New isolate genomes and global marine metagenomes resolve ecologically relevant units of SAR11

Kelle C. Freel[1], Sarah J. Tucker[1,2,3,4,5], Evan B. Freel[1], Stephen J. Giovannoni[6], A. Murat Eren,[3,4,5,7,8], & Michael S. Rappé[1*]

[1]Hawai‘i Institute of Marine Biology, University of Hawai‘i at Mānoa, Kāne‘ohe, Hawai‘i, United States; [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]Department of Microbiology, Oregon State University, 97331 Corvallis, OR, United States; [7]Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, 26129 Oldenburg, Germany; [8]Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany

*Corresponding Author. Email: rappe@hawaii.edu

Running Title: SAR11 genomes from the tropical Pacific

Abstract

The bacterial order *Pelagibacterales* (SAR11) is among the most abundant and widely distributed microbial lineages across the global surface ocean, where it forms an integral component of the marine carbon cycle. However, the limited availability of high-quality genomes has hampered comprehensive insights into the ecology and evolutionary history of this critical group. Here, we increase the number of complete SAR11 isolate genomes fourfold by describing 81 new SAR11 strains from seven distinct lineages isolated from coastal and offshore surface seawater of the tropical Pacific Ocean. We leveraged comprehensive phylogenomic insights afforded by these isolates to characterize 24 monophyletic, discrete ecotypes with unique spatiotemporal patterns of distribution across the global ocean, which we define as genera. Our data illustrate fine-scale differentiation in patterns of detection with ecologically-relevant gene content variation for some closely related genomes, demonstrating instances of ecological speciation within SAR11 genera. Our study provides unique insight into complex environmental SAR11 populations, and proposes an ecology-informed hierarchy to pave a path forward for the systematic nomenclature for this clade.

Main

SAR11 marine bacteria are a genetically diverse, order-level lineage of heterotrophs within the *Alphaproteobacteria* known as the *Pelagibacterales* (Grote et al. 2012) that numerically dominate planktonic communities across the global ocean (Morris et al. 2002; Carlson et al. 2009; Eiler et al. 2009; Schattenhofer et al. 2009; Becker et al. 2019). Associations between the spatiotemporal distribution of operationally defined subclades and environmental variables suggest the presence of distinct ecotypes within SAR11 (Carlson et al. 2009; Eren et al. 2013a; Delmont et al. 2019; Tucker et al. 2021). Previous studies further support the functional differentiation of subclades (Grote et al. 2012; Thrash et al. 2014), even across short biogeographical distances (Tucker et al. 2024a). While limited in number, the available high-quality SAR11 genomes have demonstrated that this group is a remarkably cohesive genetic assemblage (Grote et al. 2012), making it an attractive model to study the capacity of a minimalist genome to reach stunning levels of success.

Since the first observation of SAR11 through environmental 16S rRNA gene fragments over three decades ago (Giovannoni et al. 1990), microbiology has benefited from a dramatic increase in microbial sequence data recovered directly from the environment, offering representative genomes for many difficult to cultivate microbial lineages (Hug et al. 2016). However, even the most comprehensive genome-resolved surveys of marine metagenomes have failed to yield high-quality SAR11 genomes (Paoli et al. 2022), resulting in limited insights into what constitutes ecologically meaningful units within this broad group. The extensive intra-clade diversity of SAR11 (Tsementzi et al. 2016; Kiefl et al. 2023) confounds the ability to reconstruct environmental genomes from metagenomes (Delmont et al. 2018; Tully et al. 2018), which is why one of the most abundant microbial clades in marine systems suffers from poor representation in genome-resolved metagenomics surveys (Chang et al. 2024). Circumventing the need to assemble complex metagenomes first for genome recovery, single-cell sorting techniques have been much more effective in sampling environmental SAR11 populations through single-amplified genomes (SAGs). However, in an extensive effort to characterize surface ocean microbes, the estimated genome completion of SAGs that could be affiliated with SAR11 remained below 60% (Pachiadaki et al. 2019), a level that prevents robust phylogenomic insights. Such barriers have led to a reliance on isolate genomes to investigate the evolution of SAR11 populations (Vergin et al. 2007; Wilhelm et al. 2007; Thrash et al. 2011; Grote et al. 2012; Muñoz-Gómez et al. 2019), yet following this path has been impeded by another formidable challenge: the difficulty of cultivating SAR11 in the laboratory, even with genomic insights regarding its unique growth requirements (Tripp et al. 2008; Carini et al. 2013; Sun et al. 2016).

The first successful cultivation of SAR11 in 2002 resulted in the isolation of *Pelagibacter ubique* strain HTCC1062 (Rappé et al. 2002), followed by the publication of its complete genome (Giovannoni et al. 2005). Over the past two decades, additional isolate genomes have been few, with only 25 currently available. Despite their rarity, high-quality genomes from isolated strains not only shed light on SAR11 biology (Schwalbach et al. 2010; Sun et al. 2011; Carini et al. 2013) and the origins of this lineage within the *Alphaproteobacteria* (Thrash et al. 2011; Grote et al. 2012; Muñoz-Gómez et al. 2019), but also have made it possible to establish key concepts in biology such as genome streamlining (Schwalbach et al. 2010; Sun et al. 2011;

Grote et al. 2012; Viklund et al. 2012; Giovannoni et al. 2014; Giovannoni 2017) and investigate the evolutionary processes that shape protein evolution (Delmont et al. 2019; Kiefl et al. 2023).

Here we report 81 high-quality genomes from SAR11 strains, increasing the number available for SAR11 isolates by fourfold, and leverage this new collection to build a robust genome phylogeny for the order *Pelagibacterales*. By incorporating publicly available, high-quality single-cell genomes and surface ocean metagenomes from both a steep, nearshore to open-ocean local environmental gradient and elsewhere from around the globe, we reveal cohesive patterns of genomic and ecotypic diversification. We propose a framework through which to characterize and interpret genome heterogeneity at multiple stages along the evolutionary history of SAR11 marine bacteria, and establish a roadmap for future efforts to organize this globally abundant bacterial clade.

Results

Eighty-one high-quality genomes sequenced from 206 newly isolated SAR11 strains and co-cultures

Three dilution-to-extinction culturing experiments using surface seawater collected from nearshore and adjacent offshore environments of Oʻahu, Hawaiʻi, in the tropical Pacific Ocean resulted in 916 isolates from 2,102 inoculated cultures (Table 1; Supplemental Fig. 1). Using a streamlined isolate-to-genome approach, we identified 206 cultures as either pure SAR11 strains or mixed cultures with at least 50% of the total reads matching a SAR11 strain via 16S rRNA gene amplicon sequencing (Supplemental Table 1), and sequenced draft genomes from 90. Manual curation resulted in 79 high-quality SAR11 isolate genomes. The genomes from two strains (HIMB123 and HIMB109) isolated from a previous culture experiment were also added (Brandon 2006), resulting in 81 new SAR11 genomes from isolates. The majority of these (n=60) assembled into ten contigs or less, including 24 closed genomes and an additional 30 containing one to three contigs. They ranged from 1.00 to 1.54 Mbp in size and GC content of 28.5 to 30.7% (Supplemental Table 2). The median pairwise genome-wide average nucleotide identity (gANI) value across all genomes was 81.8% and none of the 81 new isolate genomes were identical. Having captured a genetically diverse array of SAR11 isolates, we used a phylogenomic approach to characterize evolutionary relationships between these genomes and to high-quality single-cell and isolate genomes previously retrieved from seawater.

Table 1. Summary of high-throughput culturing (HTC) experiments.

Site	Inoculum source	Inoculum size (# of cells)	Cultures screened	Positive cultures	SAR11 genomes
SB	raw seawater	5	576	339	53
STO1	raw seawater	5	576	126	16
STO1	cryopreserved seawater	5	480	142	9
STO1	cryopreserved seawater	100	470	343	1

A comprehensive genome phylogeny reveals a robust evolutionary backbone populated by clusters of closely related genomes

We first sought to resolve relationships between the strains isolated in this study and other publicly available high-quality *Pelagibacterales* genomes to precisely establish where the new genomes originate from within the broad spectrum of known SAR11 diversity. For this, we created a database that, in addition to the 81 genomes presented here, included 25 public SAR11 isolate genomes, 8 of which were also isolated from off the windward coast of O‘ahu, Hawai‘i, and 375 SAR11 single-amplified genomes (SAGs) estimated to be $\geq$85% complete with a redundancy <5% (Supplemental Table 3). We also included five additional SAR11 SAGs of potentially unique evolutionary origin in this collection (Vergin et al. 2013; Thrash et al. 2014), though we excluded genomes from putative SAR11 subgroups IV (Vergin et al. 2013) and V (Thrash et al. 2011) due to their unlikely or, at a minimum, uncertain shared common ancestry with SAR11 (Thrash et al. 2011; Viklund et al. 2013; Haro-Moreno et al. 2020; Muñoz-Gómez et al. 2022). This resulted in a curated collection of 481 SAR11 genomes to assess the evolutionary backbone for SAR11.

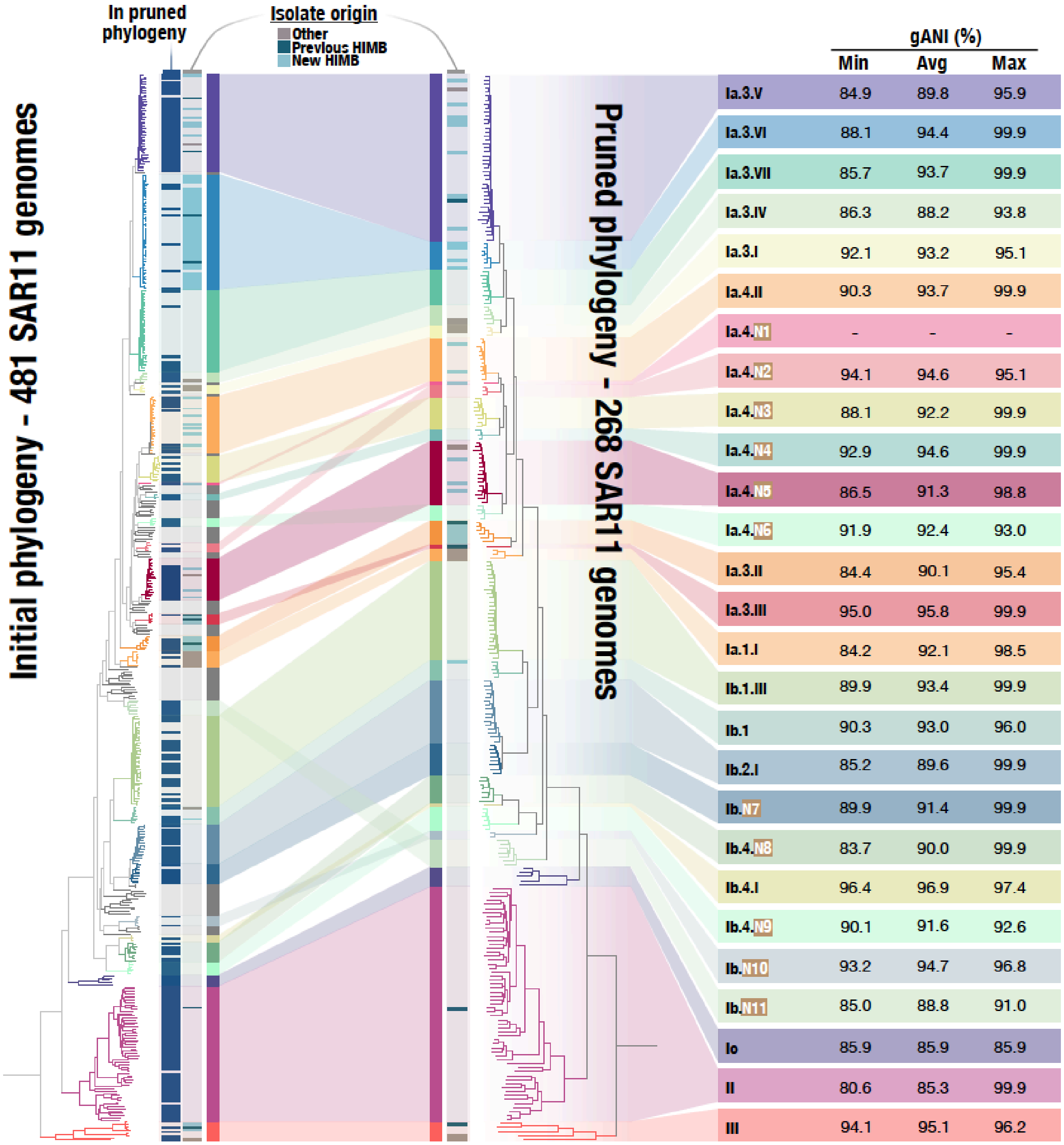
Previous studies investigating phylogenomic relationships within the *Alphaproteobacteria* utilized a curated set of 200 single-copy core genes (SCGs) for this bacterial class (Wang and Wu 2013; Muñoz-Gómez et al. 2019). We evaluated the presence of these 200 SCGs across our genome dataset, and excluded genes missing in more than 90% of the 481 SAR11 genomes. This resulted in a SAR11-specific SCG set of 165 genes for downstream phylogenomic analyses, referred to hereafter as the SAR11_165 core gene set (Supplemental Table 4).

Our analysis of the 481 genomes using the SAR11_165 gene set revealed that the SAR11 clade consists of four robust, deeply-branching sublineages (Fig. 1; Supplementary Fig. 2). Three of these branches were the previously characterized subclades Ic (Vergin et al. 2013), II (Suzuki et al. 2001), and III (Morris et al. 2005), while the fourth was a combination of established SAR11 subclades Ia and Ib (Suzuki et al. 2001), which did not form separate monophyletic subclades in this comprehensive genomic dataset and robust phylogenetic analysis. If the SAR11 clade is assigned to the taxonomic level of a bacterial order, then these four lineages logically resolve to the taxonomic level of families.

We further removed genomes from this initial tree in two steps. First we excluded SAGs that did not fall into a 90% gANI cluster of at least three genomes to focus our analyses on well-resolved regions of the tree. Second, we de-replicated the remaining genomes using a conservative cutoff of 95% gANI to minimize subsequent competitive metagenomic recruitment steps splitting reads among closely related genomes (Evans and Denef 2020). While the 95% ANI cutoff is broadly recognized in contemporary microbiology as a threshold to identify microbial species, it over-splits ecologically and evolutionarily cohesive units in SAR11 and does not delineate species-like groups. We note that the reason behind our use of the 95% ANI in this step of our analysis was solely to establish a technically robust workflow prior to competitive read recruitment rather than a biologically meaningful partitioning of our genomes, a challenge our study focuses on later.

We then turned our attention to the distal end of the phylogeny, which contained a large number of well supported clusters of closely related genomes, particularly within the Ia/Ib subgroup that contained 78 of the 81 new isolate genomes. A phylogeny of the resulting 268 genomes revealed 24 monophyletic clusters within the historical Ia/Ib subgroup that were characterized by a range of gANI values from 84% to 96% (92.1 ± 2.94%; mean ± SD) (Fig. 1, Supplemental Fig. 3). While a handful of these clusters were recognized previously, we defined an additional 11 here (Fig 1; Supplemental Table 5). Twelve of the 24 clusters contained an isolated representative, and eight contained at least one isolate from our study area in the tropical Pacific.

In summary, our extensive phylogenomic analysis of SAR11 revealed 24 monophyletic clusters within the historical Ia/Ib subgroup which included the majority of SAR11 SAGs and the new and previously published isolate genomes. The non-uniform minimum gANI estimates suggest that the application of sequence-based ANI thresholds to demarcate SAR11 diversity may obscure important evolutionary signals. Hypothesized drivers of the maintenance and partitioning of genomic diversity in SAR11 include niche-based processes, where genetically cohesive clusters also display ecological homogeneity and the underlying genetic diversity is maintained by similar forces of selection, recombination, and drift. To understand the potential eco-evolutionary forces that shape SAR11 diversification, we turned to metagenomic read recruitment analysis to recover biogeographical distribution patterns for our genomes across the globe.



	gANI (%)		
	Min	Avg	Max
Ia.3.V	84.9	89.8	95.9
Ia.3.VI	88.1	94.4	99.9
Ia.3.VII	85.7	93.7	99.9
Ia.3.IV	86.3	88.2	93.8
Ia.3.I	92.1	93.2	95.1
Ia.4.II	90.3	93.7	99.9
Ia.4.N1	-	-	-
Ia.4.N2	94.1	94.6	95.1
Ia.4.N3	88.1	92.2	99.9
Ia.4.N4	92.9	94.6	99.9
Ia.4.N5	86.5	91.3	98.8
Ia.4.N6	91.9	92.4	93.0
Ia.3.II	84.4	90.1	95.4
Ia.3.III	95.0	95.8	99.9
Ia.1.I	84.2	92.1	98.5
Ib.1.III	89.9	93.4	99.9
Ib.1	90.3	93.0	96.0
Ib.2.I	85.2	89.6	99.9
Ib.N7	89.9	91.4	99.9
Ib.4.N8	83.7	90.0	99.9
Ib.4.I	96.4	96.9	97.4
Ib.4.N9	90.1	91.6	92.6
Ib.N10	93.2	94.7	96.8
Ib.N11	85.0	88.8	91.0
Io	85.9	85.9	85.9
II	80.6	85.3	99.9
III	94.1	95.1	96.2

Fig 1. Comprehensive phylogenies of the *Pelagibacterales*. A comparison between an exhaustive phylogeny (left panel) with 481 SAR11 genomes (106 isolates and 375 SAGs) and a pruned phylogeny (right panel) with 268 genomes (50 isolates and 218 SAGs), based on a curated SAR11-specific set of 165 genes. Genomes included in the pruned phylogeny are indicated with a dark blue bar in the left panel, and the origin of isolate genomes is indicated for both phylogenies.

Global read recruitment from the surface ocean reveals broadly congruent phylogenetic and ecotypic diversification across SAR11

Our competitive metagenomic read recruitment assessed the distribution of the 268 SAR11 genomes around the globe and relied upon 950 publicly-available marine metagenomes, as well as metagenomes from the Kāneʻohe Bay Time-series (KByT), the location of isolation for the 81 new and 9 of the 25 existing isolate genomes (Supplemental Table 6; Supplemental Table 7). These data enabled us to investigate whether cohesive genomic and ecological groups, or ecotypes, could be discerned by combining SAR11 phylogeny and biogeography.

Our first priority was to establish whether genome clusters within a given SAR11 clade showed cohesive read recruitment profiles across metagenomes, or, in other words, whether the ecological patterns revealed by a single genome were similar to all genomes within the group to which it belonged. Detection values for multiple genomes within a genome cluster showed a high degree of cohesion (Fig. 2; Supplemental Table 8; Supplemental Table 9). For example, representatives from Ia.3.IV, Ia.3.I, Ia.4_II, Ia.4.N2, Ia.4.N5, Ib.1.III, and Ib.4.N9 are particularly consistent within the genome clusters (Fig. 2). Consistent overlap between SAGs and isolate genomes within the same clade demonstrated that both genome types accurately reflect distribution patterns for closely related populations as inferred by phylogeny (Fig. 2). A non-metric multidimensional scaling (NMDS) analysis of the overall detection patterns of genomes across metagenomes consistently grouped genomes within a given clade more closely compared to those that belonged to other genome clusters (Supplemental Fig. 4), further supporting a high degree of intra-clade ecological cohesion.

Our second priority was to establish insights into whether SAR11 genome clusters differed in their biogeographical patterns, and whether genome clusters identified SAR11 populations of distinct ecology. Hierarchical clustering of metagenomes based on SAR11 detection patterns revealed four groups: metagenomes that originated from (1) low-latitude samples, (2) high-latitude samples with low SAR11 diversity, (3) low-latitude samples with high SAR11 diversity, as well as (4) samples from coastal Kāneʻohe Bay (Fig. 2). Many SAR11 genome clusters were indeed differentially distributed across these metagenome groups. For example, Ia.4.II and Ia.4.N5 were only consistently found in groups 3 and 4, while Ia.3.IV was found across group 1 and only in select sites in groups 3 and 4 (Fig 2). However, in multiple cases, the environmental detection patterns of different phylogenomic genome clusters overlapped; while there was some degree of inter-clade ecological differentiation, distinct SAR11 genome clusters frequently co-occurred (Fig. 2, Supplemental Fig. 5). This observation suggests that patterns of distribution alone cannot discern the boundaries of cohesive ecological units within SAR11, a task that evidently requires the integration of biogeographical patterns through metagenomic read recruitment with ancestral relationships among genomes though phylogenomics.

Finally, we used the read recruitment analysis to assign ecological patterns to specific SAR11 genome clusters. While multiple broad patterns were clear from the pairing of the phylogenomic relationships and read recruitment data, we focused our investigation on whether the genome clusters within the SAR11 Ia/Ib lineage that appeared to be confined to the coastal end of the KByT environmental gradient (Ia.3.VI, Ia.3.II, and Ia.3.III; Supplemental Fig. 1; also see (Tucker et al. 2024a)) were similarly constrained to coastal areas globally. Indeed, two of the three genome clusters, Ia.3.II and Ia.3.III, were detected almost exclusively in metagenomes sourced from coastal environments (e.g., KByT, the north coast of Panama, the Chesapeake Bay, and the Atlantic coast of Portugal). Interestingly, while the clade Ia.3.VI was restricted to nearshore metagenomes across KByT, it was well-detected in both coastal and offshore environments in other oceanic regions (Fig. 2). Genome clusters Ia.3.II and Ia.3.III did not include any SAGs and were only composed of isolates from coastal Kāneʻohe Bay. Yet, we could detect them in other oceans, which confirms their global relevance as representatives of SAR11 populations adapted to coastal ecosystems.

Through the combination of global metagenomic read recruitment and phylogenomics, we show that SAR11 genome clusters contain genomes with a high degree of intra-clade ecological cohesion. These genome clusters were often distinguished by their ecological distributions and demonstrated notable inter-clade ecological differentiation. Finally, we applied this framework to understand how SAR11 genetic and ecological diversity partitions among ocean biomes, in particular coastal ocean and open ocean environments.

The integrated ecological and evolutionary framework here is supported by high-quality genomes that span the known diversity of the *Pelagibacterales*, providing a critical opportunity to discern distinct ecologically meaningful genome clusters within SAR11. We show that the 24 distinct genome clusters represent groups sharing cohesive ecological patterns and evolutionary relationships, not at the finest tips of the phylogenomic tree, but at relatively deeper branches that encompass gANI values ranging between 84% and 96%. This suggests it is unlikely that these genome clusters represent SAR11 diversity at the level of ‘species’. This conclusion is further supported by our companion work (Tucker et al. 2024a), which reveals systematic differences in the metabolic potential of SAR11 genome clusters that likely support distinct ecological distributions in immediately adjacent coastal and open ocean surface seawater with habitat-specific metabolic genes that are under higher selective forces. With the combined evidence presented here and in the work by Tucker et al. (2024a) that unite SAR11 diversity into distinct genome clusters with ecotype properties supported by SAR11 phylogenomics, ecology, metabolic potential, as well as population genetics, we argue that the most conceivable taxonomic rank at which SAR11 genome clusters can be described in a conventional framework emerges as the ‘genus’ level.

This genus-level designation is ideal as it encompasses a degree of diversity previously designated by SAR11 subgroups and has the flexibility to account for subtle variation in ecology recognized between closely related genomes. We identified the highest quality genome representatives (electing for isolates when possible) to assign as type genomes for each genus (Fig. 4), which establishes a roadmap to rationally designate new genera as they are identified in the future.

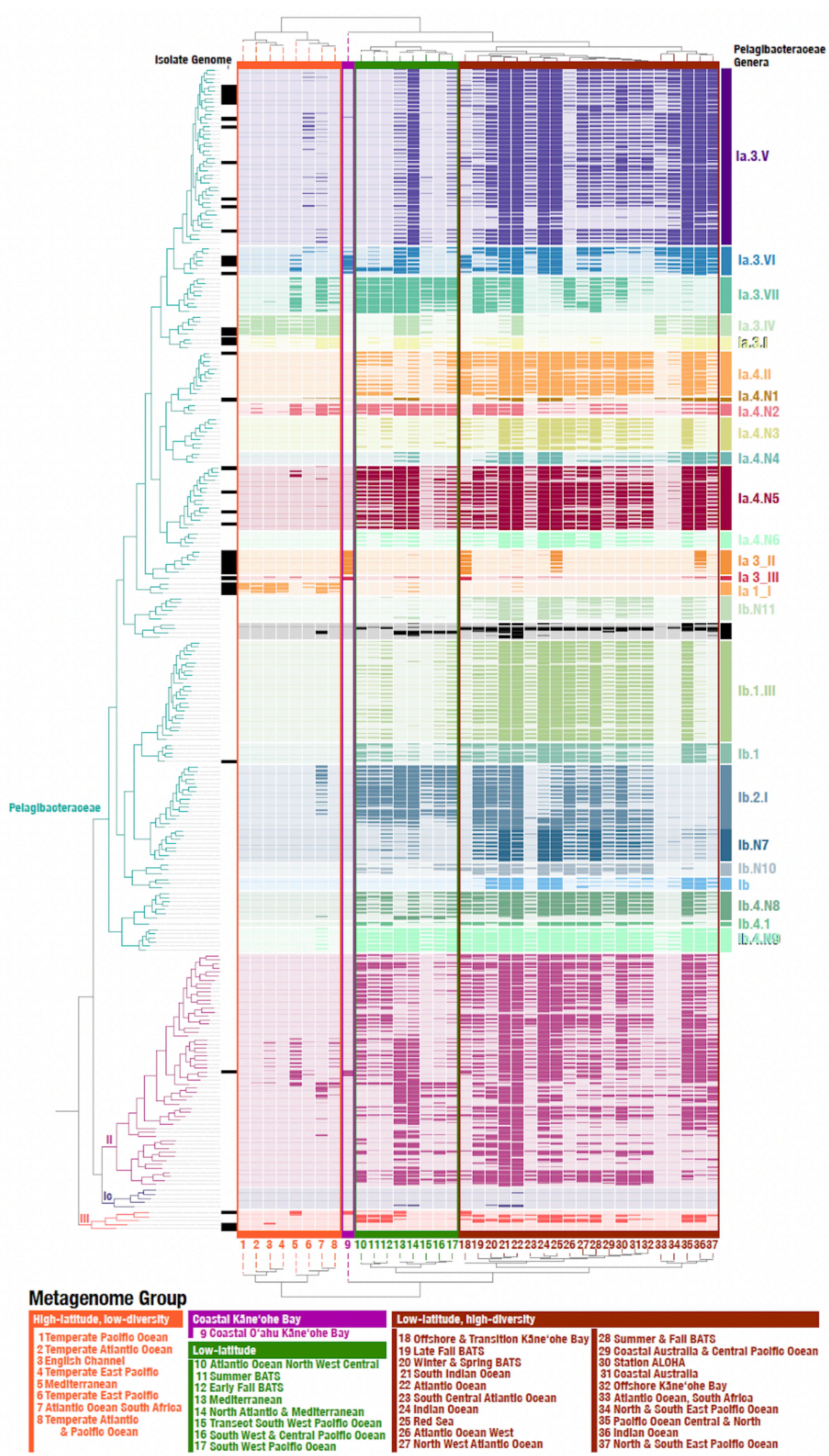

Isolate Genome
Pelagibacteraceae Genera
Ia.3.V
Ia.3.VI
Ia.3.VII
Ia.3.IV
Ia.3.I
Ia.4.II
Ia.4.N1
Ia.4.N2
Ia.4.N3
Ia.4.N4
Ia.4.N5
Ia.4.N6
Ia 3_II
Ia 3_III
Ia 1_I
Ib.N11
Ib.1.III
Ib.1
Ib.2.I
Ib.N7
Ib.N10
Ib
Ib.4.N8
Ib.4.1
Ib.4.N9
Pelagibacteraceae
II
Io
III
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37
Metagenome Group
High-latitude, low-diversity
1 Temperate Pacific Ocean
2 Temperate Atlantic Ocean
3 English Channel
4 Temperate East Pacific
5 Mediterranean
6 Temperate East Pacific
7 Atlantic Ocean South Africa
8 Temperate Atlantic & Pacific Ocean
Coastal Kāne'ohe Bay
9 Coastal O'ahu Kāne'ohe Bay
Low-latitude
10 Atlantic Ocean North West Central
11 Summer BATS
12 Early Fall BATS
13 Mediterranean
14 North Atlantic & Mediterranean
15 Transect South West Pacific Ocean
16 South West & Central Pacific Ocean
17 South West Pacific Ocean
Low-latitude, high-diversity
18 Offshore & Transition Kāne'ohe Bay
19 Late Fall BATS
20 Winter & Spring BATS
21 South Indian Ocean
22 Atlantic Ocean
23 South Central Atlantic Ocean
24 Indian Ocean
25 Red Sea
26 Atlantic Ocean West
27 North West Atlantic Ocean
28 Summer & Fall BATS
29 Coastal Australia & Central Pacific Ocean
30 Station ALOHA
31 Coastal Australia
32 Offshore Kāne'ohe Bay
33 Atlantic Ocean, South Africa
34 North & South East Pacific Ocean
35 Pacific Ocean Central & North
36 Indian Ocean
37 North & South East Pacific Ocean

Fig. 2: Global metagenome read recruitment to 268 *Pelagibacterales* genomes. A clustering analysis reveals that the distribution of metagenomes from the same geographic location have characteristic patterns of detection. Detection values from 0.25 to 0.75 are shown.

Evidence for ecological speciation within closely related genome clusters

Despite broad ecological cohesion within what we have designated as *Pelagibacterales* genera, some notable differences highlight underlying complexities in defining the finest scales of divergence. The Ia.3.VI genus includes genomes from strains of Kāneʻohe Bay origin as well as SAGs from other regions of the global ocean and encompasses significant genomic diversity (minimum gANI 88%) and phylogenomic structure (Fig. 3a). Through read recruitment, we observed notable differences in detection patterns of genomes across metagenomic samples. Isolate genomes from the bay harbored the highest detection values of the Ia.3.VI genus from metagenomes in the bay, while a SAG from the BATS site in the Atlantic Ocean (GCA_902533445.1) had the highest detection values at the BATS site (Fig. 3a), particularly in the summer and fall (Fig. 2).

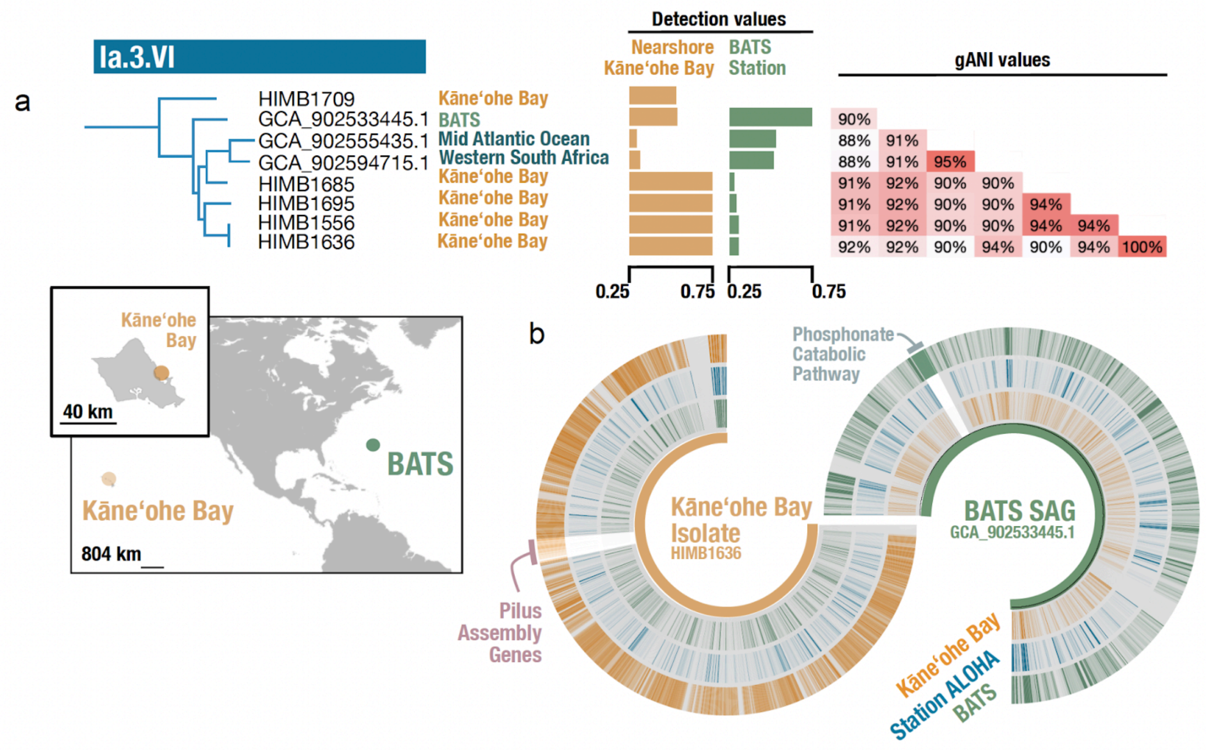


Fig. 3: Fine-scale ecological speciation between closely related SAR11 genomes. (a) Detailed view of the Ia.3.VI genus including evolutionary relationships, locations of genome origin, detection values from select locations, and within-genus gANI values. The geographic origins of two of the closely related genomes that have distinct detection patterns include Kāne'ohe Bay in the Pacific Ocean and the Bermuda Atlantic Time-series Study (BATS) in the Atlantic. **(b)** Coverage values of isolate HIMB1636 and SAG GCA_902533445.1 of metagenomes from nearshore Kāne'ohe Bay, Station ALOHA in the North Pacific Subtropical Gyre, and BATS highlighting the differential detection of genes for type IV pilus assembly and the phosphonate catabolic pathway.

Given the underlying genomic diversification between isolate HIMB1636 and BATS SAG GCA_902533445.1 and their distinct biogeographical distributions that peak in each of their respective source locations, we next surveyed the genomes for potentially unique metabolic capabilities. By inspecting the coverage of isolate HIMB1636 and BATS SAG GCA_902533445.1 using metagenomes from Kāne'ohe Bay and the BATS site (Supplemental Table 10), we found one genomic region of SAG GCA_902533445.1 that had particularly high coverage at BATS compared to the KByT samples and included 29 genes encoding the uptake (*phnCDE*) and catabolism of phosphonates via the C-P lyase pathway (*phnGHIJKLM*) (Fig. 3d) (Villarreal-Chiu et al. 2012). The *phnJ* phylogeny did not reflect the phylogenomic relationships among genomes (Supplemental Fig. 6), and the entire pathway was located on a genomic island similar to the marine bacterium HIMB59 (Molina-Pardines et al. 2023). The C-P lyase pathway is known to be enriched in phosphate-depleted systems of the Atlantic Ocean (Sosa et al. 2019; Acker et al. 2022), so the presence of the C-P lyase catabolic genes in a genome sourced from BATS, but missing from a closely related genome sourced from more phosphate-replete environments of Kāneʻohe Bay in the Pacific, suggests these genes provide an advantage in phosphate depleted systems and that the BATS SAG GCA_902533445.1 may be locally-adapted to these environments.

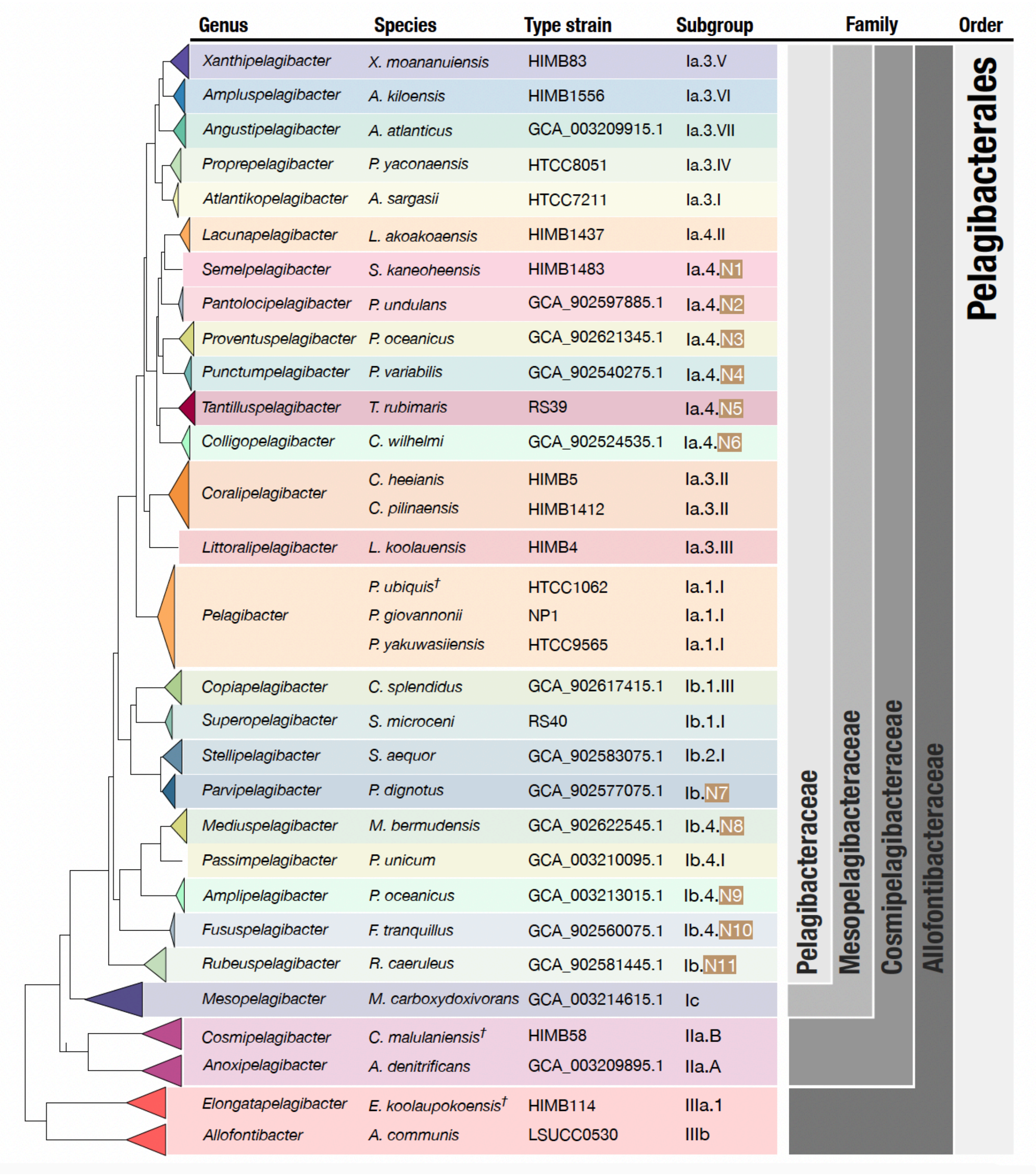
While the HIMB1636 genome lacked the C-P lyase pathway, it contained a unique genomic region with particularly high coverage that was not found in SAG GCA_902533445.1, and encoded genes for type IV pilus assembly. The role of type IV pilus assemblies in SAR11 is unclear (Zhao et al., 2017), although in other organisms it has been associated with an array of functions including DNA uptake, twitching motility, and aggregation into microcolonies (Craig and Li 2008). The presence of the type IV pilus assembly genes in the Ia.3.VI genome sourced from the nitrogen-limited Pacific Ocean, but not in genomes from relatively more nitrogen-replete waters of BATS, along with evidence that the *Pelagibacteraceae* can utilize purine nucleosides and purine-derivatives for nitrogen (Braakman et al. 2024; Tucker et al. 2024a), suggests that the presence of a type IV pilus may be advantageous for DNA uptake in nitrogen-poor environments and that HIMB1636 may be locally-adapted. Contrary to the hypothesis that genera recombine at a rate sufficient to limit the ecotypic diversification of closely related genomes (Zhao et al. 2024), our read mapping instead shows that the HIMB1636 and SAG GCA_902533445.1 genomes within cluster Ia.3.VI have sufficiently diverged at the nucleotide level to reveal clear biogeographic divergence, and that they possess sets of genes that reside in hypervariable genomic regions that are clearly associated with the differences in abundance.

We examined gANI estimates, phylogenetic branching, environmental distributions, and ecologically-relevant gene content to support the characterization of ecological diversification at the finest tips of the tree, a process that we theorize to represent speciation. This underscores the complexity of SAR11 ecology, highlights the need to include a diversity of representative genomes within even closely related genera for environmental genomics studies, and indicates that continued efforts to sample SAR11 globally are key to understanding the distribution of this ubiquitous clade.

Proposed *Pelagibacterales* classification and nomenclature

We leveraged the robust genome phylogeny, gANI metrics, and read recruitment to establish a rational classification and nomenclature system for the *Pelagibacterales* bacterial order *Pelagibacterales*. To provide a framework and vocabulary to discuss groups of SAR11 in a meaningful context, we first defined four family-level monophyletic groups as the *Pelagibacteraceae* (historical subgroups Ia and Ib), *Cosmipelagibacteraceae* (historical subgroup II), *Allofontibacteraceae* (historical subgroup III), and the *Mesopelagibacteraceae* (historical subgroup Ic) (Fig. 4). We focused our efforts primarily on classification within the

Pelagibacteraceae where the majority of cultured isolates originate. Within the *Pelagibacteraceae*, we used phylogenomics and ecological data to characterize 24 genera that represent cohesive genetic and ecological clades, and designate type species for each (Supplemental Table 11). The primary aim of these efforts is to ensure that the taxonomic hierarchy for SAR11 provides a useful and tractable reflection of the ecology and genetic diversity within this globally distributed group, and establishes a rational system that future efforts can build upon.



Genus	Species	Type strain	Subgroup
Xanthipelagibacter	*X. moananuiensis*	HIMB83	Ia.3.V
Ampluspelagibacter	*A. kiloensis*	HIMB1556	Ia.3.VI
Angustipelagibacter	*A. atlanticus*	GCA_003209915.1	Ia.3.VII
Proprepelagibacter	*P. yaconaensis*	HTCC8051	Ia.3.IV
Atlantikopelagibacter	*A. sargasii*	HTCC7211	Ia.3.I
Lacunapelagibacter	*L. akoakoaensis*	HIMB1437	Ia.4.II
Semelpelagibacter	*S. kaneoheensis*	HIMB1483	Ia.4.N1
Pantolocipelagibacter	*P. undulans*	GCA_902597885.1	Ia.4.N2
Proventuspelagibacter	*P. oceanicus*	GCA_902621345.1	Ia.4.N3
Punctumpelagibacter	*P. variabilis*	GCA_902540275.1	Ia.4.N4
Tantilluspelagibacter	*T. rubimaris*	RS39	Ia.4.N5
Colligopelagibacter	*C. wilhelmi*	GCA_902524535.1	Ia.4.N6
Coralipelagibacter	*C. heeianis* *C. pilinaensis*	HIMB5 HIMB1412	Ia.3.II Ia.3.II
Littoralipelagibacter	*L. koolauensis*	HIMB4	Ia.3.III
Pelagibacter	*P. ubiquis*† *P. giovannonii* *P. yakuwasiiensis*	HTCC1062 NP1 HTCC9565	Ia.1.I Ia.1.I Ia.1.I
Copiapelagibacter	*C. splendidus*	GCA_902617415.1	Ib.1.III
Superopelagibacter	*S. microceni*	RS40	Ib.1.I
Stellipelagibacter	*S. aequor*	GCA_902583075.1	Ib.2.I
Parvipelagibacter	*P. dignotus*	GCA_902577075.1	Ib.N7
Mediuspelagibacter	*M. bermudensis*	GCA_902622545.1	Ib.4.N8
Passimpelagibacter	*P. unicum*	GCA_003210095.1	Ib.4.I
Amplipelagibacter	*P. oceanicus*	GCA_003213015.1	Ib.4.N9
Fususpelagibacter	*F. tranquillus*	GCA_902560075.1	Ib.4.N10
Rubeuspelagibacter	*R. caeruleus*	GCA_902581445.1	Ib.N11
Mesopelagibacter	*M. carboxydoxivorans*	GCA_003214615.1	Ic
Cosmipelagibacter	*C. malulaniensis*†	HIMB58	IIa.B
Anoxipelagibacter	*A. denitrificans*	GCA_003209895.1	IIa.A
Elongatapelagibacter	*E. koolaupokoensis*†	HIMB114	IIIa.1
Allofontibacter	*A. communis*	LSUCC0530	IIIb

Figure 4. A proposed taxonomic framework for the SAR11 order *Pelagibacterales*. This schematic SAR11 phylogeny unites proposed genus and species names, proposed type strains, and historical reference labels.

Discussion

By integrating high-throughput cultivation experiments with publicly available genomes and metagenomes, our study provides key insights into a long-standing question: to what extent, and at what hierarchical levels, can the genomic and ecological diversity of SAR11 be partitioned into cohesive units? Through comprehensive phylogenomic analyses paired with global metagenomic read recruitment surveys, we reveal ecotypic differentiation at both relatively shallow, species-level and deeper, genus-level diversity within SAR11. This robust eco-evolutionary framework, which unifies independent yet complementary approaches to genomic diversity and biogeography, resolves the order *Pelagibacterales* into four families and the family *Pelagibacteraceae* into 24 genera, establishing a much-needed taxonomic framework that delineates SAR11 diversity into tractable units and provides a foundation for future investigations.

A tight relationship between the phylogeny and ecology of SAR11 has long been suggested (Brown et al. 2012; Vergin et al. 2013); however, the ability to associate specific SAR11 clades with distinct ecological patterns and explain forces that maintain SAR11 diversity has remained elusive. Focusing on sequence-discrete groups within deep ocean SAR11 lineages, a recent study concluded that recombination, rather than ecological speciation, was likely the major driver of species-level cohesion (Zhao et al., 2024). While this observation may explain forces that maintain species-level cohesion for some populations in this group, our study shows that the global sampling of environmental populations through metagenomes consistently supports ecological delineations that are congruent with phylogenomic clustering patterns, pointing towards ecotypic differentiation as the pervasive driver of the evolution within the *Pelagibacterales*. Interestingly, SAR11 genera that showed similar biogeographical distribution patterns in our analysis tended to occupy distant parts of the tree. This observation suggests an inverse correlation between the genetic similarity among SAR11 populations and their co-occurrence, a trend known as phylogenetic overdispersion. Phylogenetic overdispersion has been observed across the tree of life (Davies 2006) and is driven by forces of competitive exclusion, an overarching ecological phenomenon that limits the co-occurrence of ecologically similar, closely related organisms. Future analyses that aim to resolve specific genetic determinants of competitive exclusion or co-existence may benefit from geographically constrained time-series data, as these patterns are likely not immediately attainable from global yet spatiotemporally sparse metagenomes.

The practical need of microbiologists to find reasonable cutoffs to demarcate species boundaries from genomic data alone and the nature of SAR11 evolution do not align seamlessly. Through the analysis of genomes, a large number of anecdotal observations support 95% ANI as a reasonable means to resolve archaeal and bacterial species (Jain et al. 2018; Olm et al. 2020). However, SAR11 serves as a reminder that practical solutions do not necessarily apply to all microbial clades (Delmont et al. 2019; López-Pérez et al. 2020). One of the implications of the efforts to standardize the tree of life based on principles that work only for the majority of microbial taxa is the conflation of all SAR11 genomes into two genera in the taxonomic framework derived from genomes available on GTDB based on RED scores (Parks et al. 2022). Indeed, while the ecologically relevant units of SAR11 described in our study are in agreement with functional, evolutionary, and ecological observations, they are in disagreement with the contemporary summaries of this clade based on RED- or ANI-based demarcations. The ways in which evolutionary relationships between distinct clades of life intersect with taxonomic classification systems will unlikely be resolved in a manner that satisfies everyone in microbiology (Waite et al. 2020; Sanford et al. 2021). In this juncture, we believe that a stronger motivation to understand the biological drivers that render SAR11 incompatible with our best practical approaches will bring us closer to a unified solution to partition microbial diversity into meaningful units, rather than casting SAR11, one of the most numerous microbial clades on our planet, as a mere outlier.

Insights into the eco-evolutionary processes that shape SAR11 diversification in our study rely heavily on the contribution of 81 new isolate genomes that represent abundant and ecologically-relevant SAR11 populations across the coastal and global ocean. The ecology-informed hierarchical organization of these genomes enabled us to propose SAR11 genera with formal names here, and investigate the likely functional determinants of ecological diversification across the *Pelagibacteraceae* in our companion work (Tucker et al. 2024a). While deeper understanding of the physiological, metabolic, and genetic factors that shape SAR11 biology will require controlled experimentation of isolated strains in the laboratory, our study organizes the eco-evolutionary characteristics of known SAR11 diversity and provides a roadmap for future efforts aimed to organize and understand the ubiquitous SAR11 populations inhabiting the global ocean.

Methods

High-throughput culturing from surface seawater within and adjacent to Kāne‘ohe Bay, O‘ahu

Growth medium was prepared as previously described (Monaghan et al., 2020). Briefly, 20 L of seawater was collected first on 8 July 2017 and again on 20 September 2017 from a depth of 2 meters at station SR4 (N 21º 27.699’, W 157º 47.010’) in acid-washed polycarbonate bottles (Supplemental Fig. 1). The seawater was then filtered, autoclaved, and sparged as previously described (Monaghan et al. 2020). After processing, the sterile seawater was stored at 4ºC until use.

Two 4 L seawater samples to be used as inoculum were collected on 26 July 2017 in acid-washed polycarbonate bottles from 2 meters from stations SB (N 21° 26.181’, W 157° 46.642) and STO1 (N 21° 28.974, W 157° 45.978’) (Supplemental Fig. 1) and immediately returned to the laboratory for further processing. All of the Kāne‘ohe Bay Time series sampling sites were previously classified as ‘nearshore’, ‘transition’, or ‘offshore’, with SB and STO1 representing nearshore and offshore sites, respectively (Tucker et al. 2021). Subsamples of the raw seawater were processed as described previously (Monaghan et al. 2020). Briefly, aliquots were taken for cryopreservation in a final concentration of 10% v/v glycerol and fixed with paraformaldehyde for the enumeration of planktonic microorganisms via flow cytometry. Additionally, 0.96 L from station SB and 1.30 L from station STO1 were filtered through a 25 mm diameter, 0.1 µm pore-sized polyethersulfone membrane (Supor-100; Pall Gelman Inc., Ann

Arbor, MI), which was then submerged in 500 µL DNA lysis buffer and stored at -80ºC until DNA extraction.

Subsamples of raw seawater from SB and STO1 were enumerated using microscopy, diluted to 2.5 cells mL^{-1}, and plated in 2 mL volumes into a total of 1,152 wells (576 wells per site) of custom-fabricated 96-well Teflon microtiter plates. This experiment is referred to here as HTC2017. Plates were then sealed with breathable polypropylene microplate adhesive film and incubated in the dark at 27ºC. Plates were monitored for cellular growth at 3.5 and 8 weeks using flow cytometry as previously described (Tripp et al. 2008; Monaghan et al. 2020). Wells with positive growth (greater than 10^4 cells mL) after 24 or 57 days of incubation were further sub-cultured by transferring approximately 1 mL into 20 mL of sterile seawater media amended as previously described (Monaghan et al. 2020) with 400 µM $(NH_4)_2SO_4$, 400 µM NH_4Cl, 50 µM NaH_2PO_4, 1 µM glycine, 1 µM methionine, 50 µM pyruvate, 800 nM niacin (B3), 425 nM pantothenic acid (B5), 500 nM pyridoxine (B6), 4 nM biotin (B7), 4 nM folic acid (B9), 6 µM myo-inositol, 60 nM 4-aminobenzoic acid, and 6 µM thiamine hydrochloride (B1). These subcultures were then incubated at 27ºC in the dark for an additional 33 days and then all samples were processed and cataloged.

Cultures checked at 33 days that yielded positive growth (>10^4 cells ml^{-1}) were cryopreserved in duplicate (2 x 500 µL culture and a final concentration of 10% v/v glycerol). Each well with positive growth was assigned an HIMB culture ID and cells from the approximately 18 mL remaining volume of each culture were collected by filtration through a 13 mm diameter, 0.03 µm pore-sized polyethersulfone membrane (Sterlitech, Kent, WA, USA), which was then submerged in 250 µL DNA lysis buffer and stored at -80ºC until DNA extraction. The lysis buffer was prepared by adding the following to MilliQ water: 8 mL 1M Tris HCl (pH 8.0), 1.6 ml 0.5M EDTA (pH 8.0), and 4.8 g Triton X, for a final volume of 400 mL, which was then filter sterilized, with lysozyme added to aliquots immediately before use (at a final concentration of 20 mg ml^{-1}).

An additional experiment was performed using cryopreserved samples of seawater collected on July 26, 2017, and described previously (Monaghan et al. 2020). Briefly, the cryopreserved sample was enumerated and then diluted to two cell concentrations (2.5 and 52.5 cells ml^{-1}), and used to plate 480 and 470 2-ml dilution cultures, respectively. This experiment is referred to as HTC2018. Growth was monitored at 2, 3, and 5 weeks after inoculation with positive growth (>10^{4} cells ml^{-1}) from the 2.5 cells ml^{-1} cultures subcultured into 20 ml of sterile seawater growth medium and monitored for growth for up to 10 weeks at 27ºC in the dark. Subcultures were then cryopreserved and cells collected for DNA sequencing as described above. One well from the 52.5 cells ml^{-1} inoculation was directly collected for DNA sequencing without subculturing (Monaghan et al. 2020).

DNA extraction and 16S rRNA gene amplicon sequencing

Genomic DNA (gDNA) from all filtered cultures as well as environmental DNA (eDNA) from STO1 and SB was extracted using the Qiagen DNeasy Blood and Tissue Kit with modified manufacturer’s instructions for bacterial cells (Qiagen, Germantown, Maryland, USA). The modifications included the addition of an initial freeze-thaw step (3 cycles of 10 minutes at 65ºC followed by 10 minutes at -80ºC), the addition of 35 µL Proteinase K and 278 µL buffer AL at the appropriate pretreatment step, and finally when eluted the same 200 µL volume was passed through the membrane three times.

For the initial identification of all cultures, gDNA was used as template for the polymerase chain reaction (PCR) amplification (Bio Rad C1000 Touch, Bio Rad, Hercules, CA, USA) using barcoded 515F and 926R primers targeting the V4 region of the SSU rRNA gene (Parada et al., 2016) in a reaction volume of 25 µL composed of: 2 µL gDNA, 0.5 µL each forward and reverse primer, 10 µL 5PRIME HotMasterMix (Quantabio, Beverly, MA, USA), and 12 µL of molecular grade H_20 (Monaghan et al. 2020). The reaction was as follows: an initial denaturation step of 3 min at 94ºC, 40 cycles of 45 sec at 94ºC followed by 1 min at 50ºC and 1.5 min at 72ºC, with a final extension of 10 min at 72ºC. The PCR products were prepared for sequencing as previously described (Monaghan et al. 2020) and sequenced on a MiSeq platform by the Oregon State University Center for Genome Research and Biocomputing.

16S rRNA gene sequence analysis

Amplicon sequence data were processed as previously described (Monaghan et al. 2020). Briefly, the data was imported into QIIME2 v2019.4.0, and demultiplexed before being assessed for sequence quality and merged. DADA2 (Callahan et al. 2016) was then used for quality control. Taxonomy was assigned to all reads using a Naïve Bayes classifier trained on the Silva rRNA v132 database (Quast et al. 2013). Cultures were first classified as defined previously (Monaghan et al. 2020), with “monocultures" consisting of more than 90% of reads from a single amplicon sequence variant (ASV), “mixed cultures" with an ASV that was between 50% and 90% of the reads, and finally cultures with no dominant members. Any samples with less than 1,000 reads were not included in further analyses. We aimed to sequence all strains that included monocultures and mixed cultures of SAR11.

Genome sequencing

To prepare samples of interest for whole genome sequencing, all extractions with gDNA concentrations above 0.06 ng µL^{-1}, a total of 10 µL were aliquoted for sequencing. For samples with concentrations below 0.06 ng µL^{-1}, the remaining extraction volume (approximately 175 to 185 µL) was concentrated using a SpeedVac (ThermoFisher) to approximately 30 µl and was re-quantified (Qubit 2.0, Invitrogen). From the concentrated samples with a minimum of 0.06 ng µL^{-1}, 10 µL was aliquoted for sequencing. Samples for sequencing were prepared using a Nextera library kit and sequenced on the NextSeq500 platform via a 150 bp paired-end run.

Genomes for previously cultured strains HIMB109 and HIMB123 (Brandon 2006) were sequenced by the Joint Genome Institute. Multiple methods were used to sequence these two strains, including directly using 200 µL of cell culture for library prep as well as using multiple volumes (5, 10, or 20 µL) of culture for multiple displacement analysis (MDA) prior to library preparation. The genomes were evaluated based on completeness, length, number of reads, and total contigs post assembly using SPAdes (Bankevich et al. 2012). An additional assembly using all reads generated from various sequencing attempts per genome was also constructed using the same assembly method, the highest quality genomes based on the metrics above were manually curated and used for additional analyses.

Genome assembly and assessment

Short reads were trimmed with Trim Galore! (https://github.com/FelixKrueger/TrimGalore) and assembled using Unicycler (Wick et al. 2017), which acts as a SPAdes (Bankevich et al. 2012) optimizer with Illumina short read data. Once assembled, reference indexes were built, and read mapping was performed using Bowtie2 with default parameters (Langmead and Salzberg 2012). SAMtools (Li et al. 2009) was used to convert the SAM file to a sorted and indexed BAM file. These initial assemblies and BAM files were used to visualize genomes in anvi'o to check for possible contamination (Eren et al. 2015, 2021). For genomes with contamination, (determined visually as instances where contigs had anomalous GC content or tetranucleotide frequency), suspicious contigs were removed. Redundancy was also used as a way to flag any genomes that needed further curation. After inspection, curated contigs were exported using the program 'anvi-summarize' and reads were re-mapped to the cleaned version of assemblies. The cleaned genomes were processed again for visualization in anvi'o to ensure no erroneous contigs were included. Mapping quality was inspected visually using the Integrative Genomics Viewer (IGV) (Robinson et al. 2011) and Tablet (Milne et al. 2013) and manual curation was undertaken using mapped read data. Manual inspection was used to determine if a circular genome could be considered closed and complete. All contigs shorter than 1000 bp were removed from the genomes that were not closed after final curation, and anvi'o was used to assess final genome completeness and redundancy (Eren et al. 2021).

Phylogenomic analyses

To generate a comprehensive phylogeny of the SAR11 clade, a suite of high-quality genomes were curated. Even with an abundance of metagenomes, the high diversity among SAR11 populations makes constructing reliable MAGs currently unfeasible, so to ensure the phylogeny was as robust as possible, only isolate genomes and SAGs were included. The final set of 493 SAR11 genomes for phylogenetic reconstruction included 81 genomes sequenced in this study, 25 previously published reference genomes, and 387 previously published single amplified genomes (SAGs), in addition to 20 isolate genomes from the family Rhodobacteraceae that were used as an outgroup (Supplemental Table 3). The majority of SAGs included were equal to or greater than 85% complete according to checkM (Parks et al. 2015). However, genomes of lower quality from subclades of SAR11 with no high-quality representatives were included to produce a comprehensive phylogeny, for example SAR11 Ic genomes that ranged from 56.0 to 93.7 percent completion were also included (Thrash et al. 2014) (Supplemental Table 3). Both previously identified subgroup V and IV genomes were excluded from these analyses as subgroup V is not considered to be within SAR11 and the inclusion of subgroup IV has not been rigorously investigated and thus its relationship to the *Pelagibacterales* is questionable (Thrash et al. 2011; Viklund et al. 2013; Haro-Moreno et al. 2020; Muñoz-Gómez et al. 2022).

We compared two gene sets to determine the most appropriate genes to use for phylogenetic reconstructions of the SAR11 clade. This included the bac120 gene set utilized by GTDB-Tk to determine the bacteria guide tree, and a curated gene set of marker genes derived from the 200-genes previously demonstrated to be best fit for the *Alphaproteobacteria* (Muñoz-Gómez et al. 2019) (Supplemental Table 4). To curate the second gene set, we generated a custom HMM profile for the 200 *Alphaproteobacteria* genes with a noise cutoff term of 1×10^{-20}, ran the HMM profile on all genomes using the anvi’o program `anvi-run-hmms`, and generated a presence-absence matrix of genes in this model across genomes using the program `anvi-script-gen-hmm-hits-matrix-across-genomes`. After evaluating the model hits across the genomes matrix, we removed the genes that occurred in less than 90% of the genomes or those that were redundant in more than 2% of the genomes from the *Alphaproteobacteria* 200-gene collection, which resulted in a new collection with 165 genes, which is referred to as the `SAR11_165` throughout our study (Supplemental Table 4). To generate a concatenated alignment of the genes of interest for downstream phylogenomic analyses, a custom HMM source was generated that encompassed the SAR11_165 genes. The program `anvi-get-sequences-for-hmm-hits` with the custom HMM source was then implemented to extract and align genes of interest. The program trimAL 1.3 (Capella-Gutiérrez et al. 2009) was then used to remove all positions that were missing in more than 50% of the genomes. Phylogenies were generated with IQ-Tree v2.1.2 (Minh et al. 2020) with the best fit model (LG+F+R10) chosen using ModelFinder (Kalyaanamoorthy et al. 2017) and 1,000 ultrafast bootstraps. Phylogenies were rerooted appropriately in FigTree, and exported in NEXUS format with the options selected to "Save as currently displayed" and "Include Annotations (NEXUS & JSON only)". Once exported, phylogenies were then compared using the package phytools (Revell 2024) in R (R Development Core Team 2011).

Once the extended phylogeny was established, a subset of SAR11 genomes was used to generate a pruned phylogeny with the SAR_165 gene set. For this, we first used PyANI (Pritchard et al. 2016) to dereplicate all genomes using 95% gANI as a cutoff, then excluded SAGs that did not share at least 90% gANI with a neighboring genome, and finally included 10 genomes from the GTDB that spanned 10 families from the order *Rhodospirillales* as an outgroup prior to recomputing the final phylogenomic tree as described above. The 95% ANI dereplication cutoff was chosen to avoid read splitting during competitive read recruitment and for any clusters in which isolate genomes were available, they were chosen as preferred representatives.

Classification and nomenclature

The extended phylogeny was used to define cohesive genetic clusters at the distal end of the SAR11 tree. Single genomes that did not share at least 90% ANI with a neighboring genome were not classified into genera.

To determine how taxonomic levels across the SAR11 lineage would compare using relative evolutionary distance, we implemented this approach as previously described (Ramfelt et al. 2024). Briefly, a domain-level phylogeny was first constructed using the GTDB-Tk de_novo_workflow (Chaumeil et al. 2019) with SAR11 isolate and SAGs as well as "p__Chloroflexota" as the outgroup. Marker genes were identified from the input genomes using GTDB-Tk `identify`, and then aligned with GTDB-Tk `align` (using the "–skip_gtdb_refs" flag). Finally, a tree was constructed using FastTree v2.1.10 (model WAG+GAMMA) (Price et al. 2010), rooted with the Chloroflexota outgroup. This phylogeny was used as the input for the 'scale_tree' program in PhyloRank v0.1.11 (https://github.com/dparks1134/PhyloRank) to convert branch lengths into relative evolutionary distance (RED). RED values of 0.77 and 0.92 were used to assess how they would align with family and genus-level lineages, respectively. These values were based on the distribution of internal nodes within the SAR11 clade and values used previously for other family and genus-level lineages (Parks et al. 2018).

Read recruitment

To assess the distribution of the newly described strains described in this study and put them into context with previously sequenced genomes, we used a read recruitment approach with globally distributed metagenomes. The SAR11 genomes included in this study were grouped into clusters that shared 95% average nucleotide identity (ANI) or greater and representatives from these 95% gANI groups were then used for read recruitment (n = 314, Supplemental Table 12). Results from read recruitment were extrapolated for the other genomes included in each 95% gANI group.

Metagenomes used for recruitment included those sequenced in Kāne‘ohe Bay (Tucker et al. 2024b), the environment from which the genomes were isolated. Only samples from sites previously categorized as “nearshore” and “offshore” (Tucker et al. 2024b) were used here. Additionally, globally distributed previously published metagenomes were also used including those from TARA Oceans expeditions (Sunagawa et al. 2015), station ALOHA (Mende et al. 2017), GEOTRACERS cruises (Biller et al. 2018), the eastern coast of Japan (Kudo et al. 2018; Yoshitake et al. 2021), Monterey Bay (Mueller et al. 2015), and the ocean sampling day program (Kopf et al. 2015) (Supplemental Table 6 for a list of appropriate references and details regarding metagenomes included).

Once metagenomes were chosen, raw reads were downloaded using 'prefetch' and 'fasterq-dump' in the SRA toolkit. We automated the quality filtering of metagenomes, metagenomic read recruitment, and profiling of recruited reads using the program anvi-run-workflow (Shaiber et al. 2020) with the `--workflow metagenomics` flag, which implements snakemake (Köster and Rahmann 2012) recipes for standard analyses in anvi’o. Briefly, this workflow identified and discarded the noisy sequencing reads in metagenomes using the program `iu-filter-quality-minoche` (Eren et al. 2013b), used SAR11 genomes to competitively recruit short reads from metagenomes using Bowtie2 (Langmead and Salzberg

2012) SAMtools (Li et al. 2009) using the program `anvi-profile`, and finally merge individual profiles into an anvi'o merged profile database using the program `anvi-merge`. The resulting anvi'o merged profile database included essential data, including genome coverages and detection statistics across metagenomes, for our downstream analyses. For coverage, we primarily used the 'mean coverage Q2Q3' statistic, which represents the interquartile average of coverage values where, for any given genome, the lowest 25% and the highest 25% of individual coverage values are trimmed prior to calculating the average coverage from the remaining data points, and thus minimizing the impact of biases due to highly conserved or highly variable regions in the final coverage estimates. Visualization of read recruitment data mapped according to the phylogeny constructed was completed using the program `anvi-interactive` with the `--manual` flag.

Metagenome profile clustering

We performed a cluster analysis of metagenomes based on genome detection values from the read recruitment step using the k-means algorithm, where we determined the `k` by identifying the elbow of the curve of within-cluster sum of square values for increasing values of `k` using the R code shared by Delmont et al. (2019) at https://merenlab.org/data/sar11-saavs/. The results of the clustering analysis were visualized using anvi'o. To investigate how similar detection patterns of genomes within genome clusters were, in addition to how similar or distinct patterns were between genome-clusters, we performed a non-metric multidimensional scaling (NMDS) analysis using the vegan package in R. Any metagenomes with zero detection across all genomes were removed. The NMDS results were visualized using ggplot2 and plotly and an interactive plot was generated with ggplotly.

Investigation of C-P lyase pathway

All genomes included in the extended phylogeny (n=X) were searched using `anvi-search-functions` for the key enzyme in the C-P lyase pathway (*phnJ*) to determine the capacity among high-quality SAR11 genomes to utilize the pathway. The genes upstream and downstream of this essential gene were extracted from all 57 genomes and a pangenome was used to compare the presence and absence of other key genes in the pathway as well as the synteny of this region of the genome. The *phnJ* phylogeny (Supplemental Fig. 6) does not reflect the relationships among genomes as demonstrated by the SAR_165 phylogeny (Fig. 1), which is further evidence that this gene is located on a genomic island as previously described (Molina-Pardines et al. 2023).

Acknowledgments

We thank Aimee Saito and Kerri Lettrell for their generous help with taxonomic names, Rex Malmstrom and Nandita Nath for sequencing the genomes of isolates HIMB109 and HIMB123, SeqCoast (formerly the Microbial Genome Sequencing Center) for sequencing isolate genomes, Rachel Ouye for her help with HTC experiments, and Oscar Ramfelt for informatic support. This research was supported by funding from the National Science Foundation to MSR (OCE-1538628 and OCE-2149128). This is SOEST contribution xxx and HIMB contribution xxx.

Competing interests

The authors declare no competing interests.

Data availability

The assembled sequence data for genomes reported here are available at FigShare at https://doi.org/10.6084/m9.figshare.28087454.v1.

References

Acker, M., S. L. Hogle, P. M. Berube, T. Hackl, A. Coe, R. Stepanauskas, S. W. Chisholm, and D. J. Repeta. 2022. Phosphonate production by marine microbes: Exploring new sources and potential function. Proc. Natl. Acad. Sci. U. S. A. **119**: e2113386119.

Bankevich, A. and others. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. **19**: 455–477.

Becker, J. W., S. L. Hogle, K. Rosendo, and S. W. Chisholm. 2019. Co-culture and biogeography of Prochlorococcus and SAR11. ISME J. **13**: 1506–1519.

Biller, S. J. and others. 2018. Marine microbial metagenomes sampled across space and time. Sci Data **5**: 180176.

Braakman, R. and others. 2024. Global niche partitioning of purine and pyrimidine cross-feeding among ocean microbes. bioRxiv. doi:10.1101/2024.02.09.579562

Brandon, M. L. 2006. High-throughput isolation of pelagic marine macteria from the coastal subtropical

Pacific Ocean. University of Hawai‘i at Mānoa.

Brown, M. V. and others. 2012. Global biogeography of SAR11 marine bacteria. Mol. Syst. Biol. **8**: 595.

Callahan, B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods **13**: 581–583.

Capella-Gutiérrez, S., J. M. Silla-Martínez, and T. Gabaldón. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics **25**: 1972–1973.

Carini, P., L. Steindler, S. Beszteri, and S. J. Giovannoni. 2013. Nutrient requirements for growth of the extreme oligotroph “Candidatus Pelagibacter ubique” HTCC1062 on a defined medium. ISME J. **7**: 592–602.

Carlson, C. A., R. Morris, R. Parsons, A. H. Treusch, S. J. Giovannoni, and K. Vergin. 2009. Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. ISME J. **3**: 283–295.

Chang, T., G. S. Gavelis, J. M. Brown, and R. Stepanauskas. 2024. Genomic representativeness and chimerism in large collections of SAGs and MAGs of marine prokaryoplankton. Microbiome **12**: 126.

Chaumeil, P.-A., A. J. Mussig, P. Hugenholtz, and D. H. Parks. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics **36**: 1925–1927.

Craig, L., and J. Li. 2008. Type IV pili: paradoxes in form and function. Curr. Opin. Struct. Biol. **18**: 267–277.

Davies, T. J. 2006. Evolutionary ecology: when relatives cannot live together. Curr. Biol. **16**: R645–7.

Delmont, T. O. and others. 2018. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. Nat Microbiol **3**: 804–813.

Delmont, T. O., E. Kiefl, O. Kilinc, O. C. Esen, I. Uysal, M. S. Rappé, S. Giovannoni, and A. M. Eren. 2019. Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. Elife **8**. doi:10.7554/eLife.46497

Eiler, A., D. H. Hayakawa, M. J. Church, D. M. Karl, and M. S. Rappé. 2009. Dynamics of the SAR11 bacterioplankton lineage in relation to environmental conditions in the oligotrophic North Pacific subtropical gyre. Environ. Microbiol. **11**: 2291–2300.

Eren, A. M. and others. 2021. Community-led, integrated, reproducible multi-omics with anvi’o. Nat Microbiol **6**: 3–6.

Eren, A. M., Ö. C. Esen, C. Quince, J. H. Vineis, H. G. Morrison, M. L. Sogin, and T. O. Delmont. 2015. Anvi’o: an advanced analysis and visualization platform for 'omics data. PeerJ **3**: e1319.

Eren, A. M., L. Maignien, W. J. Sul, L. G. Murphy, S. L. Grim, H. G. Morrison, and M. L. Sogin. 2013a. Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol. Evol. **4**: 1111–1119.

Eren, A. M., J. H. Vineis, H. G. Morrison, and M. L. Sogin. 2013b. A filtering method to generate high quality short reads using illumina paired-end technology. PLoS One **8**: e66643.

Evans, J. T., and V. J. Denef. 2020. To dereplicate or not to dereplicate? mSphere **5**. doi:10.1128/mSphere.00971-19

Giovannoni, S. J. and others. 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science **309**: 1242–1245.

Giovannoni, S. J. 2017. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. Ann. Rev. Mar. Sci. **9**: 231–255.

Giovannoni, S. J., T. B. Britschgi, C. L. Moyer, and K. G. Field. 1990. Genetic diversity in Sargasso Sea bacterioplankton. Nature **345**: 60–63.

Giovannoni, S. J., J. Cameron Thrash, and B. Temperton. 2014. Implications of streamlining theory for microbial ecology. ISME J. **8**: 1553–1565.

Grote, J., J. C. Thrash, M. J. Huggett, Z. C. Landry, P. Carini, S. J. Giovannoni, and M. S. Rappé. 2012.

Streamlining and core genome conservation among highly divergent members of the SAR11 clade. MBio **3**. doi:10.1128/mBio.00252-12

Haro-Moreno, J. M. and others. 2020. Ecogenomics of the SAR11 clade. Environ. Microbiol. **22**: 1748–1763.

Hug, L. A. and others. 2016. A new view of the tree of life. Nat Microbiol **1**: 16048.

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

Kalyaanamoorthy, S., B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermiin. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods **14**: 587–589.

Kiefl, E., O. C. Esen, S. E. Miller, K. L. Kroll, A. D. Willis, M. S. Rappé, T. Pan, and A. M. Eren. 2023. Structure-informed microbial population genetics elucidate selective pressures that shape protein evolution. Sci Adv **9**: eabq4632.

Kopf, A. and others. 2015. The ocean sampling day consortium. Gigascience **4**: 27.

Köster, J., and S. Rahmann. 2012. Building and documenting workflows with python-based Snakemake. GCB 49–56.

Kudo, T. and others. 2018. Seasonal changes in the abundance of bacterial genes related to dimethylsulfoniopropionate catabolism in seawater from Ofunato Bay revealed by metagenomic analysis. Gene **665**: 174–184.

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**: 357–359.

Li, H. and others. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078–2079.

López-Pérez, M., J. M. Haro-Moreno, F. H. Coutinho, M. Martinez-Garcia, and F. Rodriguez-Valera. 2020. The evolutionary success of the marine bacterium SAR11 analyzed through a metagenomic perspective. mSystems **5**. doi:10.1128/mSystems.00605-20

Mende, D. R., J. A. Bryant, F. O. Aylward, J. M. Eppley, T. Nielsen, D. M. Karl, and E. F. DeLong. 2017. Environmental drivers of a microbial genomic transition zone in the ocean's interior. Nat Microbiol **2**: 1367–1373.

Milne, I., G. Stephen, M. Bayer, P. J. A. Cock, L. Pritchard, L. Cardle, P. D. Shaw, and D. Marshall. 2013. Using Tablet for visual exploration of second-generation sequencing data. Brief. Bioinform. **14**: 193–202.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, and R. Lanfear. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. **37**: 1530–1534.

Molina-Pardines, C., J. M. Haro-Moreno, and M. López-Pérez. 2023. Phosphate-related genomic islands as drivers of environmental adaptation in the streamlined marine alphaproteobacterial HIMB59. mSystems **8**: e0089823.

Monaghan, E. A., K. C. Freel, and M. S. Rappé. 2020. Isolation of SAR11 Marine Bacteria from Cryopreserved Seawater. mSystems **5**. doi:10.1128/mSystems.00954-20

Morris, R. M., M. S. Rappé, S. A. Connon, K. L. Vergin, W. A. Siebold, C. A. Carlson, and S. J. Giovannoni. 2002. SAR11 clade dominates ocean surface bacterioplankton communities. Nature **420**: 806–810.

Morris, R. M., K. L. Vergin, J.-C. Cho, M. S. Rappé, C. A. Carlson, and S. J. Giovannoni. 2005. Temporal and spatial response of bacterioplankton lineages to annual convective overturn at the Bermuda Atlantic Time-series Study site. Limnol. Oceanogr. **50**: 1687–1696.

Mueller, R. S. and others. 2015. Metagenome sequencing of a coastal marine microbial community from monterey bay, california. Genome Announc. **3**. doi:10.1128/genomeA.00341-15

Muñoz-Gómez, S. A., S. Hess, G. Burger, B. F. Lang, E. Susko, C. H. Slamovits, and A. J. Roger. 2019.

An updated phylogeny of the Alphaproteobacteria reveals that the parasitic Rickettsiales and Holosporales have independent origins. Elife **8**. doi:10.7554/eLife.42535

Muñoz-Gómez, S. A., E. Susko, K. Williamson, L. Eme, C. H. Slamovits, D. Moreira, P. López-García, and A. J. Roger. 2022. Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. Nat. Ecol. Evol. **6**: 253–262.

Olm, M. R., A. Crits-Christoph, S. Diamond, A. Lavy, P. B. Matheus Carnevali, and J. F. Banfield. 2020. Consistent metagenome-derived metrics verify and delineate bacterial species boundaries. mSystems **5**. doi:10.1128/mSystems.00731-19

Pachiadaki, M. G. and others. 2019. Charting the complexity of the marine microbiome through single-cell genomics. Cell **179**: 1623–1635.e11.

Paoli, L. and others. 2022. Biosynthetic potential of the global ocean microbiome. Nature **607**: 111–118.

Parks, D. H., M. Chuvochina, C. Rinke, A. J. Mussig, P.-A. Chaumeil, and P. Hugenholtz. 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. **50**: D785–D794.

Parks, D. H., M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, and P. Hugenholtz. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. **36**: 996–1004.

Parks, D. H., M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. **25**: 1043–1055.

Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS One **5**: e9490.

Pritchard, L., R. H. Glover, S. Humphris, J. G. Elphinstone, and I. K. Toth. 2016. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Anal.

Methods **8**: 12–24.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. **41**: D590–6.

Ramfelt, O., K. C. Freel, S. J. Tucker, O. D. Nigro, and M. S. Rappé. 2024. Isolate-anchored comparisons reveal evolutionary and functional differentiation across SAR86 marine bacteria. ISME J. **18**. doi:10.1093/ismejo/wrae227

Rappé, M. S., S. A. Connon, K. L. Vergin, and S. J. Giovannoni. 2002. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature **418**: 630–633.

R Development Core Team, R. 2011. R: A Language and Environment for Statistical Computing,.

Revell, L. J. 2024. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). PeerJ **12**: e16505.

Robinson, J. T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. 2011. Integrative genomics viewer. Nat. Biotechnol. **29**: 24–26.

Sanford, R. A., K. G. Lloyd, K. T. Konstantinidis, and F. E. Löffler. 2021. Microbial taxonomy run amok. Trends Microbiol. **29**: 394–404.

Schattenhofer, M., B. M. Fuchs, R. Amann, M. V. Zubkov, G. A. Tarran, and J. Pernthaler. 2009. Latitudinal distribution of prokaryotic picoplankton populations in the Atlantic Ocean. Environ. Microbiol. **11**: 2078–2093.

Schwalbach, M. S., H. J. Tripp, L. Steindler, D. P. Smith, and S. J. Giovannoni. 2010. The presence of the glycolysis operon in SAR11 genomes is positively correlated with ocean productivity. Environ. Microbiol. **12**: 490–500.

Shaiber, A. and others. 2020. Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. Genome Biol. **21**: 292.

Sosa, O. A., D. J. Repeta, E. F. DeLong, M. D. Ashkezari, and D. M. Karl. 2019. Phosphate-limited ocean regions select for bacterial populations enriched in the carbon-phosphorus lyase pathway for phosphonate degradation. Environ. Microbiol. **21**: 2402–2414.

Sunagawa, S. and others. 2015. Ocean plankton. Structure and function of the global ocean microbiome. Science **348**: 1261359.

Sun, J. and others. 2016. The abundant marine bacterium Pelagibacter simultaneously catabolizes dimethylsulfoniopropionate to the gases dimethyl sulfide and methanethiol. Nat Microbiol **1**: 16065.

Sun, J., L. Steindler, J. C. Thrash, K. H. Halsey, D. P. Smith, A. E. Carter, Z. C. Landry, and S. J. Giovannoni. 2011. One carbon metabolism in SAR11 pelagic marine bacteria. PLoS One **6**: e23973.

Suzuki, M. T., O. Béjà, L. T. Taylor, and E. F. Delong. 2001. Phylogenetic analysis of ribosomal RNA operons from uncultivated coastal marine bacterioplankton. Environ. Microbiol. **3**: 323–331.

Thrash, J. C. and others. 2011. Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. Sci. Rep. **1**: 13.

Thrash, J. C., B. Temperton, B. K. Swan, Z. C. Landry, T. Woyke, E. F. DeLong, R. Stepanauskas, and S. J. Giovannoni. 2014. Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. ISME J. **8**: 1440–1451.

Tripp, H. J., J. B. Kitner, M. S. Schwalbach, J. W. H. Dacey, L. J. Wilhelm, and S. J. Giovannoni. 2008. SAR11 marine bacteria require exogenous reduced sulphur for growth. Nature **452**: 741–744.

Tsementzi, D. and others. 2016. SAR11 bacteria linked to ocean anoxia and nitrogen loss. Nature **536**: 179–183.

Tucker, S. J., K. C. Freel, A. M. Eren, and M. S. Rappe. 2024a. Habitat-specificity in SAR11 is associated with a handful of genes under high selection. bioRxiv. doi:10.1101/2024.12.23.630198

Tucker, S. J., K. C. Freel, E. A. Monaghan, C. E. S. Sullivan, O. Ramfelt, Y. M. Rii, and M. S. Rappé. 2021. Spatial and temporal dynamics of SAR11 marine bacteria across a nearshore to offshore transect in the tropical Pacific Ocean. PeerJ **9**: e12274.

Tucker, S. J., Y. M. Rii, K. C. Freel, K. Kotubetey, A. H. Kawelo, K. B. Winter, and M. S. Rappe. 2024b. Sharp transitions in phytoplankton communities across estuarine to open ocean waters of the tropical Pacific. bioRxiv. doi: 10.1101/2024.05.23.595464v1.

Tully, B. J., E. D. Graham, and J. F. Heidelberg. 2018. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data **5**: 170203.

Vergin, K. L. and others. 2013. High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. ISME J. **7**: 1322–1332.

Vergin, K. L., H. J. Tripp, L. J. Wilhelm, D. R. Denver, M. S. Rappé, and S. J. Giovannoni. 2007. High intraspecific recombination rate in a native population of Candidatus pelagibacter ubique (SAR11). Environ. Microbiol. **9**: 2430–2440.

Viklund, J., T. J. G. Ettema, and S. G. E. Andersson. 2012. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. Mol. Biol. Evol. **29**: 599–615.

Viklund, J., J. Martijn, T. J. G. Ettema, and S. G. E. Andersson. 2013. Comparative and phylogenomic evidence that the alphaproteobacterium HIMB59 is not a member of the oceanic SAR11 clade. PLoS One **8**: e78858.

Villarreal-Chiu, J. F., J. P. Quinn, and J. W. McGrath. 2012. The genes and enzymes of phosphonate metabolism by bacteria, and their distribution in the marine environment. Front. Microbiol. **3**: 19.

Waite, D. W. and others. 2020. Proposal to reclassify the proteobacterial classes Deltaproteobacteria and Oligoflexia, and the phylum Thermodesulfobacteria into four phyla reflecting major functional capabilities. Int. J. Syst. Evol. Microbiol. **70**: 5972–6016.

Wang, Z., and M. Wu. 2013. A phylum-level bacterial phylogenetic marker database. Mol. Biol. Evol. **30**: 1258–1262.

Wick, R. R., L. M. Judd, C. L. Gorrie, and K. E. Holt. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput. Biol. **13**: e1005595.

Wilhelm, L. J., H. J. Tripp, S. A. Givan, D. P. Smith, and S. J. Giovannoni. 2007. Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. Biol. Direct **2**: 27.

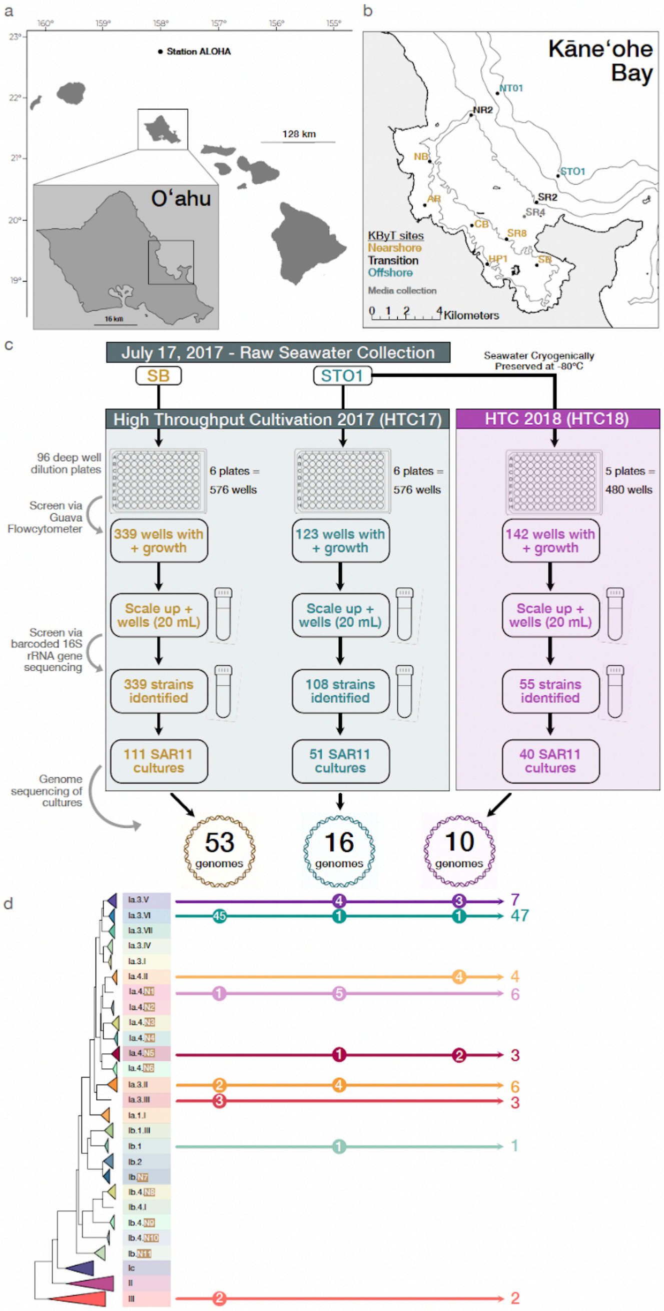
Yoshitake, K. and others. 2021. Development of a time-series shotgun metagenomics database for monitoring microbial communities at the Pacific coast of Japan. Sci. Rep. **11**: 12222.

Zhao, J. and others. 2024. Promiscuous and unbiased recombination underlies the sequence-discrete species of the SAR11 lineage in the deep ocean. bioRxiv. doi:10.1101/2024.10.30.621061

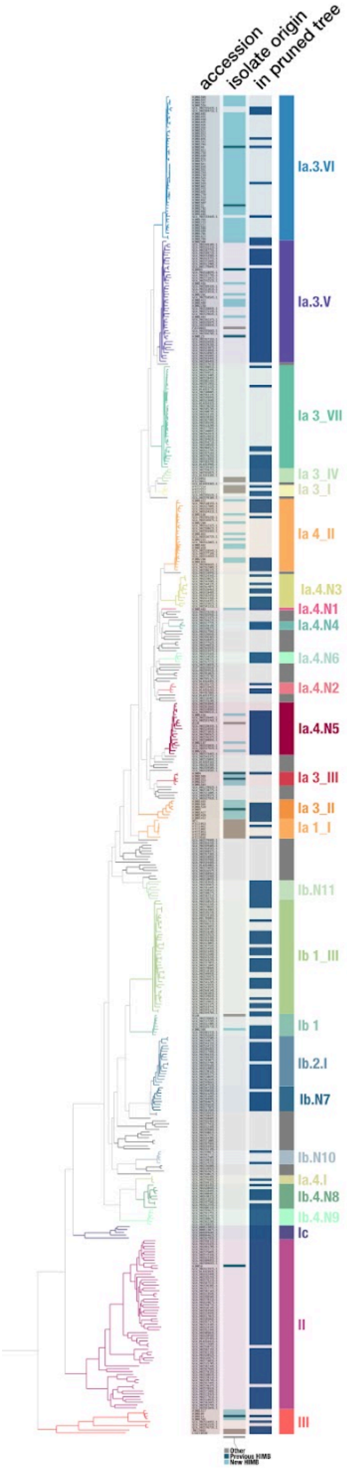
Supplemental Figures

All supplemental figures are available on FigShare at:

https://doi.org/10.6084/m9.figshare.28087760.



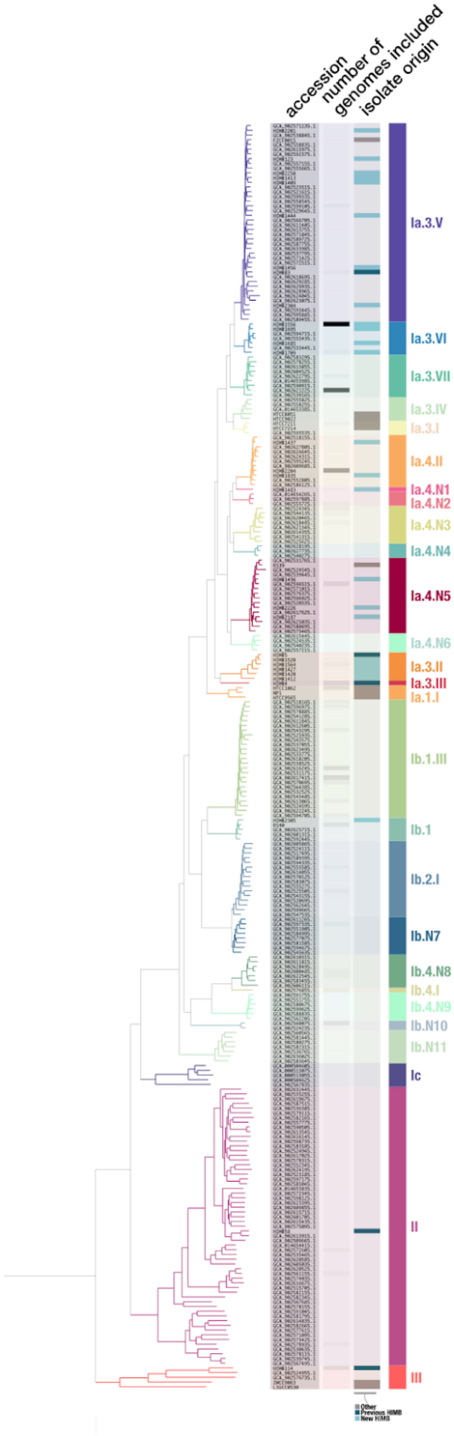
Supplemental Figure 1. **Sampling sites used for high-throughput culturing experiments. (a)** Location of O‘ahu in the Hawaiian archipelago in relations to Station ALOHA approximately 100 km north. **(b)** Map of the embayment on the windward side of O‘ahu with sites included in the Kāne‘ohe Bay Time-series with sites classified as ‘nearshore’ (orange text), ‘transition’ (black text), or ‘offshore’ (turquoise text). Site SR4 in gray from which seawater media was collected for the cultivation experiments is also indicated. Bathymetry lines are approximate. **(c)** Flowchart outlining the high throughput cultivation (HTC) experiments conducted in 2017 and 2018 leading to the isolation of hundreds of SAR11 cultures and 79 new SAR11 isolate genomes. **(d)** Schematic SAR11 phylogeny to indicate which samples harbored genomes from which subgroups.



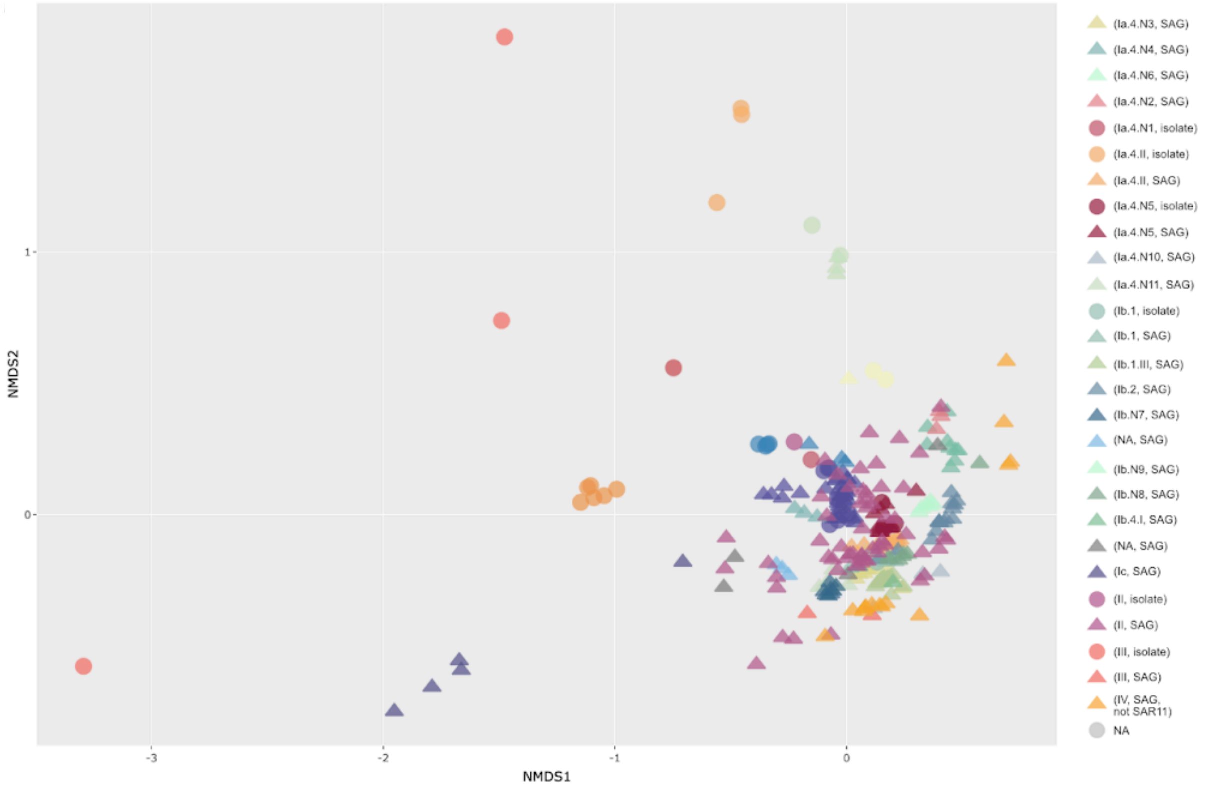
933

934

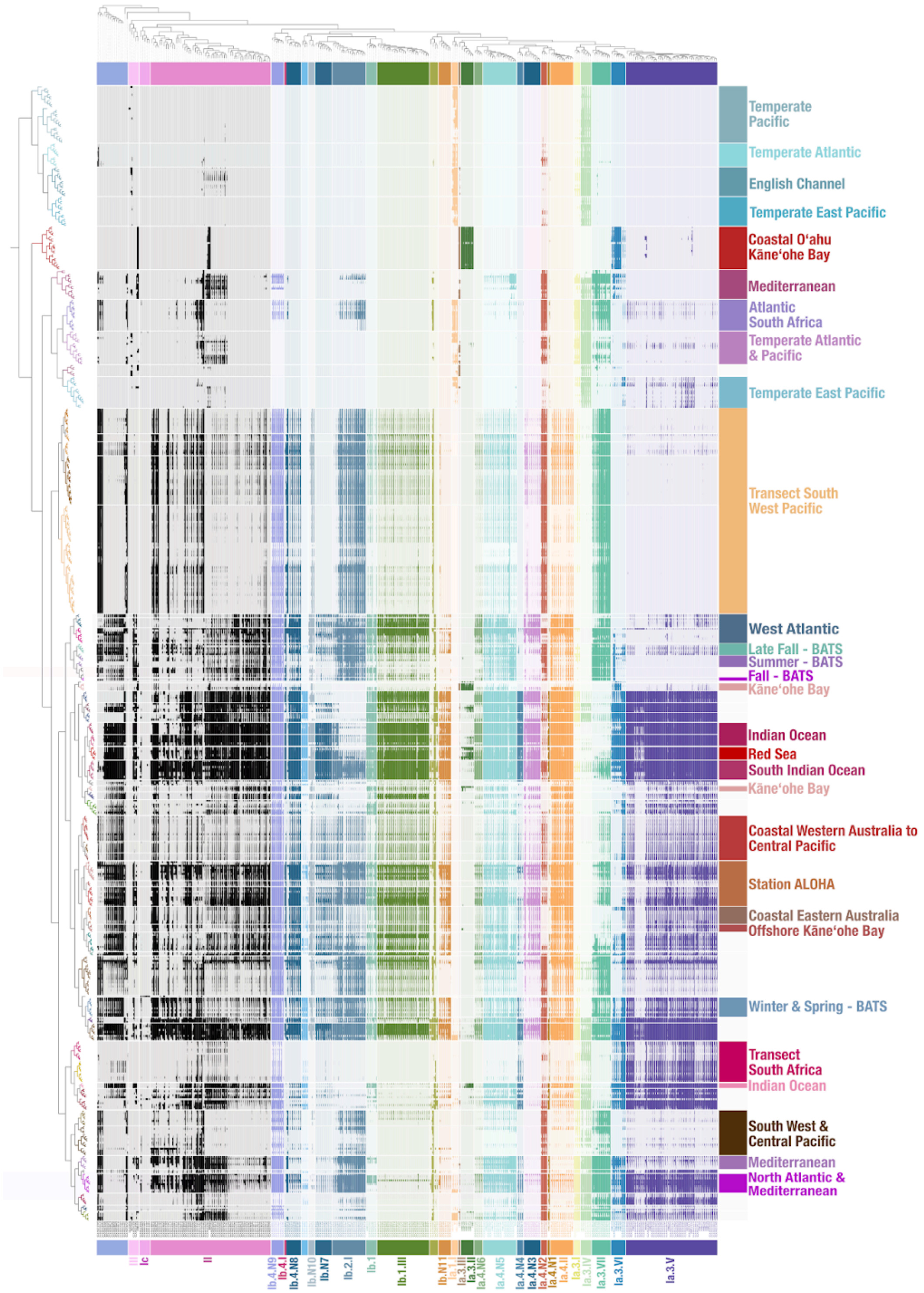
935 **Supplemental Figure 2. Phylogenomic tree of all 481 *Pelagibacterales* genomes initially included.** Of the 481 SAR11 genomes 106 were isolates and 375 SAGs and the phylogeny is based on a curated SAR11-specific set of 165 genes. Isolate origin is indicated and indicates if the genome was from this study, a previous isolate from Kāne'ohe Bay, or from another source. Genomes included in the pruned tree are also indicated.



Supplemental Figure 3. Pruned phylogenomic tree with 268 *Pelagibacterales* genomes. Of the 268 genomes 50 were isolates and 218 SAGs based on a curated SAR11-specific set of 165 genes. Isolate origin is highlighted on the tree and indicates if the genome was from this study, a previous isolate from Kāne‘ohe Bay, or from another source. Number of additional genomes in the same 95% gANI cluster are indicated as well by intensity of the bar which range from 0 to 45.



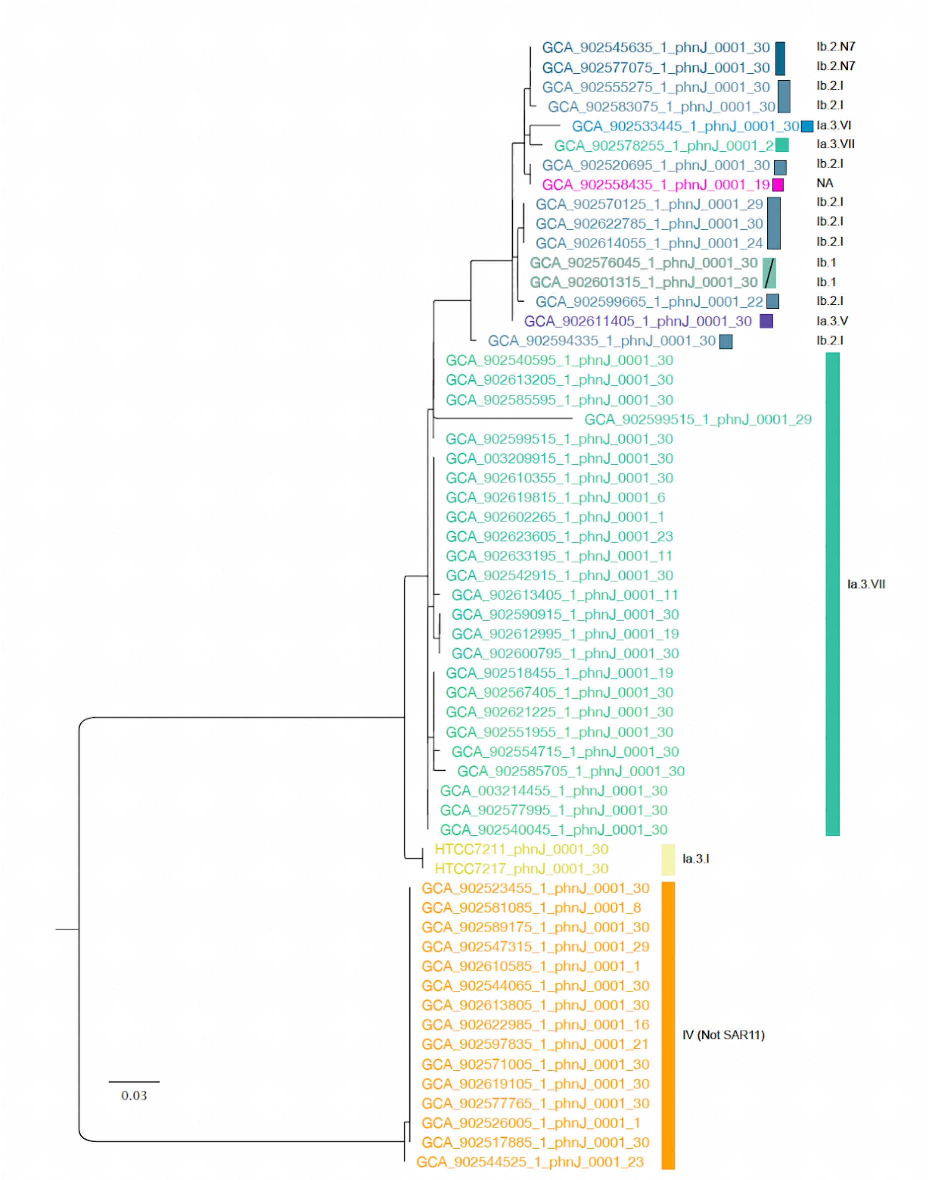
Supplemental Figure 4. NMDS of genomes with detection data. All of the genomes included in the analysis are included here, with distinction between isolate genome (circle) or SAG (triangle) indicated. Genomes not designated a subgroup noted as NA.



952

953

954 **Supplemental Figure 5. Read recruitment data including all metagenomes included in**
955 **cluster analysis of metagenomes based on genome detection values using the k-means**
956 **algorithm.** Subgroups are indicated at the bottom of the figure with labels indicating the
957 metagenome groups along the right hand side of the figure.


GCA_902545635_1_phnJ_0001_30
Ib.2.N7
GCA_902577075_1_phnJ_0001_30
Ib.2.N7
GCA_902555275_1_phnJ_0001_30
Ib.2.I
GCA_902583075_1_phnJ_0001_30
Ib.2.I
GCA_902533445_1_phnJ_0001_30
Ia.3.VI
GCA_902578255_1_phnJ_0001_2
Ia.3.VII
GCA_902520695_1_phnJ_0001_30
Ib.2.I
GCA_902558435_1_phnJ_0001_19
NA
GCA_902570125_1_phnJ_0001_29
Ib.2.I
GCA_902622785_1_phnJ_0001_30
Ib.2.I
GCA_902614055_1_phnJ_0001_24
Ib.2.I
GCA_902576045_1_phnJ_0001_30
Ib.1
GCA_902601315_1_phnJ_0001_30
Ib.1
GCA_902599665_1_phnJ_0001_22
Ib.2.I
GCA_902611405_1_phnJ_0001_30
Ia.3.V
GCA_902594335_1_phnJ_0001_30
Ib.2.I
GCA_902540595_1_phnJ_0001_30
GCA_902613205_1_phnJ_0001_30
GCA_902585595_1_phnJ_0001_30
GCA_902599515_1_phnJ_0001_29
GCA_902599515_1_phnJ_0001_30
GCA_003209915_1_phnJ_0001_30
GCA_902610355_1_phnJ_0001_30
GCA_902619815_1_phnJ_0001_6
GCA_902602265_1_phnJ_0001_1
GCA_902623605_1_phnJ_0001_23
GCA_902633195_1_phnJ_0001_11
GCA_902542915_1_phnJ_0001_30
GCA_902613405_1_phnJ_0001_11
GCA_902590915_1_phnJ_0001_30
GCA_902612995_1_phnJ_0001_19
GCA_902600795_1_phnJ_0001_30
GCA_902518455_1_phnJ_0001_19
GCA_902567405_1_phnJ_0001_30
GCA_902621225_1_phnJ_0001_30
GCA_902551955_1_phnJ_0001_30
GCA_902554715_1_phnJ_0001_30
GCA_902585705_1_phnJ_0001_30
GCA_003214455_1_phnJ_0001_30
GCA_902577995_1_phnJ_0001_30
GCA_902540045_1_phnJ_0001_30
Ia.3.VII
HTCC7211_phnJ_0001_30
HTCC7217_phnJ_0001_30
Ia.3.I
GCA_902523455_1_phnJ_0001_30
GCA_902581085_1_phnJ_0001_8
GCA_902589175_1_phnJ_0001_30
GCA_902547315_1_phnJ_0001_29
GCA_902610585_1_phnJ_0001_1
GCA_902544065_1_phnJ_0001_30
GCA_902613805_1_phnJ_0001_30
GCA_902622985_1_phnJ_0001_16
GCA_902597835_1_phnJ_0001_21
GCA_902571005_1_phnJ_0001_30
GCA_902619105_1_phnJ_0001_30
GCA_902577765_1_phnJ_0001_30
GCA_902526005_1_phnJ_0001_1
GCA_902517885_1_phnJ_0001_30
GCA_902544525_1_phnJ_0001_23
IV (Not SAR11)
0.03

958

959 **Supplemental Figure 6.** Phylogeny of the *phnJ* gene for all *Pelagibacterales* genomes in the
960 pruned phylogeny data set.

Supplemental Table Legends

All supplemental tables are available on FigShare at:

https://doi.org/10.6084/m9.figshare.28087490.v1.

Supplemental Table 1. Summary of the 16S rRNA gene amplicon data from HTC17 and HTC18.

Supplemental Table 2. Detailed information for the isolate genomes reported in this study. The genome summary information originated from checkM (v1.1.2). *Indicates the genome has been manually verified to be completely closed.

Supplemental Table 3. Summary statistics for all genomes used for analyses in this study. This includes isolates reported here, previously published isolate genomes, and high-quality single amplified genomes (SAGs) used in the extended SAR11 phylogeny. The genomes used in read-recruitment are indicated. *Indicates the accession is the IMG Genome ID not the NCBI Accession.

Supplemental Table 4. Gene sets evaluated for use in SAR11 phylogenetics. The sets evaluated include the bac120 (Parks et al., 2018) and a subset of 165 of the genes (SAR11_165) delineated for the Alphaproteobacteria (Wang and Wu 2013; Muñoz-Gómez, 2019).

Supplemental Table 5. Summary of average genome statistics for the 23 genera established in the *Pelagibacteraceae* as well as the Ic, II, and III families.

Supplemental Table 6. Studies from which metagenomes were sourced.

Supplemental Table 7. List of all metagenomes used for read recruitment and accession numbers.

Supplemental Table 8. Detection values across genomes from all metagenomes used in read recruitment.

Supplemental Table 9. Average detection across genome cluster for bins in Fig 3.

Supplemental Table 10. Coverage values across genomes from all metagenomes used in read recruitment.

Supplemental Table 11. Type genomes and classification hierarchy for the *Pelagibacterales* including proposed naming schemes.

Supplemental Table 12. All of the final 95% ANI clusters defined including the cluster number, the final representative genome for that cluster and the list of other genomes in the same cluster.