied the enrichment of some genes in one specific vent environment compared to the other (e.g., Axial Seamount compared to Mid-Cayman Rise) by assuming that the number of Axial genomes in which a gene is found in follows  $\text{Binom}(N, p)$  where  $N$  is the total number of genomes the gene is found in and  $P = 13/22$ .

**$pN/pS$  ratio and sweep analyses.** We used anvi'o (90) to conduct a gene-by-gene  $pN/pS$  ratio analysis using single codon variation (SCV), single nucleotide variation (SNV), and single amino acid variation (SAAV) counts. We generated these counts using the anvi'o command "anvi-gen-variability-profile" for each of the *Sulfurovum* MAGs with the "--engine CDN" flag to obtain SCV variability profiles and the "--engine AA" to obtain SAAV profiles. We ran the anvi'o python script "anvi-script-calculate-pn-ps-ratio" to determine the  $pN/pS$  ratio, which calculates the number of synonymous and nonsynonymous variants, normalized by the potential number of synonymous and nonsynonymous variants. Only reads that cover the full codon context were used for variant detection in this analysis. We analyzed the distribution of SNVs across the pangenome using the variability profile generated with the default option for flag "--engine". We first counted the number of SNVs in each MAG (see the above URL for the GitHub page for the code and an explanation of the code) and then calculated the number of SNVs in the contig in which each ORF was found. We assumed that the number of SNVs in the contig region follows the distribution  $\text{Poiss}(n_c)$  where  $n_c$  is the expected value of SNVs in a contig  $c$  for  $n_c = N l_c / L$ .  $N$  is the total number of SNVs in the MAG,  $L$  is the length of the MAG, and  $l_c$  is the length of contig  $c$ . We computed the Poisson cumulative distribution function (CDF) value for each gene and ascribed the value as the gene-specific sweep  $P$  value for the gene.

**Data and script accessibility.** All R, Python scripts, an explanation of code, and raw results used for this analysis are publicly available on GitHub at <https://github.com/carleton-spacehogs/pangenome-selection>.

## SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

**FIG S1**, PDF file, 0.3 MB.

**FIG S2**, PDF file, 0.1 MB.

**FIG S3**, PDF file, 0.2 MB.

**FIG S4**, PDF file, 0.4 MB.

**FIG S5**, PDF file, 0.4 MB.

**FIG S6**, PDF file, 0.1 MB.

**FIG S7**, PDF file, 0.1 MB.

**TABLE S1**, DOCX file, 0.01 MB.

**TABLE S2**, DOCX file, 0.02 MB.

**DATA SET S1**, XLSX file, 0.2 MB.

## ACKNOWLEDGMENTS

We thank Julie Reveillaud and Emily Reddington for support in the collection and generation of metagenomic data, Chip Breier, David Butterfield, Bill Chadwick, Chris German, Jim Holden, Jill McDermott, and Jeff Seewald for sample collection support at sea, and Jaclyn Saunders for discussions regarding arsenate reductase.

A.M. was supported by Carleton College. R.A. was supported by a NASA Postdoctoral Fellowship with the NASA Astrobiology Institute. This work was supported by a NASA Exobiology grant 80NSSC18K1076 to R.A. and J.A.H., a NASA Astrobiology Science and Technology for Exploring Planets (ASTEP) grant NNX-327 09AB75G and a grant from Deep Carbon Observatory's Deep Life Initiative to J.A.H., the NSF Science and Technology Center for Dark Energy Biosphere Investigations (C-DEBI) to J.A.H., and the Gordon and Betty Moore Foundation grant GBMF3297 to J.A.H. Samples were collected from the Mid-Cayman Rise with the assistance of the captains and crew of the R/V *Atlantis* and R/V *Falkor* as well as ROVs *Jason* and *Nereus*. For Mid-Cayman Rise, ship and vehicle time in 2012 were supported by the NSF-OCE great OCE-1061863 to Chris German and Jeff Seewald and in 2013 by the Schmidt Ocean Institute during cruise FK008-2013 aboard the R/V *Falkor*. Samples collected from Axial Seamount were collected with the assistance of the captains and crew of the R/V *Falkor*, R/V *Thompson*, and R/V *Brown* as well as the ROV *ROPOS* and *Jason* groups, and in 2013 the Schmidt Ocean Institute during cruise FK010-2013 aboard the R/V *Falkor*.

## REFERENCES

- Treangen TJ, Rocha E. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* 7:e1001284. <https://doi.org/10.1371/journal.pgen.1001284>.
- Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299–304. <https://doi.org/10.1038/35012500>.
- Metcalf JA, Funkhouser-Jones LJ, Briley K, Reysenbach A-L, Bordenstein SR. 2014. Antibacterial gene transfer across the tree of life. *Elife* 3:e04266. <https://doi.org/10.7554/eLife.04266>.
- Schönknecht G, Chen W-H, Ternes CM, Barbier GG, Shrestha RP, Stanke M, Bräutigam A, Baker BJ, Banfield JF, Garavito RM, Carr K, Wilkerson C, Rensing SA, Gagneul D, Dickenson NE, Oesterhelt C, Lercher MJ, Weber A. 2013. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science* 339:1207–1210. <https://doi.org/10.1126/science.1231707>.
- Brazelton WJ, Baross JA. 2010. Metagenomic comparison of two Thiomicrospira lineages inhabiting contrasting deep-sea hydrothermal environments. *PLoS One* 5:e13530. <https://doi.org/10.1371/journal.pone.0013530>.
- Welch RA, Burland V, Plunkett G, III, Redford P, Roesch P, Rasko D, Buckles EL, Liou S-R, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HLT, Donnenberg MS, Blattner FR. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci U S A* 99:17020–17024. <https://doi.org/10.1073/pnas.252529799>.
- Tettelin H, Massignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome.” *Proc Natl Acad Sci U S A* 102:13950–13955. <https://doi.org/10.1073/pnas.0506758102>.
- Tettelin H, Riley D, Cattuto C, Medini D. 2008. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* 11:472–477. <https://doi.org/10.1016/j.mib.2008.09.006>.
- Reference deleted.
- Daubin V, Ochman H. 2004. Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. *Genome Res* 14:1036–1042. <https://doi.org/10.1101/gr.2231904>.
- Andreani NA, Hesse E, Vos M. 2017. Prokaryote genome fluidity is

- dependent on effective population size. *ISME J* 11:1719–1721. <https://doi.org/10.1038/ismej.2017.36>.
12. Kimura M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, United Kingdom.
  13. Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T. 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res* 21:599–609. <https://doi.org/10.1101/gr.115592.110>.
  14. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480:241–244. <https://doi.org/10.1038/nature10571>.
  15. Campanaro S, Vezzi A, Vitulo N, Lauro FM, D'Angelo M, Simonato F, Cestaro A, Malacrida G, Bertoloni G, Valle G, Bartlett DH. 2005. Laterally transferred elements and high pressure adaptation in *Photobacterium profundum* strains. *BMC Genomics* 6:122. <https://doi.org/10.1186/1471-2164-6-122>.
  16. Martinez RJ, Wang Y, Raimondo MA, Coombs JM, Barkay T, Sobczyk PA. 2006. Horizontal gene transfer of PIB-type ATPases among bacteria isolated from radionuclide- and metal-contaminated subsurface soils. *Appl Environ Microbiol* 72:3111–3118. <https://doi.org/10.1128/AEM.72.5.3111-3118.2006>.
  17. Meyer JL, Huber JA. 2014. Strain-level genomic variation in natural populations of *Lebetimonas* from an erupting deep-sea volcano. *ISME J* 8:867–880. <https://doi.org/10.1038/ismej.2013.206>.
  18. Feil EJ. 2004. Small change: keeping pace with microevolution. *Nat Rev Microbiol* 2:483–495. <https://doi.org/10.1038/nrmicro904>.
  19. Cohan FM. 1994. The effects of rare but promiscuous genetic exchange on evolutionary divergence in prokaryotes. *Am Nat* 143:965–986. <https://doi.org/10.1086/285644>.
  20. Majewski J, Cohan FM. 1999. Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. *Genetics* 152:1459–1474.
  21. Levin BR. 1981. Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. *Genetics* 99:1–23.
  22. Koch AL. 1974. The pertinence of the periodic selection phenomenon to prokaryote evolution. *Genetics* 77:127–142.
  23. Bendall ML, Stevens SL, Chan L-K, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, Martin J, Pati A, Bushnell B, Froula J, Kang D, Tringe SG, Bertilsson S, Moran MA, Shade A, Newton RJ, McMahon KD, Malmstrom RR. 2016. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J* 10:1589–1601. <https://doi.org/10.1038/ismej.2015.241>.
  24. Cordero OX, Polz MF. 2014. Explaining microbial genomic diversity in light of evolutionary ecology. *Nat Rev Microbiol* 12:263–273. <https://doi.org/10.1038/nrmicro3218>.
  25. Rodríguez-Valera F, Martín-Cuadrado A-B, López-Pérez M. 2016. Flexible genomic islands as drivers of genome evolution. *Curr Opin Microbiol* 31:154–160. <https://doi.org/10.1016/j.mib.2016.03.014>.
  26. Caro-Quintero A, Konstantinidis KT. 2012. Bacterial species may exist, metagenomics reveal. *Environ Microbiol* 14:347–355. <https://doi.org/10.1111/j.1462-2920.2011.02668.x>.
  27. Anderson RE, Reveillaud J, Reddington E, Delmont TO, Eren AM, McDermott JM, Seewald JS, Huber JA. 2017. Genomic variation in microbial populations inhabiting the marine seafloor at deep-sea hydrothermal vents. *Nat Commun* 8:1114. <https://doi.org/10.1038/s41467-017-01228-6>.
  28. Guttman DS, Dykhuizen DE. 1994. Detecting selective sweeps in naturally occurring *Escherichia coli*. *Genetics* 138:993–1003.
  29. Papke RT, Gogarten JP. 2012. Ecology. How bacterial lineages emerge. *Science* 336:45–46. <https://doi.org/10.1126/science.1219241>.
  30. Papke RT, Koenig JE, Rodríguez-Valera F, Doolittle WF. 2004. Frequent recombination in a saltern population of *Halorubrum*. *Science* 306:1928–1929. <https://doi.org/10.1126/science.1103289>.
  31. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, Polz MF, Alm EJ. 2012. Population genomics of early events in the ecological differentiation of bacteria. *Science* 336:48–51. <https://doi.org/10.1126/science.1218198>.
  32. Rodríguez-Valera F, Martín-Cuadrado A-B, Rodríguez-Brito B, Pasić L, Thingstad TF, Rohwer F, Mira A. 2009. Explaining microbial population genomics through phage predation. *Nat Rev Microbiol* 7:828–836. <https://doi.org/10.1038/nrmicro2235>.
  33. Thingstad TF, Frede Thingstad T. 2000. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol Oceanogr* 45:1320–1328. <https://doi.org/10.4319/lo.2000.45.6.1320>.
  34. Huber JA, Mark Welch DB, Morrison HG, Proc SM, Neal PR, Butterfield DA, Sogin ML. 2007. Microbial population structures in the deep marine biosphere. *Science* 318:97–100. <https://doi.org/10.1126/science.1146689>.
  35. Akerman NH, Butterfield DA, Huber JA. 2013. Phylogenetic diversity and functional gene patterns of sulfur-oxidizing seafloor Epsilonproteobacteria in diffuse hydrothermal vent fluids. *Front Microbiol* 4:185. <https://doi.org/10.3389/fmicb.2013.00185>.
  36. Anderson RE, Beltrán MT, Hallam SJ, Baross JA. 2013. Microbial community structure across fluid gradients in the Juan de Fuca Ridge hydrothermal system. *FEMS Microbiol Ecol* 83:324–339. <https://doi.org/10.1111/j.1574-6941.2012.01478.x>.
  37. Takai K, Horikoshi K. 2000. Rapid detection and quantification of members of the archaeal community by quantitative PCR using fluorogenic probes. *Appl Environ Microbiol* 66:5066–5072. <https://doi.org/10.1128/AEM.66.11.5066-5072.2000>.
  38. Nakagawa S, Takai Y, Shimamura S, Reysenbach A-L, Takai K, Horikoshi K. 2007. Deep-sea vent epsilon-proteobacterial genomes provide insights into emergence of pathogens. *Proc Natl Acad Sci U S A* 104:12146–12150. <https://doi.org/10.1073/pnas.0700687104>.
  39. McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, Paul JH. 2010. High frequency of horizontal gene transfer in the oceans. *Science* 330:50. <https://doi.org/10.1126/science.1192243>.
  40. Bèjà O, Suzuki MT, Koonin EV, Aravind L, Hadd A, Nguyen LP, Villacorta R, Amjadi M, Garrigues C, Jovanovich SB, Feldman RA, DeLong EF. 2000. Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol* 2:516–529. <https://doi.org/10.1046/j.1462-2920.2000.00133.x>.
  41. Xie W, Wang F, Guo L, Chen Z, Sievert SM, Meng J, Huang G, Li Y, Yan Q, Wu S, Wang X, Chen S, He G, Xiao X, Xu A. 2011. Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. *ISME J* 5:414–426. <https://doi.org/10.1038/ismej.2010.144>.
  42. Ortmann AC, Suttle CA. 2005. High abundances of viruses in a deep-sea hydrothermal vent system indicates viral mediated microbial mortality. *Deep Sea Res Part I Oceanogr Res Papers* 52:1515–1527. <https://doi.org/10.1016/j.dsr.2005.04.002>.
  43. Anderson RE, Sogin ML, Baross JA. 2014. Evolutionary strategies of viruses, bacteria and archaea in hydrothermal vent ecosystems revealed through metagenomics. *PLoS One* 9:e109696. <https://doi.org/10.1371/journal.pone.0109696>.
  44. Shank TM, Fornari DJ, Von Damm KL, Lilley MD, Haymon RM, Lutz RA. 1998. Temporal and spatial patterns of biological community development at nascent deep-sea hydrothermal vents (9°50'N, East Pacific Rise). *Deep Sea Res Part II Top Stud Oceanogr* 45:465–515. [https://doi.org/10.1016/S0967-0645\(97\)00089-1](https://doi.org/10.1016/S0967-0645(97)00089-1).
  45. Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol* 55:709–742. <https://doi.org/10.1146/annurev.micro.55.1.709>.
  46. Brazelton WJ, Baross JA. 2009. Abundant transposases encoded by the metagenome of a hydrothermal chimney biofilm. *ISME J* 3:1420–1424. <https://doi.org/10.1038/ismej.2009.79>.
  47. Vrijenhoek RC. 2010. Genetic diversity and connectivity of deep-sea hydrothermal vent metapopulations. *Mol Ecol* 19:4391–4411. <https://doi.org/10.1111/j.1365-294X.2010.04789.x>.
  48. Mino S, Makita H, Toki T, Miyazaki J, Kato S, Watanabe H, Imachi H, Watsuji T-O, Nunoura T, Kojima S, Sawabe T, Takai K, Nakagawa S. 2013. Biogeography of *Persephonella* in deep-sea hydrothermal vents of the Western Pacific. *Front Microbiol* 4:107. <https://doi.org/10.3389/fmicb.2013.00107>.
  49. Price MT, Fullerton H, Moyer CL. 2015. Biogeography and evolution of *Thermococcus* isolates from hydrothermal vent systems of the Pacific. *Front Microbiol* 6:968. <https://doi.org/10.3389/fmicb.2015.00968>.
  50. Reveillaud J, Reddington E, McDermott J, Algar C, Meyer JL, Sylva S, Seewald J, German CR, Huber JA. 2016. Seafloor microbial communities in hydrogen-rich vent fluids from hydrothermal systems along the Mid-Cayman Rise. *Environ Microbiol* 18:1970–1987. <https://doi.org/10.1111/1462-2920.13173>.
  51. Meier DV, Pjevac P, Bach W, Hourdez S, Girguis PR, Vidoudez C, Amann R, Meyerdierrks A. 2017. Niche partitioning of diverse sulfur-oxidizing bacteria at hydrothermal vents. *ISME J* 11:1545–1558. <https://doi.org/10.1038/ismej.2017.37>.
  52. Schmidtova J, Hallam SJ, Baldwin SA. 2009. Phylogenetic diversity of



- transition and anoxic zone bacterial communities within a near-shore anoxic basin: Nitinat Lake. *Environ Microbiol* 11:3233–3251. <https://doi.org/10.1111/j.1462-2920.2009.02044.x>.
53. Grote J, Schott T, Bruckner CG, Glockner FO, Jost G, Teeling H, Labrenz M, Jurgens K. 2012. Genome and physiology of a model Epsilonproteobacterium responsible for sulfide detoxification in marine oxygen depletion zones. *Proc Natl Acad Sci U S A* 109:506–510. <https://doi.org/10.1073/pnas.1111262109>.
  54. Galambos D, Anderson RE, Reveillaud J, Huber JA. 2019. Genome-resolved metagenomics and metatranscriptomics reveal niche differentiation in functionally redundant microbial communities at deep-sea hydrothermal vents. *Environ Microbiol* 21:4395–4410. <https://doi.org/10.1111/1462-2920.14806>.
  55. Zhang Y, Sievert SM. 2014. Pan-genome analyses identify lineage- and niche-specific markers of evolution and adaptation in Epsilonproteobacteria. *Front Microbiol* 5:110. <https://doi.org/10.3389/fmicb.2014.00110>.
  56. German CR, Bowen A, Coleman ML, Honig DL, Huber JA, Jakuba MV, Kinsey JC, Kurz MD, Leroy S, McDermott JM, de Lépinay BM, Nakamura K, Seewald JS, Smith JL, Sylva SP, Van Dover CL, Whitcomb LL, Yoerger DR. 2010. Diverse styles of submarine venting on the ultraslow spreading Mid-Cayman Rise. *Proc Natl Acad Sci U S A* 107:14020–14025. <https://doi.org/10.1073/pnas.1009205107>.
  57. McDermott JM, Seewald JS, German CR, Sylva SP. 2015. Pathways for abiotic organic synthesis at submarine hydrothermal fields. *Proc Natl Acad Sci U S A* 112:7668–7672. <https://doi.org/10.1073/pnas.1506295112>.
  58. Fortunato CS, Larson B, Butterfield DA, Huber JA. 2018. Spatially distinct, temporally stable microbial populations mediate biogeochemical cycling at and below the seafloor in hydrothermal vent fluids. *Environ Microbiol* 20:769–784. <https://doi.org/10.1111/1462-2920.14011>.
  59. Opatkiewicz AD, Butterfield DA, Baross JA. 2009. Individual hydrothermal vents at Axial Seamount harbor distinct subseafloor microbial communities. *FEMS Microbiol Ecol* 70:413–424. <https://doi.org/10.1111/j.1574-6941.2009.00747.x>.
  60. Meyer JL, Akerman NH, Proskurowski G, Huber JA. 2013. Microbiological characterization of post-eruption “snowblower” vents at Axial Seamount, Juan de Fuca Ridge. *Front Microbiol* 4:153. <https://doi.org/10.3389/fmicb.2013.00153>.
  61. Butterfield DA, Lilley MD, Huber JA, Roe KK, Embley RE, Baross JA, Massoth GJ. 2004. Mixing, reaction and microbial activity in the sub-seafloor revealed by temporal and spatial variation in diffuse flow vents at Axial Volcano, p 269–289. In Wilcock WSD, Kelley DS, Baross JA, DeLong E, Cary C (ed), *The sub-seafloor biosphere at mid-ocean ridges*. Geophysical Monograph. American Geophysical Union, Washington, DC.
  62. McHugh JP, Rodríguez-Quinoñes F, Abdul-Tehrani H, Svistunenko DA, Poole RK, Cooper CE, Andrews SC. 2003. Global iron-dependent gene regulation in *Escherichia coli*. A new mechanism for iron homeostasis. *J Biol Chem* 278:29478–29486. <https://doi.org/10.1074/jbc.M303381200>.
  63. Vernikos G, Medini D, Riley DR, Tettelin H. 2015. Ten years of pangenome analyses. *Curr Opin Microbiol* 23:148–154. <https://doi.org/10.1016/j.mib.2014.11.016>.
  64. Hamilton TL, Jones DS, Schaperdoth I, Macalady JL. 2014. Metagenomic insights into S(0) precipitation in a terrestrial subsurface lithoautotrophic ecosystem. *Front Microbiol* 5:756. <https://doi.org/10.3389/fmicb.2014.00756>.
  65. Pjevac P, Kamysnyy A, Jr, Dyksma S, Mussmann M. 2014. Microbial consumption of zero-valence sulfur in marine benthic habitats. *Environ Microbiol* 16:3416–3430. <https://doi.org/10.1111/1462-2920.12410>.
  66. Vieira G, Sabarly V, Bourguignon P-Y, Durot M, Le Fèvre F, Mornico D, Vallenet D, Bouvet O, Denamur E, Schachter V, Médigue C. 2011. Core and panmetabolism in *Escherichia coli*. *J Bacteriol* 193:1461–1472. <https://doi.org/10.1128/JB.01192-10>.
  67. Delmont TO, Murat Eren A. 2018. Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ* 6:e4320. <https://doi.org/10.7717/peerj.4320>.
  68. Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, Steglich C, Church GM, Richardson P, Chisholm SW. 2007. Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. *PLoS Genet* 3:e231. <https://doi.org/10.1371/journal.pgen.0030231>.
  69. Anderson RE, Kouris A, Seward CH, Campbell KM, Whitaker RJ. 2017. Structured populations of Sulfobolus acidocaldarius with susceptibility to mobile genetic elements. *Genome Biol Evol* 9:1699–1710. <https://doi.org/10.1093/gbe/evx104>.
  70. Cuadros-Orellana S, Martin-Cuadrado A-B, Legault B, D’Auria G, Zhaxybayeva O, Papke RT, Rodriguez-Valera F. 2007. Genomic plasticity in prokaryotes: the case of the square haloarchaeon. *ISME J* 1:235–245. <https://doi.org/10.1038/ismej.2007.35>.
  71. Stewart LC, Algar CK, Fortunato CS, Larson BI, Vallino JJ, Huber JA, Butterfield DA, Holden JF. 2019. Fluid geochemistry, local hydrology, and metabolic activity define methanogen community size and composition in deep-sea hydrothermal vents. *ISME J* 13:1711–1721. <https://doi.org/10.1038/s41396-019-0382-3>.
  72. Martiny AC, Coleman ML, Chisholm SW. 2006. Phosphate acquisition genes in Prochlorococcus ecotypes: evidence for genome-wide adaptation. *Proc Natl Acad Sci U S A* 103:12552–12557. <https://doi.org/10.1073/pnas.0601301103>.
  73. Coleman ML, Chisholm SW. 2010. Ecosystem-specific selection pressures revealed through comparative population genomics. *Proc Natl Acad Sci U S A* 107:18634–18639. <https://doi.org/10.1073/pnas.1009480107>.
  74. Sunda WG. 2012. Feedback interactions between trace metal nutrients and phytoplankton in the ocean. *Front Microbiol* 3:204. <https://doi.org/10.3389/fmicb.2012.00204>.
  75. Conkright ME, Gregg WW, Levitus S. 2000. Seasonal cycle of phosphate in the open ocean. *Deep Sea Res Part I Oceanogr Res Papers* 47:159–175. [https://doi.org/10.1016/S0967-0637\(99\)00042-4](https://doi.org/10.1016/S0967-0637(99)00042-4).
  76. Wu J, Sunda W, Boyle EA, Karl DM. 2000. Phosphate depletion in the western North Atlantic Ocean. *Science* 289:759–762. <https://doi.org/10.1126/science.289.5480.759>.
  77. Wheat CG, Feely RA, Mottl MJ. 1996. Phosphate removal by oceanic hydrothermal processes: an update of the phosphorus budget in the oceans. *Geochim Cosmochim Acta* 60:3593–3608. [https://doi.org/10.1016/0016-7037\(96\)00189-5](https://doi.org/10.1016/0016-7037(96)00189-5).
  78. Sanders JG, Windom HL. 1980. The uptake and reduction of arsenic species by marine algae. *Estuar Coast Mar Sci* 10:555–567. [https://doi.org/10.1016/S0302-3524\(80\)80075-2](https://doi.org/10.1016/S0302-3524(80)80075-2).
  79. Smedley PL, Kinniburgh DG. 2002. A review of the source, behaviour and distribution of arsenic in natural waters. *Appl Geochem* 17:517–568. [https://doi.org/10.1016/S0883-2927\(02\)00018-5](https://doi.org/10.1016/S0883-2927(02)00018-5).
  80. Douville E, Charlou JL, Donval JP, Hureau D, Appriou P. 1999. As and Sb behaviour in fluids from various deep-sea hydrothermal systems. *Earth Planet Sci Lett* 328:97–104. [https://doi.org/10.1016/S1521-8050\(99\)80004-4](https://doi.org/10.1016/S1521-8050(99)80004-4).
  81. Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. *PLoS Genet* 4:e1000304. <https://doi.org/10.1371/journal.pgen.1000304>.
  82. Cooper VS, Vohr SH, Wrocklage SC, Hatcher PJ. 2010. Why genes evolve faster on secondary chromosomes in bacteria. *PLoS Comput Biol* 6:e1000732. <https://doi.org/10.1371/journal.pcbi.1000732>.
  83. Bohlin J, Eldholm V, Pettersson JHO, Brynildsrud O, Snipen L. 2017. The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. *BMC Genomics* 18:151. <https://doi.org/10.1186/s12864-017-3543-7>.
  84. Mosquera-Rendón J, Rada-Bravo AM, Cárdenas-Brito S, Corredor M, Restrepo-Pineda E, Benítez-Páez A. 2016. Pangenome-wide and molecular evolution analyses of the *Pseudomonas aeruginosa* species. *BMC Genomics* 17:45. <https://doi.org/10.1186/s12864-016-2364-4>.
  85. Cohan FM, Perry EB. 2007. A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol* 17:R373–R386. <https://doi.org/10.1016/j.cub.2007.03.032>.
  86. Takeuchi N, Cordero OX, Koonin EV, Kaneko K. 2015. Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. *BMC Biol* 13:20. <https://doi.org/10.1186/s12915-015-0131-7>.
  87. Eren AM, Vineis JH, Morrison HG, Sogin ML. 2013. A filtering method to generate high quality short reads using illumina paired-end technology. *PLoS One* 8:e66643. <https://doi.org/10.1371/journal.pone.0066643>.
  88. Peng Y, Leung HCM, Yiu SM, Chin F. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28:1420–1428. <https://doi.org/10.1093/bioinformatics/bts174>.
  89. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25. <https://doi.org/10.1186/gb-2009-10-3-r25>.
  90. Murat Eren A, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi’o: an advanced analysis and visualization platform for ‘omics data. *PeerJ* 3:e1319. <https://doi.org/10.7717/peerj.1319>.
  91. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. <https://doi.org/10.1186/1471-2105-11-119>.

92. Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol* 7:e1002195. <https://doi.org/10.1371/journal.pcbi.1002195>.
93. Campbell JH, O'Donoghue P, Campbell AG, Schwientek P, Sczyrba A, Woyke T, Söll D, Podar M. 2013. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc Natl Acad Sci U S A* 110:5540–5545. <https://doi.org/10.1073/pnas.1303090110>.
94. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu W-T, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499: 431–437. <https://doi.org/10.1038/nature12352>.
95. Darling AE, Jospin G, Lowe E, Matsen FA, IV, Bik HM, Eisen JA. 2014. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* 2:e243. <https://doi.org/10.7717/peerj.243>.
96. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>.
97. Benedict MN, Henriksen JR, Metcalf WW, Whitaker RJ, Price ND. 2014. ITEP: an integrated toolkit for exploration of microbial pan-genomes. *BMC Genomics* 15:8. <https://doi.org/10.1186/1471-2164-15-8>.
98. Van Dongen S. 2008. Graph clustering via a discrete uncoupling process. *SIAM J Matrix Anal Appl* 30:121–141. <https://doi.org/10.1137/040608635>.
99. R Core Team. 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.