Habitat-specificity in SAR11 is associated with a handful of genes under high selection

Sarah J. Tucker[1,2,3,4,5], Kelle C. Freel[1], A. Murat Eren[3,4,5,6,7,#,*], Michael S. Rappé[1,#,*]

[1]Hawai'i Institute of Marine Biology, School of Ocean and Earth Science and Technology, University of Hawai'i at Mānoa, Kāne'ohe, HI 96744, USA; [2]Marine Biology Graduate Program, University of Hawai‘i at Mānoa, Honolulu, Hawai‘i, United States; [3]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543, USA; [4]Helmholtz Institute for Functional Marine Biodiversity, 26129 Oldenburg, Germany; [5]Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, 27570 Bremerhaven, Germany; [6]Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, 26129 Oldenburg, Germany; [7]Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany

[#]Co-senior Authors

*Corresponding Authors. Email: meren@hifmb.de, rappe@hawaii.edu

Running Title: Ecotype differentiation in SAR11

Abstract

The order *Pelagibacterales* (SAR11) is the most abundant group of heterotrophic bacteria in the global surface ocean, where individual sublineages likely play distinct roles in oceanic biogeochemical cycles. Yet, understanding the determinants of niche partitioning within SAR11 has been a formidable challenge due to the high genetic diversity within individual SAR11 sublineages and the limited availability of high-quality genomes from both cultivation and metagenomic reconstruction. Here, we take advantage of 71 new SAR11 genomes from strains we isolated from the tropical Pacific Ocean to evaluate the distribution of metabolic traits across the *Pelagibacteraceae,* a recently classified family within the order *Pelagibacterales* encompassing subgroups Ia and Ib. Our analyses of metagenomes generated from stations where the strains were isolated reveals distinct habitat preferences across SAR11 genera for coastal or offshore environments, and subtle but systematic differences in metabolic potential that support these observations. We also observe higher levels of selective forces acting on habitat-specific metabolic genes linked to SAR11 fitness and polyphyletic distributions of habitat preferences and metabolic traits across SAR11 subclades, suggesting that contrasting lifestyles have emerged across multiple lineages independently. Together, these insights reveal niche-partitioning within sympatric and parapatric populations of SAR11 and demonstrate that the immense genomic diversity of SAR11 bacteria naturally segregates into ecologically and genetically cohesive units, or ecotypes, that vary in spatial distributions in the tropical Pacific.

Main

Bacterial populations typically harbor extensive variation in gene content that can be difficult to delineate into discrete ecological or evolutionary units, despite the significant impact such differences may have on the biogeochemical roles, ecological interactions, and fitness of these microorganisms[1–3]. Comparative genomics has enabled the identification of gene content variation that underpins distinct ecologies and contributes to the process of niche diversification[4–6]. However, because these analyses most often use a collection of genomes isolated from diverse environments where genes have evolved under a range of selective pressures[7], associating variations in gene content with specific forces of selection and identifying crucial genes that likely serve as determinants of fitness under specific conditions remains a formidable challenge.

The bacterial order *Pelagibacterales* (SAR11) is one of the most abundant and diverse bacterial groups in the surface oceans[8–10], where even a single SAR11 population can exhibit substantial sequence diversity in the environment[11]. Defining the genomic variation underlying distinct spatial distributions of SAR11 populations and how gene content differs across environmental conditions is vital to further understanding ocean biogeochemical cycling, eco-evolutionary relationships, and the biology of a ubiquitous and dominant lineage in the surface oceans. This task remains difficult because assessing the loss or gain of ecologically-relevant genes or metabolic pathways in SAR11 necessitates a robust collection of high quality genome sequences that represent a spectrum of both ecological and genomic diversity. SAR11 cells are historically recalcitrant to laboratory cultivation, and as such only a small portion of the total diversity of naturally occurring SAR11 populations can be interrogated

via high quality genomes such as those from isolated strains[12]. Yet even with a small set of isolate genomes, SAR11 has become a key model to help elucidate fundamental processes of ecology and evolution such as genome streamlining[13], marine oligotrophy [14], co-evolution[15,16], drivers of genetic diversity and evolution within organisms of large population sizes[11,17], structure-aware investigations of microbial population genetics[18], and ocean biogeochemical cycling[19–21].

In our companion work[22], we address the gap in the available genomic resources with novel SAR11 isolates that originate from a geographically constrained environment within the tropical Pacific, and develop a phylogenomic backbone that unites evolution with ecology to understand this diverse clade. Freel *et al.*,[22] delineated four families within SAR11, including the family *Pelagibacteraceae* which encompasses the historical SAR11 subgroups Ia and Ib. Here, we illuminate metabolic diversity found within *Pelagibacteraceae* and identify the determinants of niche differentiation within co-occurring *Pelagibacteraceae* populations by combining 71 of the new isolates with 21 previously sequenced *Pelagibacteraceae* isolate genomes. By taking advantage of metagenomes from the prominent source of isolation, we apply an integrated 'omics analysis to unveil differences in *Pelagibacteraceae* metabolic features across coastal and offshore habitats and identify metabolic genes under high selective pressures.

Results

Ecology and evolution of *Pelagibacteraceae* reveals strong habitat-specificity

83 By analyzing 92 high quality genomes from isolates obtained predominantly from coastal
84 Kāneʻohe Bay, Oʻahu, Hawaiʻi, and the adjacent offshore (**Fig. 1, Supplementary file 1a**), we
85 reveal that the genomic diversity of cultured *Pelagibacteraceae* partitions into eleven discrete
86 monophyletic clusters, seven of which include isolates from the Kāneʻohe Bay Time-series
87 (KByT) (**Fig. 1, Extended Data Fig. 1a**). The 92 genomes belong to both previously-described
88 and new clades, which we recently described as eleven genera[22] (**Fig. 1**, **Supplementary file 1a**).
89 The within-genus genome-wide average nucleotide identity (gANI) differed among genera, with
90 the lowest within-genus average gANI found in *Proprepelagibacter* (87.1±0.0%, mean ± sd) and
91 the highest within-genus average gANI found in *Littoralipelagibacter* (95.8±0.0;
92 **Supplementary file 1b**). The average of all within-genus genome-wide average nucleotide
93 identities was 92.3±2.7%.
94 We investigated the biogeography of *Pelagibacteraceae* isolate genomes through read
95 recruitment with metagenomes from KByT and Station ALOHA in the adjacent North Pacific
96 Subtropical Gyre. This revealed that the seven genera within the *Pelagibacteraceae* that
97 contained KByT isolate genomes were commonly detected within the KByT system, while the
98 three genera not containing KByT isolates were rarely detected, if at all (**Extended Data Fig.
99 1b**). Congruent with k-means cluster analysis that grouped the biogeochemical parameters from
100 the metagenomic samples into coastal and offshore clusters (**Extended Data Fig. 3,
101 Supplementary file 1c, Supplementary Information Note 1**), the patterns of genome
102 distribution revealed a clear dichotomy where each genus was either highly present in the
103 offshore or in the coastal environment, but not both (**Fig. 1c, Supplementary file 1d**). Genera
104 with higher abundances in the coastal environment, hereafter referred to as ‘coastal genera’, did

105 not occur within a single monophyletic clade and instead appeared to have evolved multiple
106 times across the evolutionary history of SAR11 (**Fig. 1c**).

107 For the most part, the patterns of distribution were consistent for all genomes within the
108 same genus (**Extended Data Fig. 3**), and genomes that belonged to the same genus generally
109 recruited similar proportions of reads from the environment. One exception was the genome
110 HIMB1412 from the genus *Coralipelagibacter* which had the highest relative abundance among
111 all *Pelagibacteraceae* genomes in the coastal environment (19.9±9.4%, mean ± sd, **Extended**
112 **Data Fig. 3**). While we consistently detected all isolate genomes of *Xanthipelagibacter* in the
113 offshore environment, three of eight genomes (HIMB83, FZCC0015, HIMB1456) were also
114 frequently detected in the coastal environment (**Extended Data Fig. 3**). Regardless, their relative
115 abundances were low in coastal samples (0.4±0.5%, mean ± sd) and sharply increased offshore
116 (3.5±2.0%, mean ± sd; **Extended Data Fig. 3**), further supporting the characterization of
117 *Xanthipelagibacter* as an offshore genus.

118 Overall, our metagenomic read recruitment reveals that *Pelagibacteraceae* communities
119 partition between coastal and offshore habitats in the tropical Pacific, that these distinct
120 communities are driven by differences in the distribution of individual *Pelagibacteraceae*
121 genera, and that coastal genera (*Ampluspelagibacter*, *Coralipelagibacter*, *Littoralipelagibacter*)
122 are distributed polyphyletically across the *Pelagibacteraceae*.

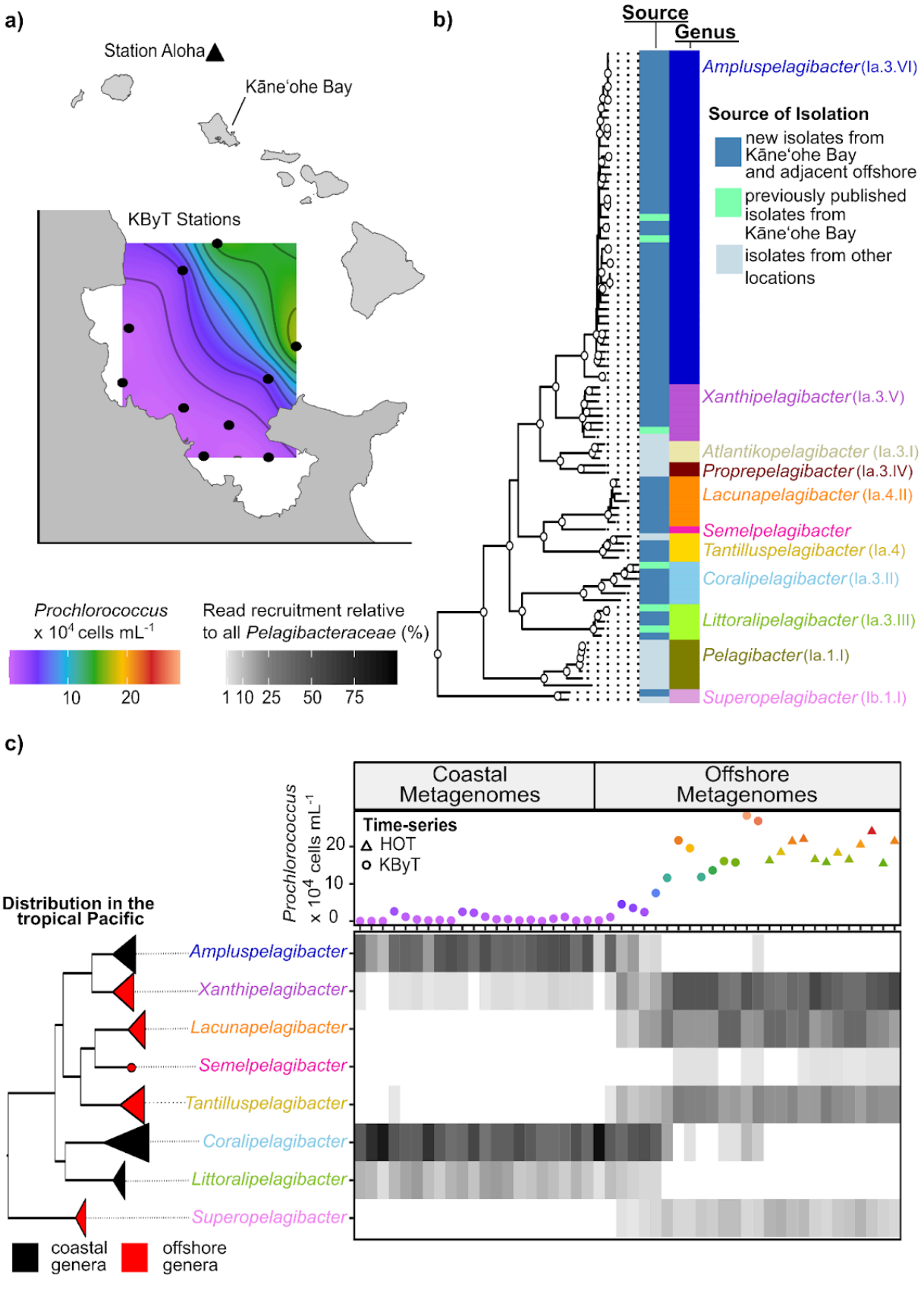

a)
Station Aloha
Kāneʻohe Bay
KByT Stations
Prochlorococcus
x 10^4 cells mL^-1
10
20
Read recruitment relative
to all Pelagibacteraceae (%)
1 10 25 50 75
b)
Source
Genus
Ampluspelagibacter (Ia.3.VI)
Source of Isolation
new isolates from Kāneʻohe Bay and adjacent offshore
previously published isolates from Kāneʻohe Bay
isolates from other locations
Xanthipelagibacter (Ia.3.V)
Atlantikopelagibacter (Ia.3.I)
Proprepelagibacter (Ia.3.IV)
Lacunapelagibacter (Ia.4.II)
Semelpelagibacter
Tantilluspelagibacter (Ia.4)
Coralipelagibacter (Ia.3.II)
Littoralipelagibacter (Ia.3.III)
Pelagibacter (Ia.1.I)
Superopelagibacter (Ib.1.I)
c)
Coastal Metagenomes
Offshore Metagenomes
Prochlorococcus
x 10^4 cells mL^-1
20
10
0
Time-series
HOT
KByT
Distribution in the tropical Pacific
Ampluspelagibacter
Xanthipelagibacter
Lacunapelagibacter
Semelpelagibacter
Tantilluspelagibacter
Coralipelagibacter
Littoralipelagibacter
Superopelagibacter
coastal genera
offshore genera

Fig. 1. ***Pelagibacteraceae* show polyphyletic habitat preferences for coastal or offshore environments across KByT. a**) The location of the Kāneʻohe Bay Time-series (KByT) sampling stations and the Hawaii Ocean Time-series at Station ALOHA. KByT spans a steep biogeochemical gradient, as noted by the dramatic increase in *Prochlorococcus* cellular abundances immediately offshore. **b**) Eleven genera are represented by genome-sequenced *Pelagibacteraceae* isolates, with the majority isolated from coastal Kāneʻohe Bay and the adjacent offshore. Nodes with circles represent ≥90% bootstrap support. Historically referenced clade names for *Pelagibacteraceae* genera are provided in parentheses. **c**) The relative read recruitment of *Pelagibacteraceae* isolate genomes summed at the level of genera for metagenomes from KByT and Station ALOHA. Metagenomes grouped into two clusters by environmental parameters, coastal and offshore, coinciding with the gradient in *Prochlorococcus* cellular abundances. *Pelagibacteraceae* genera showed distinct distribution patterns in the coastal and offshore waters that are distributed polyphyletically. Read recruitment <1% not shown.

Pelagibacteraceae pangenome includes differentially enriched genes and functions within coastal and offshore genera

Next, we sought to characterize the genomic diversity and functional gene content of *Pelagibacteraceae* isolates in order to investigate the potential determinants of fitness that may explain the distinct ecological patterns of distribution. We utilized a pangenomic approach that partitioned all genes across all *Pelagibacteraceae* genomes into *de novo* gene families, or ‘gene clusters’, based on amino acid sequence homology. This analysis resulted in 8,242 gene clusters across all 92 genomes, of which 784 were core among all *Pelagibacteraceae* and made up the majority (52-63%) of gene content found in any individual genome (**Fig. 2a; Supplementary file 2a**). Half of the pangenome was composed of gene clusters that were singletons (4,201 of 8,242; **Fig. 2a**). Between the extremes of gene clusters that occur in every genome and those that occur only in one, we found 206 accessory gene clusters that were ‘genus-specific’ (21±27, mean±sd per genus; **Fig. 2a**; **Supplementary file 2b**) as they were present among all representatives of a single genus, but absent from genomes belonging to any other genus. The 206 genus-specific gene clusters were not distributed uniformly across genera, as 95 were unique

155 to the genus *Superopelagibacter* alone (**Supplementary file 2b**). However, this number is likely
156 an overestimation of the true *Superopelagibacter* genus-specific traits since this particular genus
157 was represented by only two genomes, and the number of genus-specific gene-clusters
158 negatively correlated with the number of genomes in a genus (r^2adj = 0.37, F=6.29, p = 0.037;
159 Pearson's correlation). Gene clusters that were core to multiple genera (e.g., present in every
160 genome for two or more genera, but absent in every other) were also not a dominant feature in
161 the pangenome (Fig. 2a). They only represented 182 gene clusters, and were generally uniformly
162 distributed across many different groupings of genera (**Fig. 2a**; **Supplementary file 2b**). We
163 could further assign a subset of the gene clusters to coastal or offshore categories based on the
164 environmental preference of the genomes in which they occurred, finding 32 and 125 gene
165 clusters that were specific to coastal genera and offshore genera, respectively. (**Fig. 2a,**
166 **Supplementary file 2b**).

167 We next focused on genes with well-characterized roles in known metabolisms by
168 examining those that share orthology with genes described in the Kyoto Encyclopedia of Genes
169 and Genomes (KEGG)[23] through KOfam models[24] and generating a 'functional pangenome'
170 (**Fig. 2b**). The *Pelagibacteraceae* functional pangenome grouped genes based on functional
171 identity rather than amino acid sequence similarity and consisted of 1,276 KOfam functions,
172 with a large core (612 KOfams; **Fig. 2b**). Many of the remaining non-core functions were still
173 shared widely across all eleven genera with a relatively uniform distribution (n=136; **Fig. 2b**) or
174 were singletons (n=70). Similar to the conventional pangenome, the functional pangenome
175 contained very few genus-specific functions (14 total among all genera; 0-5 KOfams per genus;
176 **Fig. 2b**) and a limited number of functions that were core to multiple genera (e.g. present in
177 every genome for two or more genera, but absent in every other; 32 KOfams total; **Fig. 2b;**

178 **Supplementary file 2c**). Only three KOfam functions were specific to coastal genera, while 13
179 KOfam functions were specific to offshore genera (**Fig. 2a; Supplementary file 2c**).

180 Pangenomic analysis using conventional and functional approaches revealed a large
181 overlap in gene content among ecologically distinct genera, with subtle but consistent
182 genus-specific features. The high overlap in gene content across the *Pelagibacteraceae* and small
183 number genus-specific gene clusters suggests differentiation between genera of the
184 *Pelagibacteraceae* may be determined by only a small portion of the genome. The polyphyletic
185 distribution of habitat preference (**Fig. 1**) supports that genes and functions that are linked to
186 success in these different habitats likely evolved multiple times through independent processes.
187 Given this evolutionary trajectory, it is highly plausible that instead of vertical inheritance,
188 horizontal gene transfer led to the acquisition of most of these habitat-specific genes and
189 subsequent high rates of recombination and strong selective forces resulted in the increased
190 prevalence among *Pelagibacteraceae* in a given ecological niche. This process, referred to as
191 gene-specific sweeps, frequently results in a more subtle display of habitat-specific traits[25] as is
192 observed here. To further examine the potential importance of these environmentally-distinct
193 functions in providing advantages to *Pelagibacteraceae* prevalent in coastal or offshore habitats,
194 we next focused on the subset of the environmentally-distinct functions characterized here that
195 are involved in nutrient resource utilization and cellular stress. Furthermore, in examining the
196 functions involved in nutrient resource utilization and cellular stress, we broadened our
197 investigation beyond the functions that were strictly found in coastal genera or strictly found in
198 offshore genera (e.g. environmentally-distinct functions), to also include functions that were
199 enriched in either coastal or offshore genera.

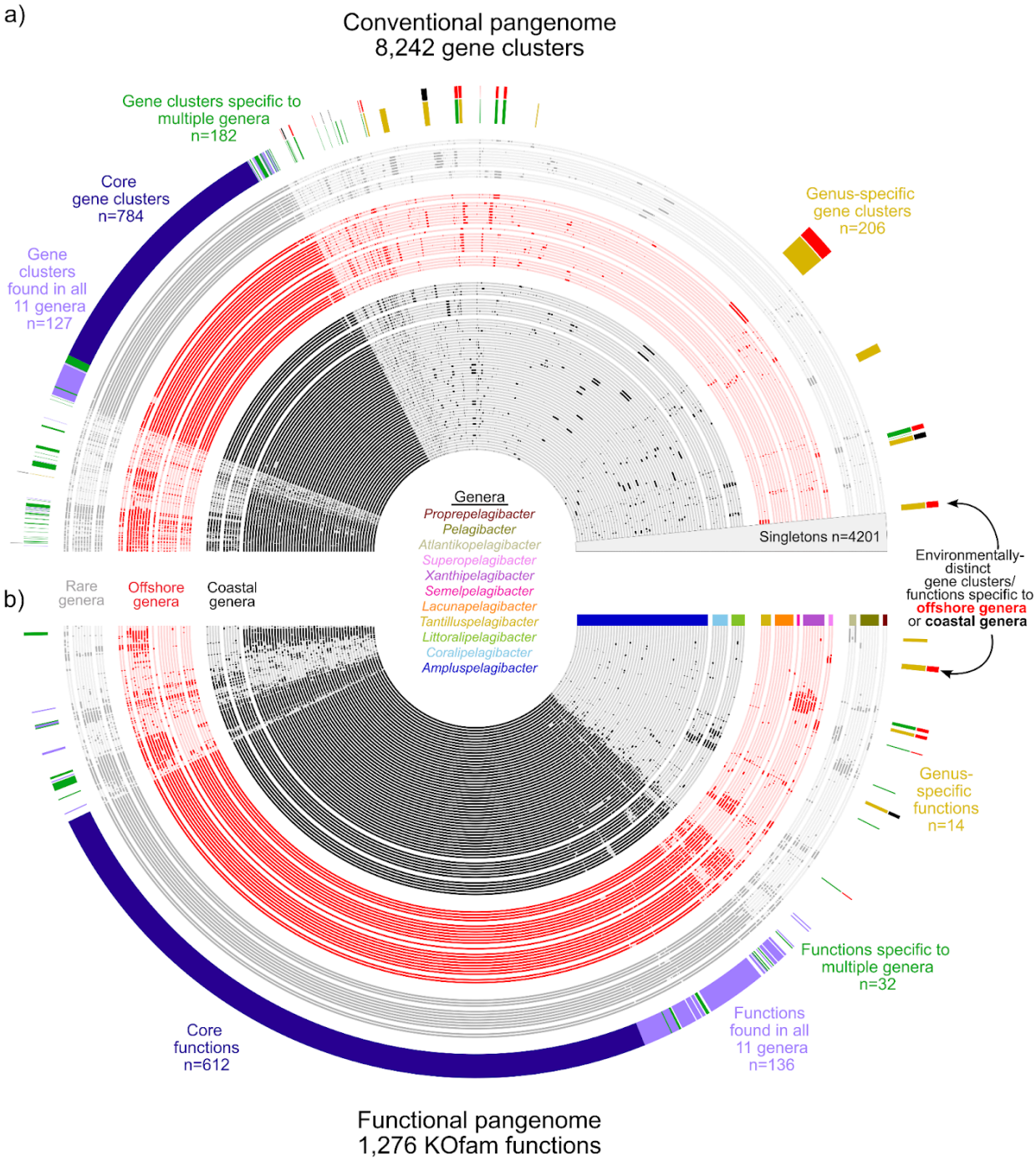


Fig 2. *Pelagibacteraceae* genomes share high overlap in genes and functions among ecologically-distinct groups. The presence and absence of gene clusters and KOfams are shown across **a)** a conventional pangenome and **b)** a functional pagenome, respectively. Genomes (inner rings) are grouped by genus and further grouped by ecological distribution in the tropical Pacific as coastal, offshore, or rare. The distribution of the gene pool and functions in the second most outer layer (e.g., core, genus-specific) show that the gene pool and functions are shared

extensively across all 11 genera and 92 genomes. The outer layer shows the sparse environmentally-distinct portions of the gene pool and functional pool that are specific to coastal or offshore genera.

Potential determinants of habitat-specificity in *Pelagibacteraceae* include molybdenum utilization and nitrogen and carbon metabolisms

Despite a high amount of water exchange, coastal Kāneʻohe Bay and the adjacent offshore vary in nutrient availability, osmotic conditions (e.g. salinity, temperature, pH), and phytoplankton communities that determine the organic carbon pool[26,27], suggesting that this would enforce distinct environmental pressures on coastal and offshore *Pelagibacteraceae* genera. To characterize ecologically-relevant genes that may support coastal and offshore habitat-preferences of *Pelagibacteraceae* across the KByT system, we used an enrichment analysis to examine metabolic traits related to nutrient acquisition and cellular stress that were differentially distributed across genomes from either coastal or offshore genera. Genes that were core across genera of the *Pelagibacteraceae* or did not show easily distinguishable distributions (e.g. driven by phylogeny or environment) are further discussed in **Supplementary Information Note 2**. Unless noted otherwise, the following metabolic traits were found outside of hyper-variable regions of genomes and thus represent genes that are encoded within the relatively stable genomic backbones of a given genus or set of genera.

Molybdenum enzyme utilization.

Molybdenum enzymes are nearly ubiquitous among organisms from all domains of life, catalyze important oxidation-reduction reactions involved in carbon, sulfur, and nitrogen

230 metabolisms, and require a cofactor scaffold to hold the molybdenum in place[28]. While most
231 *Pelagibacteraceae* genomes contained genes involved in molybdenum cofactor biosynthesis as
232 well as molybdenum cofactor-dependent enzymes (**Fig. 3a, Supplementary file 2d**), these genes
233 were missing from two offshore genera, *Lacunapelagibacter* and *Tantilluspelagibacter*. In
234 contrast to the lack of molybdenum enzymes found in the offshore *Lacunapelagibacter* and
235 *Tantilluspelagibacter*, other offshore genera, *Xanthipelagibacter* and *Semelpelagibacter*,
236 appeared to increase reliance on molybdenum-dependent enzymes by uniquely harboring genes
237 encoding molybdenum-dependent enzymes associated with the oxidation of purines (xanthine;
238 *xdhAB*) (**Fig. 3a, Supplementary file 2d, Supplementary Note 3**). The distribution of
239 molybdenum enzymes are polyphyletically distributed among offshore genera, where some
240 offshore groups possess no molybdenum enzymes while others harbor all the molybdenum
241 enzymes detected within *Pelagibacteraceae* (**Fig. 3a**). We hypothesize that the absence of
242 molybdenum enzymes within the offshore *Lacunapelagibacter* and *Tantilluspelagibacter*, which
243 have some of the smallest genomes within *Pelagibacteraceae* (~1.2Mbp; **Supplemental file 1a**),
244 is likely driven by streamlining selection to minimize metabolic requirements and genome
245 size[13,29].

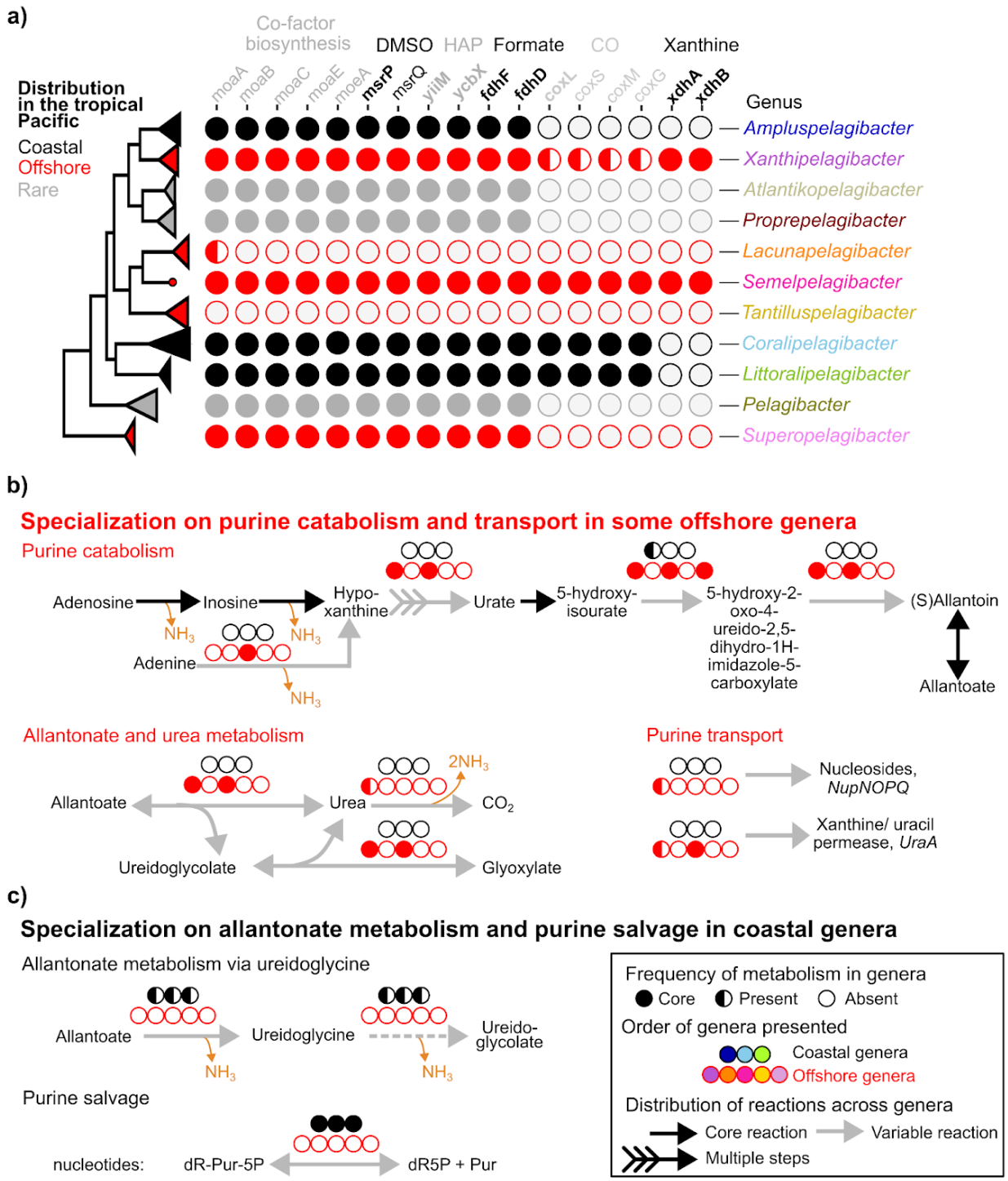


Fig. 3. *Pelagibacteraceae* genera fill distinct niches in molybdenum enzyme utilization. a) The distribution of molybdenum cofactor (MoCo) biosynthesis genes and genes involved in metabolisms that require a molybdenum cofactor across *Pelagibacteraceae* genera. Metabolisms within *Pelagibacteraceae* that require a molybdenum cofactor included those involved in the repair of methionine sulfoxides and the catalytic subunit of periplasmic dimethylsulfoxide

(DMSO; *msrPQ*), the detoxification of 6-N-hydroxylaminopurine (HAP) to adenine (*yiiM*, *ycbX*), the oxidation of formate to carbon dioxide (*fdhFD*), the oxidation of carbon monoxide (*coxSMLG*), and the oxidation of purines (*xdhAB*). Genes that are predicted to represent proteins with molybdenum cofactor binding sites are noted in bold. **b)** Genes encoding for purine catabolism and transport are only present in some offshore genera**. c)** Genes encoding for allantonate metabolism and purine salvage are present only in coastal genera. MoCo: molybdenum cofactor, DMSO: dimethylsulfoxide, HAP: hydroxylaminopurine, CO: carbon monoxide, NH_3: ammonia, dR-Pur-5P: purine 2'-deoxyribonucleoside 5'-monophosphate, Pur: purine base, dRib5P: 2'-deoxyribonucleoside 5'-monophosphate, (d)R-Pur-1P: beta-(deoxy)ribonucleosides.

Nitrogen metabolisms.

Purines are abundant in aquatic environments[30], contain a high number of nitrogen atoms per molecule[31], and likely contribute to niche-differentiation between coastal and offshore SAR11[32]. Our metabolic reconstructions show that offshore genera *Xanthipelagibacter* and *Semelpelagibacter* harbored genes to catabolize purines through the utilization of a molybdenum cofactor-dependent xanthine dehydrogenase (**Fig. 3b, Supplementary file 2e**), as well as genes with functions predicted to support the efficient transport of nucleotides (*Xanthipelagibacter*: *nupNOPQ* system; *Xanthipelagibacter* and *Semelpelagibacter*: xanthine/uracil permease *uraA*). To make hypoxanthine available for degradation, *Semelpelagibacter* can convert adenine to hypoxanthine using adenine deaminase (*ade*) and both genera *Semelpelagibacter* and *Xanthipelagibacter* also uniquely possess a tRNA(adenine34) deaminase (*tadA*), which can convert adenine to hypoxanthine (**Fig. 3b**). Most *Xanthipelagibacter* genomes (6 of 8) also harbor a urease system (*ureABCDEFG*) to liberate ammonia in the final steps of the xanthine degradation pathway.

In contrast to the purine degradation pathways found within the offshore genera, we found coastal genera specialize on the degradation of allatonate, a bi-product of purine metabolism, as well as the recycling of nucleotides. Members of the coastal genus

280 *Ampluspelagibacter* (and a single *Coralipelagibacter* genome, HIMB5) have the functional
281 capacity to degrade allantoate to (S)-ureiodoglycine and ammonia via an allantoate deiminase
282 and ureidoglycine to ureidoglycolate and ammonia, via a nucleophile (Ntn)-hydrolase (**Fig. 3c,**
283 **Supplementary file 2e**). It is important to note that the allantoin degradation genes found in
284 coastal genomes characterized here appear to provide a source of nitrogen, and differ from those
285 found in the offshore genera. All coastal genera possess a deoxynucleoside 5-monophosphate
286 N-glycosidase (*rcl*) that putatively breaks the N-glycosidic bond of purine nucleotides (i.e.
287 purine 2'-deoxyribonucleoside 5'-monophosphates; **Fig. 3c**). This differs from the purine
288 nucleoside salvage pathway core to both coastal and offshore genera, which requires
289 orthophosphate to cleave the N-glycosidic bond of beta-(deoxy)ribonucleosides to yield
290 alpha-(deoxy)ribose 1-phosphate (**Supplementary file 2e**). All *Pelagibacteraceae* are missing
291 genes to utilize deoxyribose-5-phosphate or deoxyribose-1-phosphate via a
292 deoxyribose-phosphate aldolase (*deoC*). However, ribose-1-phosphate could be catabolized to
293 D-fructose-6P and D-glyceraldehyde-3P (**Supplementary file 2e**).
294 In addition to differences in purine metabolism, offshore genera uniquely harbor a
295 transaminase (*agxt*) that may help these cells to overcome an auxotrophy for the biosynthesis of
296 the amino acid glycine (**Fig. 4, Supplementary file 2f**), which is a common feature across
297 *Pelagibacteraceae*, but likely restricts important biosynthetic pathways including protein
298 synthesis and central carbon metabolism[33]. Offshore genera also uniquely contain genes involved
299 in glutamate deamination, glutathione and carnitine metabolism, and the synthesis of
300 glucosylglycerate, which is used as an alternative osmolyte under nitrogen-limited conditions in
301 cyanobacteria[34]. In contrast, the coastal genera contain nitrogen metabolisms generally involved
302 in the deamination of various amino acids (e.g. hydroxyprolines, proline, ornithine, and

303 threonine). A subset of coastal *Ampluspelagibacter* genomes also have the potential to
304 metabolize the polyamines putrescine and spermidine via *puuBCD* (**Fig. 4**). The puuD gene was
305 located in hypervariable regions within coastal *Ampluspelagibacter* genomes.

306 Phosphate & Metals.

307 Some offshore *Pelagibacteraceae* genera exhibit a reduced genomic capacity to transport
308 phosphate relative to coastal genera. High-affinity phosphate acquisition genes (*pstSCAB-phoU*;
309 *phoB*; *phoR*) were core among coastal genera but, while present in some genomes, were not core
310 among the offshore genera. Offshore genomes lacking this system appear to rely solely on the
311 *pitA* gene, a presumably lower-affinity proton-motive force permease of phosphate or sulfur
312 (**Fig. 4, Supplementary file 2f**). Some offshore genera contain an increased genomic capacity to
313 transport or capture iron relative to coastal genera. While all *Pelagibacteraceae* contain an iron
314 (III) transport system (*afuAB*), the offshore *Semelpelagibacter* genus alone possesses a
315 high-affinity iron permease (*ftr1*) (**Fig. 4, Supplementary file 2f)**. Offshore genera
316 *Superopelagibacter*, *Lacunapelagibacter*, and *Tantilluspelagibacter* also contain a heme oxidase,
317 which could catalyze the oxidative degradation of the heme porphyrin ring to release iron.
318 Unique genes to utilize copper, potassium, and nickel were found among the coastal genera (**Fig.
319 4)**. We observed a copper-containing plastocyanin gene within a small subset of coastal genomes
320 (**Fig. 4; Supplementary file 2f**). Plastocyanins serve as an electron carrier from cytochrome *f* to
321 photosystem I and are generally found within eukaryotic phytoplankton and cyanobacteria[35].
322 Future studies are needed to understand the functional role of plastocyanins in
323 *Pelagibacteraceae*, although one potential use could be the storage of copper to cope with
324 periods of reduced availability[36].

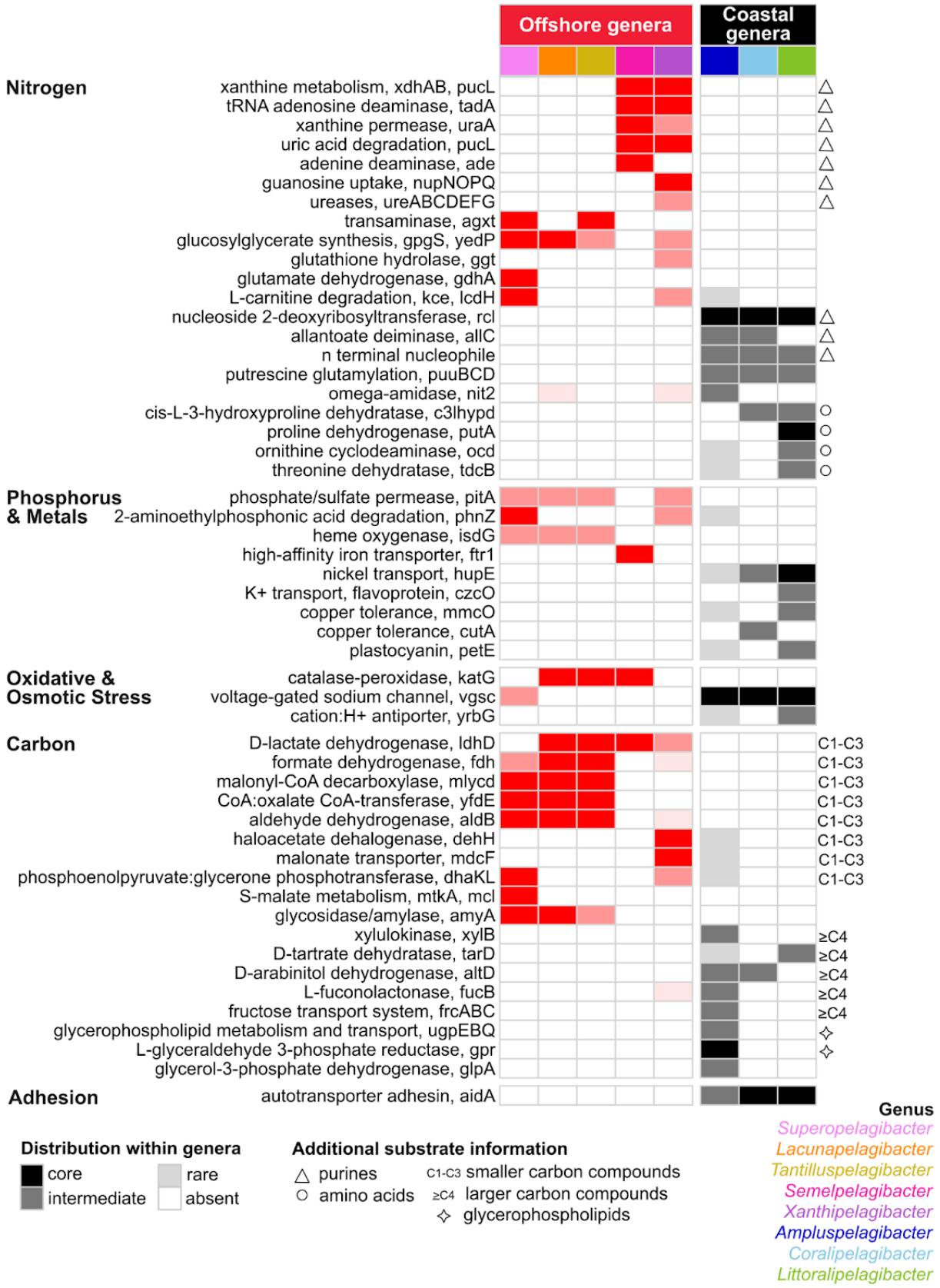
Cellular stress.

Differences in the capacities to respond to osmotic and oxidative stresses were found among offshore and coastal *Pelagibacteraceae* genera. Most offshore genera shared a catalase-peroxidase (*katG*) that may help to cope with oxidative stress and has also been suggested to play a role in the co-evolution of marine microbial communities[15] (**Fig. 4, Supplementary file 2f**). Coastal genera contained two genes involved in osmotic stress that were not prevalent among offshore genomes: a voltage-gated sodium channel (*vgsc*) and a cation:H+antiporter (*yrbG*) (**Fig. 4, Supplementary file 2f**). Voltage-gated sodium channels utilize a structurally complex gating system to physically open and close the pore in response to changes in fluid shear or membrane stretch from dehydration[37,38], and may be particularly useful in coastal Kāneʻohe Bay where spatiotemporal shifts in salinity are common[39].

Carbon metabolisms.

Carbon metabolisms unique to coastal and offshore genera were remarkably distinct. SAR11's specialization on low-molecular weight, labile carbon sources has likely facilitated the success of this marine oligotroph in open oceans where carbon substrates are limited in availability and under high competition[40–42]. Consistent with these observations, the carbon metabolisms shared by offshore genera did not rely on glycolysis and involved low-molecular-weight compounds (one-carbon to three-carbon; C1-C3) that are common molecules and/or typical waste products in the marine environment. This includes the genetic potential to metabolize organic acids such as D-lactate, oxalate, haloacetate, and formate, as well as diverse aldehydes (**Fig. 4**, **Supplementary file 2f**). In contrast, coastal genera contain gene content to specialize in the degradation of relatively larger carbon compounds (C4-C6), including

347 the sugars D-xylulose and D-ribulose via *xylB* and sugar alcohols arabinitol and mannitol via an
348 arabinitol dehydrogenase, both of which subsequently enter glycolysis through the
349 non-phosphorylative Entner–Doudoroff (np-ED) pathway. Some coastal groups also have the
350 capacity to metabolize the sugar acids D-tartrate and malate via *tarD* (**Fig. 4; Supplementary**
351 **file 2f**). Genomes belonging to the coastal genus *Ampluspelagibacter*, which have the largest
352 genomes among known *Pelagibacteraceae*, harbored gene content potentially involved in the
353 transport and metabolism of glycerophospholipids and sn-glycerol-3-phosphate (*glpA*, *upgEBQ*,
354 *gpr*), the capacity to metabolize L-fucose, D-arabinose, and L-xylose to pyruvate through a
355 L-fucono-1,5-lactonase (*fucB*) and the glycolytic Entner-Doudoroff (ED) glycolytic pathway, and
356 a fructose transport system (*frcABC*) (**Fig.4; Supplementary file 2f**). The ED glycolytic
357 pathway was core among *Ampluspelagibacter*, *Xanthipelagibacter*, and *Proprepelagibacter*
358 genomes and variable across genomes in other genera (**Supplementary file 2f**). The increased
359 prevalence of sugar metabolisms among coastal genera, especially the *Ampluspelagibacter*
360 genus, suggests broadened metabolic versatility among coastal *Pelagibacteraceae* in response to
361 a diverse and/or more readily available set of organic carbons found within coastal Kāneʻohe
362 Bay relative to offshore waters. The distinct distributions of carbon metabolisms resolved here
363 also suggests fine-scale niche partitioning along the sugar-acid spectrum within
364 *Pelagibacteraceae*[43].


Offshore genera
Coastal genera
Nitrogen
xanthine metabolism, xdhAB, pucL
tRNA adenosine deaminase, tadA
xanthine permease, uraA
uric acid degradation, pucL
adenine deaminase, ade
guanosine uptake, nupNOPQ
ureases, ureABCDEFG
transaminase, agxt
glucosylglycerate synthesis, gpgS, yedP
glutathione hydrolase, ggt
glutamate dehydrogenase, gdhA
L-carnitine degradation, kce, lcdH
nucleoside 2-deoxyribosyltransferase, rcl
allantoate deiminase, allC
n terminal nucleophile
putrescine glutamylation, puuBCD
omega-amidase, nit2
cis-L-3-hydroxyproline dehydratase, c3lhypd
proline dehydrogenase, putA
ornithine cyclodeaminase, ocd
threonine dehydratase, tdcB
Phosphorus & Metals
phosphate/sulfate permease, pitA
2-aminoethylphosphonic acid degradation, phnZ
heme oxygenase, isdG
high-affinity iron transporter, ftr1
nickel transport, hupE
K+ transport, flavoprotein, czcO
copper tolerance, mmcO
copper tolerance, cutA
plastocyanin, petE
Oxidative & Osmotic Stress
catalase-peroxidase, katG
voltage-gated sodium channel, vgsc
cation:H+ antiporter, yrbG
Carbon
D-lactate dehydrogenase, ldhD
formate dehydrogenase, fdh
malonyl-CoA decarboxylase, mlycd
CoA:oxalate CoA-transferase, yfdE
aldehyde dehydrogenase, aldB
haloacetate dehalogenase, dehH
malonate transporter, mdcF
phosphoenolpyruvate:glycerone phosphotransferase, dhaKL
S-malate metabolism, mtkA, mcl
glycosidase/amylase, amyA
xylulokinase, xylB
D-tartrate dehydratase, tarD
D-arabinitol dehydrogenase, altD
L-fuconolactonase, fucB
fructose transport system, frcABC
glycerophospholipid metabolism and transport, ugpEBQ
L-glyceraldehyde 3-phosphate reductase, gpr
glycerol-3-phosphate dehydrogenase, glpA
Adhesion
autotransporter adhesin, aidA
C1-C3
≥C4
Distribution within genera
core
intermediate
rare
absent
Additional substrate information
purines
amino acids
C1-C3 smaller carbon compounds
≥C4 larger carbon compounds
glycerophospholipids
Genus
Superopelagibacter
Lacunapelagibacter
Tantilluspelagibacter
Semelpelagibacter
Xanthipelagibacter
Ampluspelagibacter
Coralipelagibacter
Littoralipelagibacter

365

Fig. 4. Metabolic determinants of habitat-specificity in *Pelagibacteraceae*. Metabolic genes associated with nutrient acquisition and cellular stress vary among coastal and offshore *Pelagibacteraceae* genera.

Adhesion.

Finally, while investigating the functional differences between coastal and offshore genera, we observed homologs of a large gene (~5,000 to 9,500 bp) annotated as an autotransporter adhesin that was almost always present among coastal genomes and absent in all offshore genomes (**Fig. 4**). Autotransporters provide a simple and relatively minimal mechanism for the delivery of a passenger protein to the surface of Gram-negative bacteria[44], and can vary greatly in size due to changes in the number of repeating sequences in the passenger protein[45]. Because repeating regions cause difficulty for short-read assemblers such as those used in our genome assemblies[22], we further established confidence in the assembly of this gene in our coastal genomes by successfully finding homologs in long-read sequencing libraries from the coastal Kāneʻohe Bay environment and comparing protein structure models of genes annotated from the long-read data to a known autotransporter (**Extended Data Fig. 4; Supplementary Information Note 4**). We also examined the genome context of the gene, which was always positioned next to genes involved in type IV pilus systems and/or type II secretory systems and sometimes, but not always, within hyper-variable regions (**Extended Data Fig. 4, Supplementary Information Note 4, Supplementary file 2f**).

Adhesion to surfaces has not been observed in *Pelagibacteraceae* in culture or through environmental studies, although long filaments believed to be pili were observed within dividing cells of *Pelagibacteraceae*[46] and are associated with multiple processes, including adhesion to surfaces, transformation competence, and DNA uptake[47,48]. While speculative, it is possible that

390 adhesion by the autotransporter adhesin gene may be involved in the uptake of nutrients also in
391 *Pelagibacteraceae*[49]. The coastal distribution of the autotransporter gene suggests that the
392 advantage of having this large of a gene is worth the trade-off of maintenance in more
393 nutrient-rich coastal environments, but potentially not in nutrient-limited offshore environments.
394 In summary, our investigation of the distribution of genes related to nutrient acquisition
395 and cellular stress that are unique to coastal and offshore genera reveals numerous subtle
396 differences in metabolic traits driven by the presence or absence of single genes (e.g. high
397 affinity iron transporters, voltage-gated sodium channel genes), but few differences in multi-gene
398 metabolic pathways (e.g. purine metabolism, molybdenum cofactor biosynthesis, fucose
399 metabolism). Differences in genes related to nutrient acquisition were predominantly related to
400 organic carbon and nitrogen metabolisms, and help to further define the unique and diverse roles
401 *Pelagibacteraceae* play in oceanic biogeochemical cycles. The distribution of these metabolic
402 traits generally do not follow phylogenetic distributions, but environmental patterns, suggesting
403 that these metabolisms may be driven by selective pressures from the environment.
404

405 Gene determinants of habitat-specificity are under relatively higher 406 selective pressures compared to non-diagnostic genes

407 Our analyses indicated a relatively small number of functions in the *Pelagibacteraceae*
408 appear to be associated with its strict niche partitioning. Assuming that the requirement to
409 perform optimally may be higher for genes that determine fitness to particular lifestyles, we
410 hypothesized that metabolic traits that have distinct environmental distributions should be under
411 higher selective pressures compared to other genes in the environment. To test this hypothesis we
412 examined the proportion of non-synonymous to synonymous (pN/pS) sites per gene across

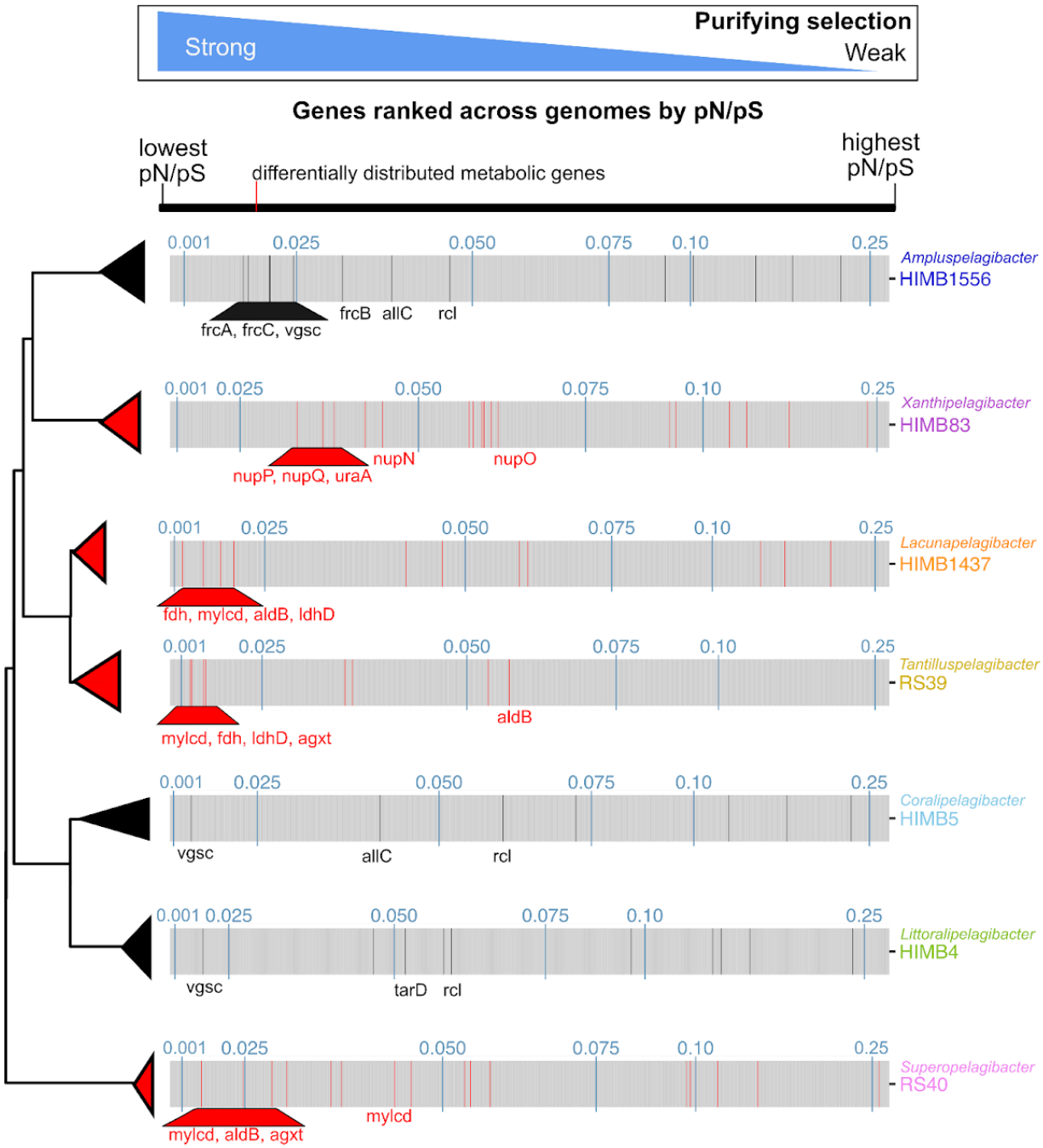
413 *Pelagibacteraceae* genomes using four deeply sequenced metagenomic samples from the KByT
414 system.

415 To define the set of genomes to be included in this analysis, we first evaluated whether
416 the gene-level selective pressures were uniform across genomes from the same genus. We
417 examined variation in pN/pS across genes with shared KOfam assignments for genomes
418 belonging to the genus *Coralipelagibacter,* which has a low minimum gANI (84.5%, n=6) and is
419 abundant in the tropical Pacific. Gene-to-gene variation explained 82% of the variation, while
420 sample-to-sample variation and genome-to-genome variation only explained 0.26 and 0.09%,
421 respectively (ANOVA, **Supplemental file 3a**). The little genome-to-genome variation showed
422 that genomes within the same genus likely experience similar patterns of selection on shared
423 genes, and supported using a single genome as a representative per genus. Thus, our downstream
424 analyses utilized a single genome for each genus in *Pelagibacteraceae*, except
425 *Atlantikopelagibacter*, *Pelagibacter*, *Proprepelagibacter*, and *Semelpelagibacter*, which lacked
426 sufficient coverage in the KByT system (**Extended Data Fig. 1**, **Supplemental file 3b**).

427 Across the *Pelagibacteraceae* genomes examined, strong purifying selection (pN/pS $\ll$ 1)
428 was a general characteristic of most genes as observed previously by Kiefl et al.,[18], where over
429 75% of genes in each genome representative had a pN/pS value of $\leq$ 0.1 (**Supplemental file 3c**).
430 However, not all genes experienced similar magnitudes of purifying selection, with gene-to-gene
431 variation in pN/pS explaining between 88-99% of the variation in pN/pS across genomes (n=7,
432 ANOVA, **Supplemental file 3d**). Sample-to-sample variation in pN/pS values varied for most
433 genomes, however it explained a very small proportion of the overall variation (<0.5%, n=7,
434 ANOVA). Gene coverage showed no correlation between per gene values of pN/pS (absolute
435 value of r=0.02-0.28; Pearson correlation coefficient; **Extended Data Fig. 5**). Gene coverage

436 explained an extremely small portion of the overall pN/pS variation (<0.07%, n=7, ANOVA,
437 **Supplemental file 3d**), which indicates that pN/pS values are unlikely to be driven by artifacts
438 associated with variation in coverage.

439 We next evaluated the selective pressures on genes that were differentially distributed
440 across coastal and offshore *Pelagibacteraceae* genera (**Fig. 5; Extended Data Fig. 6)**. The
441 differentially distributed genes with the highest selective pressures in offshore genera included a
442 malonyl-CoA decarboxylase (*mlycd;* pN/pS= 0.019±0.010, mean ± sd), an aldehyde
443 dehydrogenase (*aldB*; pN/pS= 0.033±0.018), a formate dehydrogenase
444 (*fdh;*pN/pS=0.0127±0.001), a D-lactate dehydrogenase (*ldhD*; pN/pS=0.018±0.003) and an
445 alanine-glyoxylate / serine-glyoxylate / serine-pyruvate transaminase (*agxt*; pN/pS=
446 0.022±0.008) (**Fig. 5**). These genes are generally involved in the metabolism of small and
447 common carbon substrates and the capacity to produce glycine, which is an auxotrophy in other
448 *Pelagibacteraceae*. Within *Xanthipelagibacter*, genes involved in purine uptake (*uraA*,
449 pN/pS=0.040±0.001; *nupNOPQ*, pN/pS=0.045±0.010) experienced higher purifying selection
450 than purine degradation pathways (*xdhAB*, pN/pS=0.084±0.028), which is in line with previous
451 observations that high-affinity transport is key to *Pelagibacteraceae*'s success as an
452 oligotroph[50,51]. The low pN/pS values for these distinct carbon and nitrogen related genes support
453 the hypothesis that these metabolisms provide a fitness advantage in offshore waters.



Ecologically-relevant genes for **offshore** and **coastal** genera

agxt	alanine-glyoxylate/ serine-glyoxylate/ serine-pyruvate transaminase	allC	allantoate deiminase
aldB	aldehyde dehydrogenase	frcABC	fructose transport system
fdh	formate dehydrogenase	rcl	nucleoside 2-deoxyribosyltransferase
ldhD	d-lactate dehydrogenase	tarD	d-tartrate dehydrogenase
mlycd	malonyl-CoA decarboxylase	vgsc	voltage-gated sodium channel
nupNOPQ	guanosine transporter		
uraA	xanthine permease		

455 **Fig. 5**. **Habitat-specific genes are under relatively higher selective pressures compared to**
456 **non-diagnostic genes.** pN/pS values per gene ranked from lowest to highest across each
457 genome. pN/pS values of 0.01, 0.025, 0.05, 0.075, 0.1, and 0.25 are shown in blue. Genes that
458 are differentially distributed between offshore and coastal *Pelagibacteraceae* genera in **Fig. 4** are
459 colored red (offshore) or black (coastal) across the genomes. For genes that are discussed in the
460 text, the gene names are noted and the description of the gene provided in the key. The full suite
461 of gene names and descriptions can be found in **Extended Data Fig. 6**.
462
463 Among the genes unique to coastal genera, the voltage-gated sodium channel was under
464 high selective pressure (*vgsc*, pN/pS=0.016±0.004). Coastal Kāneʻohe Bay harbors a larger range
465 of salinity, temperature, and pH conditions compared to the adjacent offshore[26,27,39], and thus the
466 voltage-gated sodium channels may be critical for coastal genera as they rapidly regulate ionic
467 composition[37]. The allantoin deiminase (*allC*, pN/pS=0.039±0.002) and nucleoside
468 2-deoxyribosyltransferase (*rcl*, pN/pS=0.054±0.006) were also under moderately high selective
469 pressures in coastal genomes, further supporting the specialization of coastal genera on
470 allantonate and nucleotides and a separation in purine metabolism niche from offshore genera.
471 Three sugar transporter genes (*frcABC*, pN/pS=0.024±0.006) within *Ampluspelagibacter* and a
472 tartrate dehydratase (*tarD*, pN/pS=0.051±0.001) within *Littoralipelagibacter* exhibited low
473 pN/pS values suggesting that sugar metabolisms and transporters are not only a more common
474 genomic feature in the coastal genera, but also likely essential for their fitness.
475 Finally, to understand which *Pelagibacteraceae* core genes were under the highest
476 selective pressures in the KByT system, we examined the mean pN/pS values for core genes
477 found across the same seven genomes. Not surprisingly, ribosomal proteins were found among
478 the genes with the highest purifying selection (low pN/pS values; **Table 2**). Genes related to
479 baseline functions of the cell including general energy production and conversation, translational,
480 ribosomal structure and biogenesis, transcription, and nucleotide transport and metabolism were
481 also under high selective pressure (**Table 2**). Two genes involved in regulatory systems for

482 responses to acidity and glycine were among the top 25 core genes with the most purifying
483 selection. The low pN/pS values for these genes supports that although *Pelagibacteraceae* may
484 have few regulatory genes generally[29], the ones it does have likely impact the the performance
485 and survival of *Pelagibacteraceae* cells.

486 Only two genes related to nutrient acquisition were found among the top 25 core genes
487 with the highest purifying selection. This included both subunits of the adenylylsulfate reductase
488 (*aprAB*), which is involved in dissimilatory sulfur metabolism, taking sulfite to adenylyl sulfate
489 (APS). Incomplete dissimilatory and assimilatory sulfate reduction pathways were core among
490 all *Pelagibacteraceae* examined in this study (**Extended Data Fig. 7**), in line with past analysis
491 of *Pelagibacteraceae* genomes[52]. As reported by Smith et al.[53], adenylylsulfate reductase likely
492 removes sulfite that accumulates during organic sulfur compound degradation. Further
493 supporting this hypothesis, we found multiple core or near core genes involved in the production
494 of sulfite from reduced organosulfur compounds: (2R)-sulfolactate sulfo-lyase (*suyAB*), a
495 sulfoacetaldehyde acetyltransferase (*xsc*), and a sulfur dioxygenase (**Extended Data Fig. 7**). Our
496 analyses reveal the genetic potential to metabolize a wide diversity of reduced organosulfur
497 compounds as alternative sources of sulfur and that adenylylsulfate reductases (*aprAB*) likely
498 plays an important role in accommodating *Pelagibacteraceae*'s unique sulfur metabolisms.
499
500
501
502
503
504

Table 2. **Core *Pelagibacteraceae* genes under high selective pressures in the KByT system**. The lowest 25 mean pN/pS values for core *Pelagibacteraceae* genes with shared KOfam annotations examined from genomes representing the seven *Pelagibacteraceae* genera that are highly abundant in the tropical Pacific. pN/pS: proportion of non-synonymous to synonymous. Enzymes involved in regulatory processes are denoted by a pound sign (#).

Accession	Description	pN/pS (mean±s.d.)
K02946	small subunit ribosomal protein S10	0.012±0.004
K00340	NADH-quinone oxidoreductase subunit K	0.013±0.008
K03560	biopolymer transport protein TolR	0.014±0.007
K03628	transcription termination factor Rho	0.014±0.004
K02078	acyl carrier protein	0.014±0.007
K03544	ATP-dependent Clp protease, subunit ClpX	0.017±0.004
K00053	ketol-acid reductoisomerase	0.017±0.005
K00395	adenylylsulfate reductase, subunit B	0.018±0.008
K00394	adenylylsulfate reductase, subunit A	0.018±0.005
K00412	ubiquinol-cytochrome c reductase cytochrome b subunit	0.018±0.006
K02274	cytochrome c oxidase subunit I	0.018±0.005
K02982	small subunit ribosomal protein S3	0.019±0.007
K03566#	LysR family transcriptional regulator, glycine cleavage system transcriptional activator	0.019±0.011
K01494	dCTP deaminase	0.019±0.006
K02878	large subunit ribosomal protein L16	0.019±0.008
K03798	cell division protease FtsH	0.019±0.005
K03702	excinuclease ABC subunit B	0.020±0.008
K00330	NADH-quinone oxidoreductase subunit A	0.020±0.010
K03569	rod shape-determining protein MreB	0.020±0.011
K02874	large subunit ribosomal protein L14	0.020±0.011
K14981#	two-component system, OmpR family, response regulator ChvI	0.020±0.01
K02988	small subunit ribosomal protein S5	0.021±0.005
K03046	DNA-directed RNA polymerase subunit beta	0.021±0.003
K03701	excinuclease ABC subunit A	0.021±0.003
K00331	NADH-quinone oxidoreductase subunit B	0.021±0.006

Discussion

Through genomic and metagenomic characterizations of a recently expanded collection of SAR11 isolates[22], our study reveals a polyphyletic distribution of coastal ocean versus offshore habitat specialists among closely related SAR11 populations, where the emergence of key genetic features that underlie this strong habitat preferences seem to have occurred through independent evolutionary events rather than transfer from a single common ancestor. Genera that share the same environment (e.g. coastal or offshore) have also accumulated shared and unique gene content that is predominantly involved in the metabolism of organic carbon and the acquisition of nitrogen from organic sources. A subset of these metabolic genes were under high purifying selection, emphasizing the importance of these functions to the fitness of distinct *Pelagibacteraceae* genera and highlighting potential determinants of niche differentiation of *Pelagibacteraceae* in coastal and offshore environments.

The cohesion of both ecological and genetic diversity observed within genera of *Pelagibacteraceae* most closely resembles ecotypes: ecologically homogeneous groups of closely related bacteria whose genetic diversity is guided by cohesive forces of selection, recombination, and genetic drift[54,55]. Ecotypic differentiation has previously been suggested to describe the spatial and temporal variation partitioning genetic and genomic SAR11 diversity[9,12,22,26,56–61], however the underlying evolutionary processes or functional differences contributing to this diversification have rarely been linked to patterns of distribution. Our findings reveal evolutionary drivers behind ecotypic differentiation by explaining distinct ecological distribution patterns with differences in gene content, metabolic potential, and selective pressures across ecotypes. The use of a constrained model system that provided both

535 metagenomic sequence data as well as isolate genomes from local *Pelagibacteraceae*
536 populations was paramount to being able to resolve signatures of environmental selection in
537 sympatric and parapatric SAR11 populations. The power of this approach is further demonstrated
538 in the fact that most of the *Pelagibacteraceae* ecotypes defined in this constrained system as
539 having preferences for coastal oceans or the open ocean appear to hold when global read
540 recruitment data were examined[22].

541 Our results suggest that *Pelagibacteraceae* has transitioned between offshore and coastal
542 environments multiple times, with a handful of genes that support these lifestyles likely acquired
543 through multiple independent processes. The small number of adaptive genes are shared among
544 ecologically similar, but polyphyletically-distributed genera, where genomes have maintained a
545 substantial amount of genetic diversity. Given the distribution patterns of the adaptive genes and
546 the genetic diversity maintained within the populations that carry them, the proliferation of
547 adaptive genes unlikely occurred via genome-wide selective sweep[62], a process that describes the
548 clonal expansion of an individual subpopulation that carries an adaptive gene leading to a
549 genome-wide reduction in genetic diversity. Instead, the maintenance of genetic diversity and the
550 relatively strong purifying selection to maintain the functional identity of these adaptive genes is
551 consistent with a gene-specific selective sweep[25,63–66]. Consistent with previous studies that have
552 identified a handful of genes responsible for habitat-specific adaptation[25,66,67], the gene content
553 differences attributed to stable ecological speciation here are generally found within the
554 relatively stable genomic backbone, rather than within hyper-variable genomic islands. This is a
555 critical observation as it links eco-evolutionary processes to specific portions of the immense
556 genomic diversity that exists within *Pelagibacteraceae*, and underscores the importance of
557 analyses informed by genomic architecture. Future studies that examine the contribution of gene

558 content found within hyper-variable islands to stable ecological speciation as opposed to
559 incipient speciation or intraspecific diversity would provide a better understanding of the role of
560 hyper-variable regions in long-term evolutionary processes.

561 In closing, our analyses broaden the metabolic diversity known from *Pelagibacteraceae*
562 genomes and suggest that metabolic versatility has contributed to SAR11's success in the global
563 ocean. While metabolic reconstructions are subject to uncertainties that necessitate physiological
564 exploration with controlled experimentation in the laboratory, the metabolic analysis of
565 high-quality genomes from cultivated isolates provides the opportunity to delineate metabolic
566 specialization that could advance cultivation approaches and allow for more targeted media
567 recipes[68–71]. SAR11 is estimated to oxidize between 6% to 37% of gross ocean primary
568 productivity[20]. Future work combining these metabolic predictions with *in situ* measures will
569 greatly improve our understanding of how SAR11 impacts the dissolved organic matter pool and
570 biogeochemical cycles in the ocean.

571 Methods

572 Pangenome analyses

573 Publicly-available *Pelagibacteraceae* isolate genomes downloaded from the National
574 Center for Biotechnology Information (NCBI) or the Joint Genome Institute (JGI) were
575 combined with genomes sequenced from newly isolated strains[22] (**Supplementary file 1a**).
576 Genome completion and contamination were examined with checkM v1.1.2[72] and only isolates
577 above 95% completion with less than 5% redundancy[73] and fewer than 50 contigs were kept
578 (**Supplementary file 1a).** High-quality *Pelagibacteraceae* genomes were used to construct a

579 pangenome using anvi'o v8.0[74] following previously described pipelines[67]. Briefly, anvi'o
580 database was created using `anvi-gen-contigs-db` and Prodigal v2.6.3[75] was used to identify
581 open reading frames (ORFs) from contigs. Single-copy core genes were identified using
582 HMMER v3.2.1[76]. ORFs with associated functions were annotated from NCBI's Clusters of
583 Orthologous Groups (COGs)[77] and KOfams[23,24] (a customized HMM database of KEGG
584 orthologs). The pangenome was created with `anvi-pan-genome`, which uses NCBI's Basic
585 Local Alignment Search Tool (BLAST)[78] to quantify the similarity between pairs of gene clusters
586 and the Markov Cluster algorithm (MCL)[79] to define homologous gene clusters with a MCL
587 inflation parameter of 2. The pangenomes were visualized using the command
588 `anvi-display-pan` and summary tables exported using the command `anvi-summarize`.

589 Genome-wide average nucleotide identity (gANI) was estimated using FastANI in
590 anvi'o[80]. A phylogenomic tree of the *Pelagibacteraceae* was estimated using IQ-Tree v2.12[81]
591 using 1000 ultrafast bootstraps and model LG+F+R10 from a concatenated alignment of a
592 custom gene set for SAR11 (SAR11_165[22]). Phylogenies were rooted and branches trimmed in R
593 v 4.4.1[82] using treeio v1.28.0[83] and visulized in R using phytools v 2.3.0[84] to examine the
594 phylogenomic relationships between *Pelagibacteraceae* groups detected in our environmental
595 study.

596 Gene clusters found in 100% of *Pelagibacteraceae* genomes were considered core while
597 gene clusters found in only one genome were considered singletons. Gene clusters were also
598 assessed as genus specific (shared among all representatives of a genus and not found in other
599 genera) and multi-genus specific (shared among all representatives of two or more genera and
600 not found among all others). Shared gene content between genera was assessed using
601 ComplexUpset v 1.3.3[85].

Metagenomic read recruitment and environmental analyses

To examine the distribution of *Pelagibacteraceae* genera within the environments from where the majority were isolated, metagenomic read recruitment was conducted using surface ocean metagenomes from 10 stations within and adjacent to Kānʻeohe Bay, Hawaiʻi (**Fig. 1**) collected as part of the Kāneʻohe Bay Time-series (KByT; PRJNA971314)[26,27]. In addition, metagenomes collected in the surrounding North Pacific Subtropical Gyre at Station ALOHA, a sampling location within the Hawaii Ocean Time-series (HOT)[86,87] (**Fig. 1; Supplementary file 1c**; PRJNA352737[88]) were also used. Metagenomes were competitively mapped with Bowtie2 v 2.3.5[89] to the anvi'o contig database of *Pelagibacteraceae* isolate genomes. The `anvi-profile` function stored coverage and detection statistics of each *Pelagibacteraceae* genome found in the KByT and Station ALOHA metagenomic samples and the `anvi-meta-pan-genome` function[67] was used to bring together the pangenomic information with the read recruitment data.

To evaluate the distribution of individual genomes, a detection metric- the proportion of the nucleotides in a given sequence that are covered by at least one short read- was used to define whether a population was detected in a metagenomic sample. A detection value of at least 0.25 was used as criteria to eliminate false positives that could arise if an isolate genome was falsely found within a sample[90].

The average depth of coverage excluding nucleotide positions with coverages in the 1st and 4th quartiles (mean coverage Q2Q3) was mapped for each genome in each sample. To avoid biased estimates of coverage that can occur due to highly recruiting accessory genes, read recruitment was analyzed using only single-copy genes core to each genus. Using a custom script in R, gene clusters that were found within all genomes of the genus (e.g. core), but found only in a single copy within each genome were identified. Then the mean coverage Q2Q3 read

recruitment data per gene were then subsetted for only single copy core genes. The mean coverage Q2Q3 of each single-copy core gene was summed per genome per sample and then this value was normalized per genome per sample, based on whether the genome was detected at >0.25[90]. Next, when evaluating read recruitment at the genus-level, the read recruitment was summed for all genomes in a genus per sample. These outputs were then divided across the total *Pelagibacteraceae* read recruitment for the sample to yield a relative estimate of each genome or genus within a sample. For the genus with only a single genome representative (*Semelpelagibacter*), the single-copy core genes shared with the HIMB1483 genome and nearest neighbor, genus *Lacunapelagibacter*, were used.

K-means clustering analysis of a scaled matrix of biogeochemical conditions was used to characterize the environmental background from which the metagenomic data were derived. Metadata from Station ALOHA were downloaded from https://hahana.soest.hawaii.edu/hot/hot-dogs/ (accessed 11 Oct 2021). All KByT (n=158) and Station ALOHA (n=34) surface seawater samples (depth of <30m) with data for flow cytometrically-determined cellular abundance, chlorophyll *a* concentrations, silicate concentrations, temperature, pH, and salinity were included in the analyses (**Supplementary file 1c**). The number of clusters were estimated using the `kmeansruns` command in the R fpc v 2.2.13 package[91] using a Calinski Harabasz index. Principal Component Analysis (PCA) was conducted using the `prcomp` function in R v 4.4.1[82] and visualized using 'ggbiplot' in the ggplot2 v 3.5.1[92].

The sampling locations of the Kāneʻohe Bay Time-series and Station ALOHA were mapped using R with `geom_sf` from ggplot2 with geospatial data of the main Hawaiian Islands (USGS Digital Line Graphs). To visualize *Prochlorococcus* cellular abundance across the

648 Kāneʻohe Bay Time-series, `mba.surf` from MBA v 0.1.2[93] was used to interpolate data over the
649 KByT stations.

650 Functional inferences from *Pelagibacteraceae* genera

651 Functional enrichment analysis was performed in anvi'o and has been described
652 previously[94]. Briefly, the `anvi-compute-functional-enrichment` command, as well as the
653 `anvi-script-gen-function-matrix-across-genomes`, was used to assess the enrichment of
654 Clusters of Orthologous Groups (COG), KEGG orthologs (KOs), and KEGG Modules across
655 genomes and genus affiliations. The degree of completeness of individual KEGG modules[95,96] in
656 the genomes and genera was evaluated using `anvi-estimate-metabolism`[97], prior to
657 `anvi-compute-functional-enrichment`. The functional pangenome was visualized using
658 `anvi-display-functions`. Central and alternative carbon metabolisms, amino acid, vitamin, and
659 cofactor biosynthesis, and genes to cope with nutrient, osmotic, and oxidative stress were
660 examined across the *Pelagibacteraceae* isolate genomes. The distribution of functional genes
661 were evaluated for each genus as follows: core genes found in all genomes of a genus,
662 intermediate genes found in 15-99% of genomes, and rare genes found in less than 15%.
663 Heatmaps of the metabolic gene distribution across genera were made in the R package
664 pheatmap v 1.0.12[98].

665 Selection within *Pelagibacteraceae* genes

666 To evaluate the selective pressures on individual metabolic genes of interest, the
667 proportion of non-synonymous to synonymous (pN/pS) sites per gene across genomes was
668 evaluated. The pipeline followed those developed in Kiefl et al.,[18]. Briefly, contig databases from

669 the genomes of type specimens from each of the *Pelagibacteraceae* genera[22] were first
670 functionally annotated in anvi'o (see above) and metagenomic reads from deeply sequenced
671 metagenomes from either coastal Kāne'ohe Bay or the adjacent offshore were non-competitively
672 recruited using Bowtie2 v 2.3.5[89] (**Supplementary file 1c**).

673 Next, the`anvi-profile` function with the flag `--profile-SCVs` was used to characterize
674 single codon variants across the read recruitment data and then `anvi-gen-variability-profile`
675 was used to export the single codon variants per gene per sample using the flags --engine CDN
676 --include-site-pN/pS --kiefl-mode[18]. Finally, pN/pS values were reported per gene with at least
677 20x coverage in each sample using 'anvi-get-pn-ps-ratio` with the -m 20 flag. Per gene coverage
678 statistics were exported using `anvi-summarize` with the `--init-gene-coverages` flag. Functional
679 annotations were exported using `anvi-export-functions`.

680 To evaluate whether a single genome representative contained similar pN/pS gene values
681 as other genomes within a given genus, these steps were repeated with all representatives of the
682 genus sharing the lowest minimum gANI (among the genera that were abundant in the tropical
683 Pacific): genus *Coralipelagibacter*. The exported data files were brought into R, where variation
684 in pN/pS per sample, per genome, and across genomes of the same genus was evaluated.
685 Relationships between pN/pS and coverage were examined and genes per genome representative
686 were ranked from lowest pN/pS value to highest and visualized in R using ggplot2 v 3.5.1[92].
687 Figures were made in R v 4.4.1[82] or anvi'o[74] and further edited with Inkscape
688 (https://inkscape.org/).

689

690 # Data availability

Accession numbers for genome sequences used to conduct pangenomic analyses and read recruitment in this study are found in **Supplementary file 1a**. Short-read metagenome accession IDs and environmental data used in this study are provided in **Supplementary file 1c**. Long-read metagenomes from coastal Kāneʻohe Bay are available under NCBI Project PRJNA1201851.

Code availability

Code to run analyses and create figures can be found at https://github.com/tucker4/SAR11_metabolism.

Competing interests

The authors declare no competing interests.

Acknowledgements

We thank Oscar Ramfelt and Evan B. Freel for their insights on computational analyses. This research was funded by the Simons Postdoctoral Fellowships in Marine Microbial Ecology (grant #989028), NOAA Margaret A. Davidson Fellowship, the Colonel Willys E. Lord, DVM and Sadina L. Lord Scholarship, the ARCS Foundation George Orton and Mona Marie Elmore Award in Marine Biology, the Research Corporation of the University of Hawai‘i Graduate Research Fellowship, the EECB Hampton and Meredith Carson Fellowship, and the Center for Microbial Analysis through Island Knowledge & Investigation Research Grant to SJ Tucker, and NSF grants OCE-1538628 and OCE- 2149128 to MS Rappé. AME acknowledges support from the Center for Chemical Currencies of a Microbial Planet (C-CoMP) (NSF Award OCE-2019589, C-CoMP publication #065), and Simons Foundation (grant #687269)

References

1. Louca, S. *et al.* Integrating biogeochemistry with multiomic sequence information in a model oxygen minimum zone. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E5925–E5933 (2016).
2. Hao, L. *et al.* Novel syntrophic bacteria in full-scale anaerobic digesters revealed by genome-centric metatranscriptomics. *ISME J.* **14**, 906–918 (2020).
3. Wang, C. *et al.* Bacterial genome size and gene functional diversity negatively correlate with taxonomic diversity along a pH gradient. *Nat. Commun.* **14**, 7437 (2023).
4. Cordero, O. X. & Polz, M. F. Explaining microbial genomic diversity in light of evolutionary ecology. *Nat. Rev. Microbiol.* **12**, 263–273 (2014).
5. Sheppard, S. K., Guttman, D. S. & Fitzgerald, J. R. Population genomics of bacterial host adaptation. *Nat. Rev. Genet.* **19**, 549–565 (2018).
6. Whelan, F. J., Hall, R. J. & McInerney, J. O. Evidence for selection in the abundant accessory gene content of a prokaryote pangenome. *Mol. Biol. Evol.* **38**, 3697–3708 (2021).
7. Conrad, R. E. *et al.* Toward quantifying the adaptive role of bacterial pangenomes during environmental perturbations. *ISME J.* **16**, 1222–1234 (2022).
8. Morris, R. M. *et al.* SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**, 806–810 (2002).
9. Brown, M. V. *et al.* Global biogeography of SAR11 marine bacteria. *Mol. Syst. Biol.* **8**, 595 (2012).
10. Grote, J. *et al.* Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *MBio* **3**, (2012).
11. Delmont, T. O. *et al.* Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *Elife* **8**, (2019).
12. Haro-Moreno, J. M. *et al.* Ecogenomics of the SAR11 clade. *Environ. Microbiol.* **22**, 1748–1763 (2020).

13. Giovannoni, S. J. *et al.* Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**, 1242–1245 (2005).

14. Noell, S. E. *et al.* Differences in the regulatory strategies of marine oligotrophs and copiotrophs reflect differences in motility. *Environ. Microbiol.* **25**, 1265–1280 (2023).

15. Morris, J. J., Lenski, R. E. & Zinser, E. R. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *MBio* **3**, (2012).

16. Braakman, R., Follows, M. J. & Chisholm, S. W. Metabolic evolution and the self-organization of ecosystems. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E3091–E3100 (2017).

17. Vergin, K. L. *et al.* High intraspecific recombination rate in a native population of Candidatus pelagibacter ubique (SAR11). *Environ. Microbiol.* **9**, 2430–2440 (2007).

18. Kiefl, E. *et al.* Structure-informed microbial population genetics elucidate selective pressures that shape protein evolution. *Sci Adv* **9**, eabq4632 (2023).

19. Grant, S. R., Church, M. J., Ferrón, S., Laws, E. A. & Rappé, M. S. Elemental composition, phosphorous uptake, and characteristics of growth of a SAR11 strain in batch and continuous culture. *mSystems* **4**, (2019).

20. White, A. E., Giovannoni, S. J., Zhao, Y., Vergin, K. & Carlson, C. A. Elemental content and stoichiometry of SAR11 chemoheterotrophic marine bacteria. *Limnol. Oceanogr. Lett.* **4**, 44–51 (2019).

21. Acker, M. *et al.* Phosphonate production by marine microbes: Exploring new sources and potential function. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2113386119 (2022).

22. Freel, K. C., Tucker, S. J., Freel, E. B., Eren, A. M. & Rappé, M. S. New isolate genomes and global marine metagenomes resolve ecologically relevant units of SAR11. *In prep*.

23. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

24. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).

25. Shapiro, B. J. *et al.* Population genomics of early events in the ecological differentiation of bacteria. *Science* **336**, 48–51 (2012).

26. Tucker, S. J. *et al.* Spatial and temporal dynamics of SAR11 marine bacteria across a nearshore to offshore transect in the tropical Pacific Ocean. *PeerJ* **9**, e12274 (2021).

27. Tucker, S. J. *et al.* Sharp transitions in phytoplankton communities across estuarine to open ocean waters of the tropical Pacific. *bioRxiv* (2024) doi:10.1101/2024.05.23.595464.

28. Zhang, Y. & Gladyshev, V. N. Molybdoproteomes and evolution of molybdenum utilization. *J. Mol. Biol.* **379**, 881–899 (2008).

29. Giovannoni, S. J., Cameron Thrash, J. & Temperton, B. Implications of streamlining theory for microbial ecology. *ISME J.* **8**, 1553–1565 (2014).

30. Berman, T. & Bronk, D. A. Dissolved organic nitrogen: a dynamic participant in aquatic ecosystems. *Aquat. Microb. Ecol.* **31**, 279–305 (2003).

31. Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. *Science* **184**, 868–871 (1974).

32. Braakman, R. *et al.* Global niche partitioning of purine and pyrimidine cross-feeding among ocean microbes. *bioRxiv* (2024) doi:10.1101/2024.02.09.579562.

33. Tripp, H. J. *et al.* Unique glycine-activated riboswitch linked to glycine-serine auxotrophy in SAR11. *Environ. Microbiol.* **11**, 230–238 (2009).

34. Klähn, S., Steglich, C., Hess, W. R. & Hagemann, M. Glucosylglycerate: a secondary compatible solute common to marine cyanobacteria from nitrogen-poor environments. *Environ. Microbiol.* **12**, 83–94 (2010).

35. Castell, C., Rodríguez-Lumbreras, L. A., Hervás, M., Fernández-Recio, J. & Navarro, J. A. New insights into the evolution of the electron transfer from cytochrome f to photosystem I in the Green and red branches of photosynthetic eukaryotes. *Plant Cell Physiol.* **62**, 1082–1093 (2021).

36. Peers, G. & Price, N. M. Copper-containing plastocyanin used for electron transport by an oceanic diatom. *Nature* **441**, 341–344 (2006).

37. Ren, D. *et al.* A prokaryotic voltage-gated sodium channel. *Science* **294**, 2372–2375 (2001).

38. Strege, P. R. *et al.* Mechanosensitive pore opening of a prokaryotic voltage-gated sodium channel. *Elife* **12**, (2023).

39. Yeo, S. K., Huggett, M. J., Eiler, A. & Rappé, M. S. Coastal bacterioplankton community dynamics in response to a natural disturbance. *PLoS One* **8**, e56207 (2013).

40. Sun, J. *et al.* One carbon metabolism in SAR11 pelagic marine bacteria. *PLoS One* **6**, e23973 (2011).

41. Giovannoni, S. J. SAR11 bacteria: The most abundant plankton in the oceans. *Ann. Rev. Mar. Sci.* **9**, 231–255 (2017).

42. Moore, E. R., Weaver, A. J., Davis, E. W., Giovannoni, S. J. & Halsey, K. H. Metabolism of key atmospheric volatile organic compounds by the marine heterotrophic bacterium Pelagibacter HTCC1062 (SAR11). *Environ. Microbiol.* **24**, 212–222 (2022).

43. Gralka, M., Pollak, S. & Cordero, O. X. Genome content predicts the carbon catabolic preferences of heterotrophic bacteria. *Nat. Microbiol.* **8**, 1799–1808 (2023).

44. Leyton, D. L., Rossiter, A. E. & Henderson, I. R. From self sufficiency to dependence: mechanisms and factors important for autotransporter biogenesis. *Nat. Rev. Microbiol.* **10**, 213–225 (2012).

45. Doyle, M. T., Tran, E. N. H. & Morona, R. The passenger-associated transport repeat promotes virulence factor secretion efficiency and delineates a distinct autotransporter subtype. *Mol. Microbiol.* **97**, 315–329 (2015).

46. Zhao, X. *et al.* Three-Dimensional Structure of the Ultraoligotrophic Marine Bacterium 'Candidatus Pelagibacter ubique'. *Appl. Environ. Microbiol.* **83**, (2017).

47. Hazes, B. & Frost, L. Towards a systems biology approach to study type II/IV secretion systems. *Biochim. Biophys. Acta* **1778**, 1839–1850 (2008).

48. Ellison, C. K. *et al.* Retraction of DNA-bound type IV competence pili initiates DNA uptake during natural transformation in Vibrio cholerae. *Nat. Microbiol.* **3**, 773–780 (2018).

49. Meuskens, I., Saragliadis, A., Leo, J. C. & Linke, D. Type V secretion systems: An overview of passenger domain functions. *Front. Microbiol.* **10**, 1163 (2019).

821 50. Noell, S. E. & Giovannoni, S. J. SAR11 bacteria have a high affinity and multifunctional glycine
822 betaine transporter. *Environ. Microbiol.* **21**, 2559–2575 (2019).
823 51. Clifton, B. E., Alcolombri, U., Uechi, G.-I., Jackson, C. J. & Laurino, P. The ultra-high affinity
824 transport proteins of ubiquitous marine bacteria. *Nature* **634**, 721–728 (2024).
825 52. Tripp, H. J. *et al.* SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature* **452**,
826 741–744 (2008).
827 53. Smith, D. P. *et al.* Proteome Remodeling in Response to Sulfur Limitation in 'Candidatus
828 Pelagibacter ubique'. *mSystems* **1**, (2016).
829 54. Cohan, F. M. Towards a conceptual and operational union of bacterial systematics, ecology, and
830 evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 1985–1996 (2006).
831 55. Koeppel, A. *et al.* Identifying the fundamental units of bacterial diversity: a paradigm shift to
832 incorporate ecology into bacterial systematics. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 2504–2509
833 (2008).
834 56. Field, K. G. *et al.* Diversity and depth-specific distribution of SAR11 cluster rRNA genes from
835 marine planktonic bacteria. *Appl. Environ. Microbiol.* **63**, 63–70 (1997).
836 57. Schwalbach, M. S., Tripp, H. J., Steindler, L., Smith, D. P. & Giovannoni, S. J. The presence of the
837 glycolysis operon in SAR11 genomes is positively correlated with ocean productivity. *Environ.*
838 *Microbiol.* **12**, 490–500 (2010).
839 58. Vergin, K. L. *et al.* High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series
840 Study site by phylogenetic placement of pyrosequences. *ISME J.* **7**, 1322–1332 (2013).
841 59. Tsementzi, D. *et al.* Ecogenomic characterization of widespread, closely-related SAR11 clades of the
842 freshwater genus 'Candidatus Fonsibacter' and proposal of Ca. Fonsibacter lacus sp. nov. *Syst. Appl.*
843 *Microbiol.* **42**, 495–505 (2019).
844 60. Kraemer, S., Ramachandran, A., Colatriano, D., Lovejoy, C. & Walsh, D. A. Diversity and
845 biogeography of SAR11 bacteria from the Arctic Ocean. *ISME J.* **14**, 79–90 (2020).
846 61. Larkin, A. A., Hagstrom, G. I., Brock, M. L., Garcia, N. S. & Martiny, A. C. Basin-scale

847 biogeography of Prochlorococcus and SAR11 ecotype replication. *ISME J.* **17**, 185–194 (2023).

848 62. Shapiro, B. J., Leducq, J.-B. & Mallet, J. What is speciation? *PLoS Genet.* **12**, e1005860 (2016).

849 63. Falush, D. *et al.* Recombination and mutation during long-term gastric colonization by Helicobacter
850 pylori: estimates of clock rates, recombination size, and minimal age. *Proc. Natl. Acad. Sci. U. S. A.*
851 **98**, 15056–15061 (2001).

852 64. Whitaker, R. J., Grogan, D. W. & Taylor, J. W. Recombination shapes the natural population
853 structure of the hyperthermophilic archaeon Sulfolobus islandicus. *Mol. Biol. Evol.* **22**, 2354–2361
854 (2005).

855 65. Rosen, M. J., Davison, M., Bhaya, D. & Fisher, D. S. Microbial diversity. Fine-scale diversity and
856 extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science* **348**,
857 1019–1023 (2015).

858 66. Bendall, M. L. *et al.* Genome-wide selective sweeps and gene-specific sweeps in natural bacterial
859 populations. *ISME J.* **10**, 1589–1601 (2016).

860 67. Delmont, T. O. & Eren, A. M. Linking pangenomes and metagenomes: the Prochlorococcus
861 metapangenome. *PeerJ* **6**, e4320 (2018).

862 68. Rappé, M. S., Connon, S. A., Vergin, K. L. & Giovannoni, S. J. Cultivation of the ubiquitous SAR11
863 marine bacterioplankton clade. *Nature* **418**, 630–633 (2002).

864 69. Stingl, U., Tripp, H. J. & Giovannoni, S. J. Improvements of high-throughput culturing yielded novel
865 SAR11 strains and other abundant marine bacteria from the Oregon coast and the Bermuda Atlantic
866 Time Series study site. *ISME J.* **1**, 361–371 (2007).

867 70. Song, J., Oh, H.-M. & Cho, J.-C. Improved culturability of SAR11 strains in dilution-to-extinction
868 culturing from the East Sea, West Pacific Ocean. *FEMS Microbiol. Lett.* **295**, 141–147 (2009).

869 71. Henson, M. W. *et al.* Artificial seawater media facilitate cultivating members of the microbial
870 majority from the Gulf of Mexico. *mSphere* **1**, (2016).

871 72. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the
872 quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*

25, 1043–1055 (2015).

73. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).

74. Eren, A. M. *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nat. Microbiol.* **6**, 3–6 (2021).

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

76. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

77. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).

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

79. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).

80. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).

81. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

82. R Core Team. *R: A Language and Environment for Statistical Computing*. (2023).

83. Wang, L.-G. *et al.* Treeio: An R package for phylogenetic tree input and output with richly annotated and associated data. *Mol. Biol. Evol.* **37**, 599–603 (2020).

84. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).

85. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of

intersecting sets. *IEEE Trans. Vis. Comput. Graph.* **20**, 1983–1992 (2014).

86. Karl, D. M. & Lukas, R. The Hawaii Ocean Time-series (HOT) program: Background, rationale and field implementation. *Deep Sea Res. Part 2 Top. Stud. Oceanogr.* **43**, 129–156 (1996).

87. Karl, D. M. & Church, M. J. Microbial oceanography and the Hawaii Ocean Time-series programme. *Nat. Rev. Microbiol.* **12**, 699–713 (2014).

88. Mende, D. R. *et al.* Environmental drivers of a microbial genomic transition zone in the ocean's interior. *Nat. Microbiol.* **2**, 1367–1373 (2017).

89. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

90. Utter, D. R., Borisy, G. G., Eren, A. M., Cavanaugh, C. M. & Mark Welch, J. L. Metapangenomics of the oral microbiome provides insights into habitat adaptation and cultivar diversity. *Genome Biol.* **21**, 293 (2020).

91. Hennig, C. *Fpc: Flexible Procedures for Clustering. R Package Version 2.2-13*. (2024).

92. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York*. (2016).

93. Finley, A., Banerjee, S. & Hjelle, Ø. *MBA: Multilevel B-Spline Approximation. R Package Version 0.1-2*. (2017).

94. Shaiber, A. *et al.* Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol.* **21**, 292 (2020).

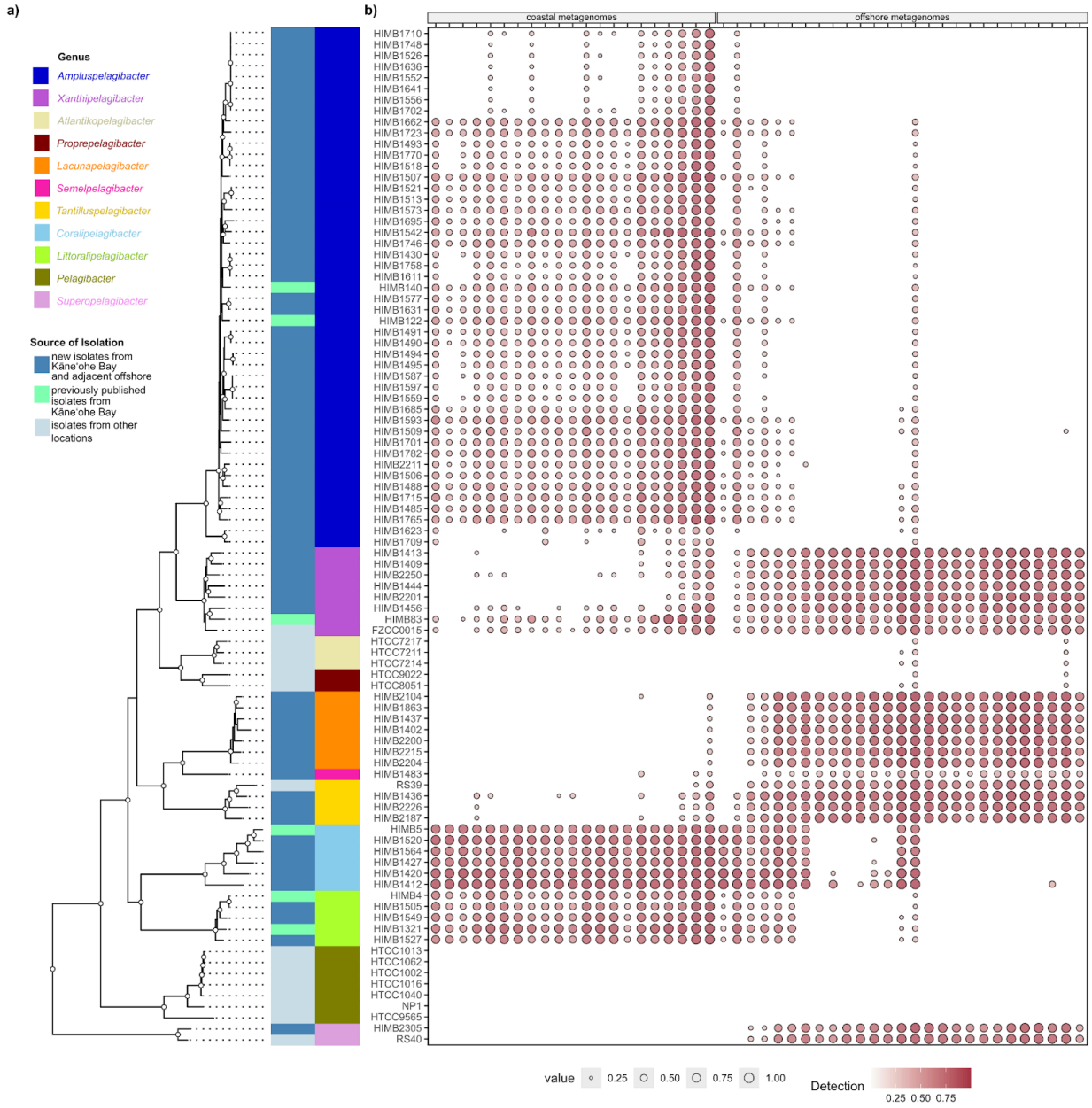
95. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199–205 (2014).

96. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).

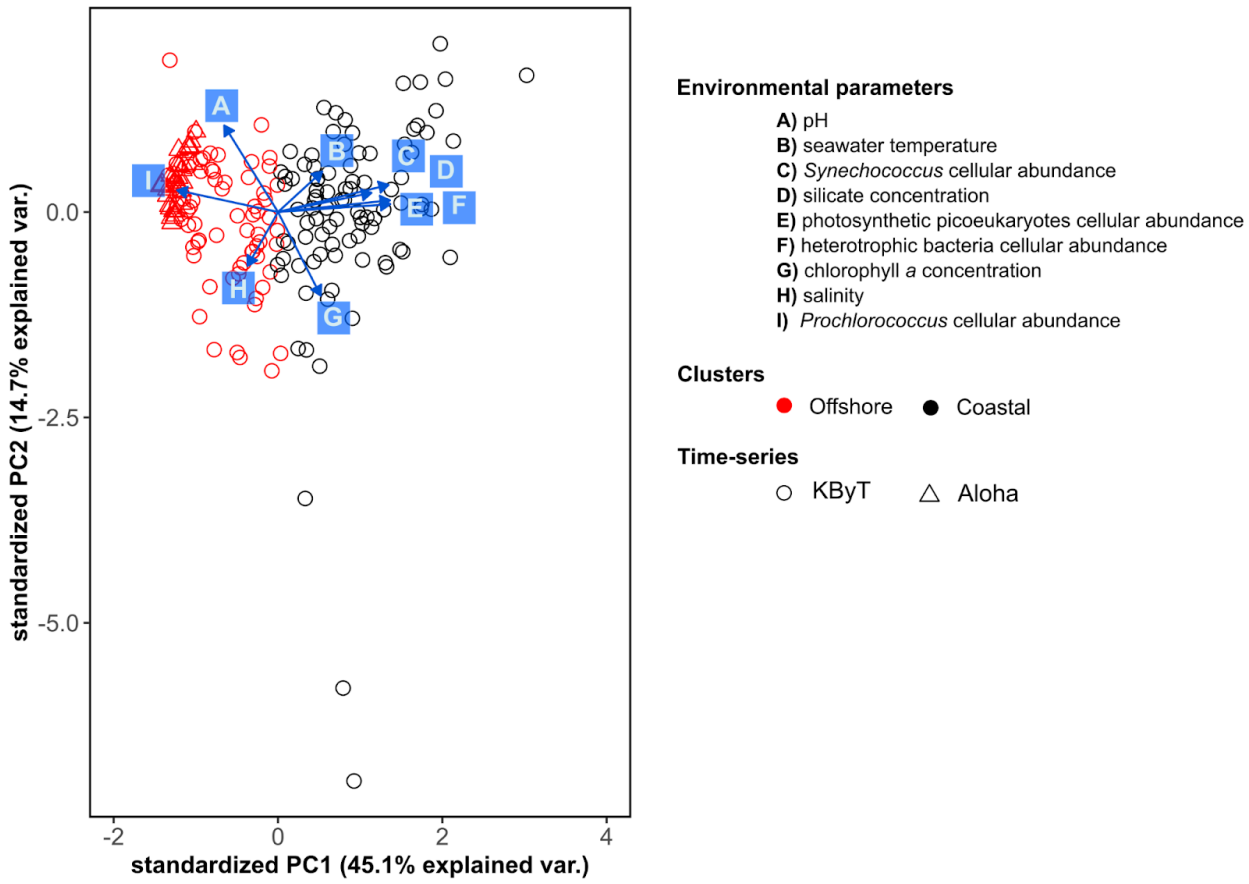
97. Veseli, I. *et al.* Microbes with higher metabolic independence are enriched in human gut microbiomes under stress. (2023) doi:10.7554/elife.89862.

98. Kolde, R. *Pheatmap: Pretty Heatmaps. R Package Version 1.0.12*. (2019).

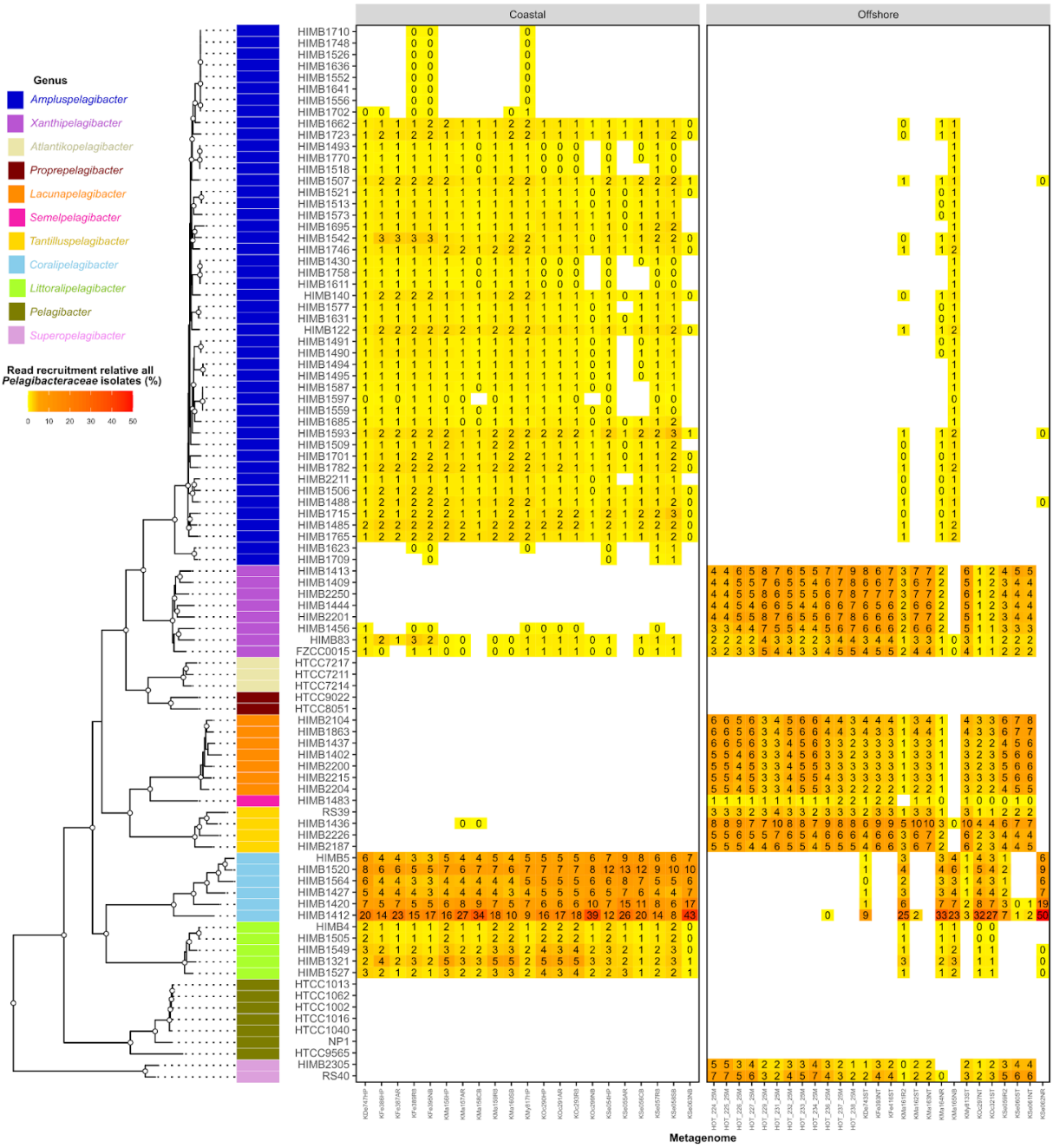
Extended Data Figures



Extended Data Fig. 1. a) Phylogenomic tree showing 11 monophyletic clusters within *Pelagibacteraceae*, their genus assignments, and source of isolation. Circles at nodes indicate ultrafast bootstrap support values ≥90% from 1000 replicates. **b)** Detection of all *Pelagibacteraceae* genomes in coastal and offshore metagenomic samples.



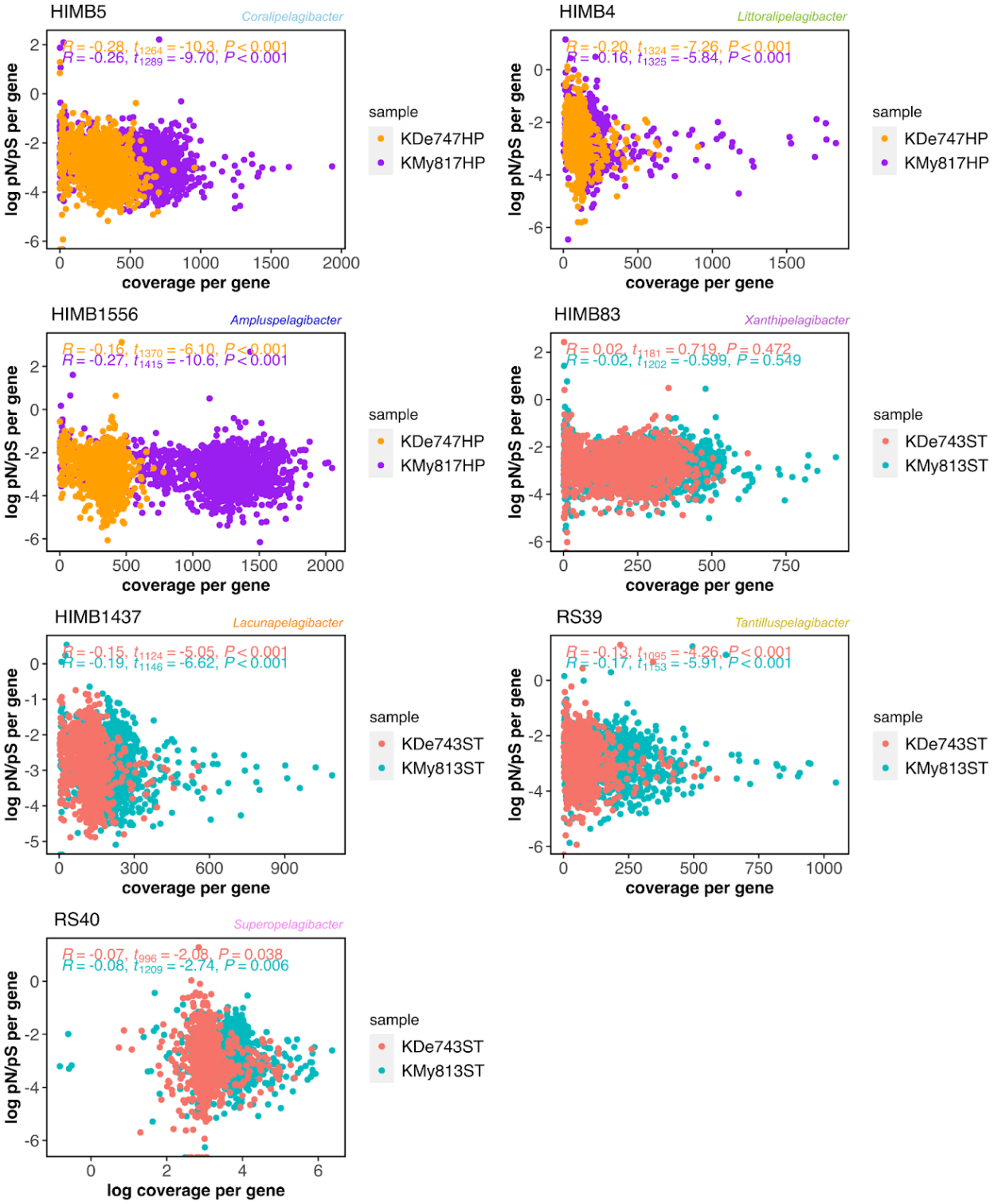
Extended Data Fig. 2. K-means analyses of environmental parameters from 192 surface ocean samples from KByT and Station ALOHA grouped into two clusters, herein referred to as coastal and offshore.



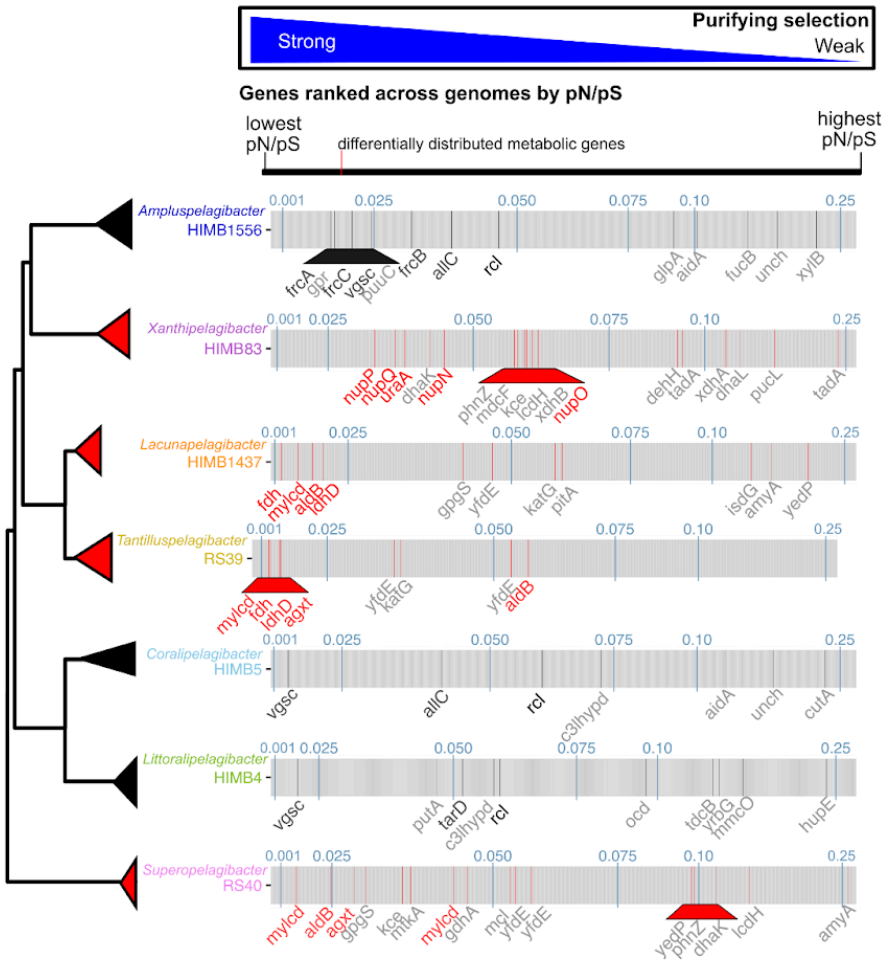
Extended Data Fig. 3. Relative read recruitment of each isolate genome across the coastal and offshore environment. Data for genomes that are not detected or have a relative recruitment less than 0.25% are not shown.

Extended Data Fig. 4. **a)** A protein structure model of a gene annotated as an autotransporter adhesin from KByT long-read metagenomes and sharing close sequence similarity to gene sequences in *Pelagibacteraceae* genomes shows characteristics expected of autotransporter proteins: 1) α-helical linker, 2) β-barrel translocator domain, and 3) passenger domains containing β helices (Leyton et al., 2012). **b)** Homolog searches to recover autotransporter adhesin genes in long-read metagenomes recovered nine non-redundant gene sequences that were near exact matches to most gene sequences in *Pelagibacteraceae* genomes (33 of 43). **c)**

Patterns of detection of short-read metagenomic read recruitment data across genes in two genomes of *Pelagibacteraceae* show that autotransporter adhesion genes are sometimes located in hypervariable regions and sometimes not. Genes for each *Pelagibacteraceae* genome are ordered by genome synteny with detection values of read recruitment data per gene shown as a bar per metagenomic sample. Areas of low detection are indicative of hypervariable regions (HVRs), and the autotransporter adhesin genes were located in HVRs within the HIMB5 genome, but not within HIMB4 genome. The autotransport adhesion genes were always positioned next to genes involved in type IV pilus systems and/or type II secretory systems.



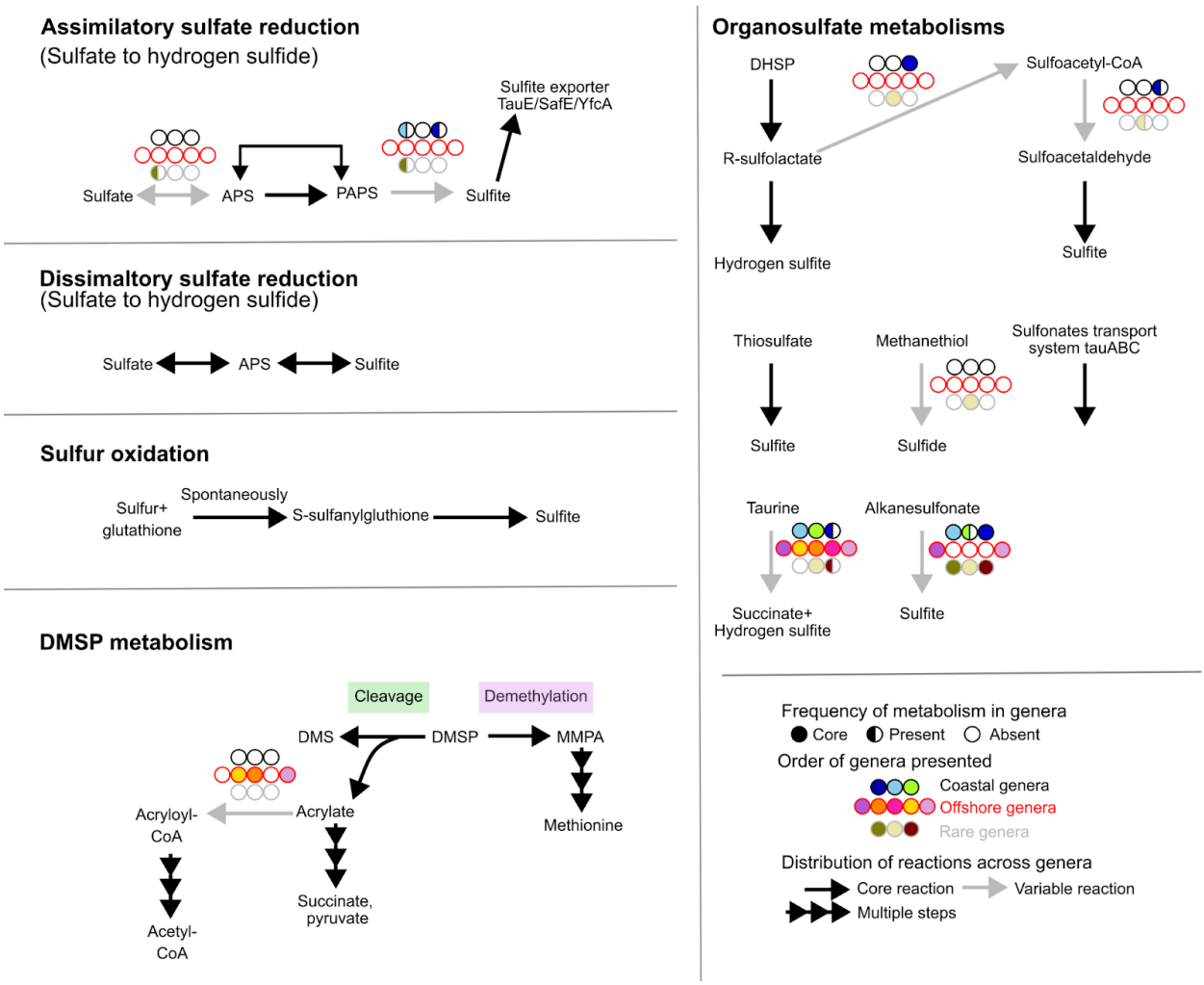
Extended Data Fig. 5. The coverage of genes from metagenomic read recruitment did not correlate with the pN/pS values. pN/pS: proportion of non-synonymous to synonymous.



aidA	fibronectin-binding autotransporter adhesin
agxt	transaminase
aldB	aldehyde dehydrogenase
allC	allantoate deiminase
amyA	glycosidase/amylase
c3lhypd	cis-L-3-hydroxyproline dehydratase
cutA	copper tolerance
dehH	haloacetate dehalogenase
dhaK, dhaL	phosphoenolpyruvate: glycerone phosphotransferase
fdh	formate dehydrogenase
frcABC	fructose transport system
fucB	L-fuconolactonase
gdhA	glutamate dehydrogenase
glpA	glycerol-3-phosphate dehydrogenase
gpgS, yedP	glucosylglycerate synthesis
gpr	L-glyceraldehyde 3-phosphate reductase
hupE	nickel transport
isdG	heme oxygenase
katG	catalase-peroxidase
kce, lcdH	L-carnitine degradation
ldhD	D-lactate dehydrogenase
mcl, mtkA	S-malate metabolism
mdcF	malonate transporter
mmcO	copper tolerance
mylcd	malonyl-CoA decarboxylase
nit2	omega-amidase
nupNOPQ	guanosine uptake
ocd	ornithine cyclodeaminase
phnZ	2-aminoethylphosphonic acid degradation
pitA	phosphate/sulfate permease
pucL	uric acid degradation
putA	proline dehydrogenase
puuC	putrescine glutamylation
rcl	nucleoside 2-deoxyribosyltransferase
tadA	tRNA adenosine deaminase
tarD	D-tartrate dehydratase
tdcB	threonine dehydratase
unch	n terminal nucleophile
uraA	xanthine permease
vgsc	voltage-gated sodium channel
xdhAB	xanthine dehydrogenase
xylB	xylulokinase
yedP	mannosyl-3-phosphoglycerate phosphatase
yfdE	CoA:oxalate CoA-transferase
yrbG	cation:H+ antiporter

961

962 **Extended Data Fig. 6**. All genes in a given genome are ordered by lowest pN/pS value to
963 highest, with genes that are differentially distributed between offshore and coastal
964 Pelagibacteraceae genera in Fig. 4 colored red (offshore) or black (coastal) across the genomes.
965 pN/pS values of 0.01, 0.025, 0.05, 0.075, 0.1, and 0.25 are shown in blue. pN/pS: proportion of
966 non-synonymous to synonymous.



Extended Data Fig. 7. *Pelagibacteraceae* has incomplete dissimilatory and assimilatory sulfur reduction pathways and the capacity to utilize a wide diversity of organosulfate compounds. APS: adenylyl sulfate, PAPS: 3'-phosphoadenylyl sulfate, DHSP: 2,3-dihydroxypropane-1-sulfonate, DMSP: dimethylsulfoniopropionate, DMS: dimethyl sulfide, MMPA: methylmercaptopropionate, CoA: coenzyme A.

Supplementary Information

All supplementary infomation text and data files are available on Figshare at:

10.6084/m9.figshare.28087925.