UCSF UC San Francisco Previously Published Works

Title

Global ecotypes in the ubiquitous marine clade SAR86

Permalink

https://escholarship.org/uc/item/3406n1cb

Journal

The ISME Journal: Multidisciplinary Journal of Microbial Ecology, 14(1)

ISSN 1751-7362

Authors

Hoarfrost, Adrienne Nayfach, Stephen Ladau, Joshua <u>et al.</u>

Publication Date

2020

DOI

10.1038/s41396-019-0516-7

Peer reviewed

1	Global ecotypes in the ubiquitous marine clade SAR86
2	
3	Adrienne Hoarfrost ^{1,8*} , Stephen Nayfach ^{2,3} , Joshua Ladau ² , Shibu Yooseph ⁴ , Carol Arnosti ¹ ,
4	Chris L Dupont ⁵ , Katherine S. Pollard ^{3,6,7}
5	
6	¹ Dept. of Marine Sciences, University of North Carolina, Chapel Hill, NC
7	² Joint Genome Institute, Walnut Creek, CA
8	³ Gladstone Institutes, San Francisco, CA
9	⁴ Department of Computer Science, University of Central Florida, Orlando, FL
10	⁵ J. Craig Venter Institute, La Jolla, CA
11	⁶ Dept. of Epidemiology & Biostatistics, Institute for Human Genetics, Computational Health
12	Sciences Institute, and Quantitative Biology Institute, University of California, San Francisco,
13	CA
14	⁷ Chan-Zuckerberg Biohub, San Francisco, CA
15	⁸ Present address: Department of Biochemistry & Microbiology, Rutgers University, New
16	Brunswick, NJ
17	
18	* Corresponding author: adrienne.hoarfrost@rutgers.edu
19	
20	

21 Abstract

22

23 SAR86 is an abundant and ubiquitous heterotroph in the surface ocean that plays a central 24 role in the function of marine ecosystems. We hypothesized that despite its ubiquity, different 25 SAR86 subgroups may be endemic to specific ocean regions and functionally specialized for 26 unique marine environments. However, the global biogeographical distributions of SAR86 27 genes, and the manner in which these distributions correlate with marine environments, have not 28 been investigated. We quantified SAR86 gene content across globally-distributed metagenomic 29 samples and modeled these gene distributions as a function of 51 environmental variables. We 30 identified five distinct clusters of genes within the SAR86 pangenome, each with a unique 31 geographic distribution associated with specific environmental characteristics. Gene clusters are 32 characterized by strong taxonomic enrichment of distinct SAR86 genomes and partial 33 assemblies, as well as differential enrichment of certain functional groups, suggesting differing 34 functional and ecological roles of SAR86 ecotypes. We then leveraged our models and high-35 resolution, remote sensing-derived environmental data to predict the distributions of SAR86 gene 36 clusters across the world's oceans, creating global maps of SAR86 ecotype distributions. Our 37 results reveal that SAR86 exhibits previously unknown, complex biogeography, and provide a 38 framework for exploring geographic distributions of genetic diversity from other microbial 39 clades.

40

41 Introduction

42 Marine microbes are important drivers of biogeochemical cycling and ecological function 43 [1, 2]. Many studies have demonstrated the link between microbial genetic diversity and 44 functional capacities [e.g. 3–7], as well as the dependence of microbial community structure and 45 function on environmental variables [5, 8, 9]. However, the complexity of microbial 46 communities and of their interactions with their environment limit our ability to link microbial 47 genetic and functional variation across environments [10]. Furthermore, we have only limited 48 understanding of the geographic distributions of genetic diversity within key taxa, the 49 relationship of gene distributions to environmental conditions, and the manner in which these 50 distributions may result in distinct ecotypes across different environments and regions. Our 51 limitations in mapping microbial genetic diversity to geographic distributions restrict our ability 52 to predict microbial ecotypes across the environment. Accurate models linking environmental 53 and microbial variables may improve our current ability to incorporate biological inputs into 54 ecosystem models, which often rely on simplified biological systems utilizing incomplete 55 environmental relationships or imprecise evaluations of the functional capabilities of microbial 56 communities at different locations [11, 12]. 57 In microbial ecology, an ecotype [13] is often identified in practice as a group of closely

related lineages that co-occur on the same spatial or temporal scale and are associated with particular environmental conditions. This contrasts with the classical ecological definition, which additionally specifies that an ecotype must be genotypically adapted to the environmental conditions it is associated with [14]. In microbial ecology, where community members often lack cultured representatives and experiments directly measuring adaptive capacity to manipulated environmental conditions are challenging to conduct, adaptation is often difficult to demonstrate

64 conclusively. In this study, we define an ecotype to be a group of lineages within a clade whose 65 genomes contain a similar set of genes with a common geographic distribution associated with 66 distinct environmental conditions. This definition is consistent with previous studies of microbial 67 ecotypes [15]. Additionally, we require an ecotype to be taxonomically and functionally 68 differentiated from other ecotypes, which may indicate an adaptive strategy specific to that 69 ecotype, although we do not explicitly test for genetic signatures of adaptation. 70 The biogeography of marine microbes has been observed at scales from single depth 71 profiles [4] to global surveys [16, 17], revealing spatial and temporal patterns in microbial 72 community structure [16, 18], function [8, 19], and diversity [17]. Many marine microbial clades 73 exhibit population structure that correlates with their differential geographic distributions [20]. 74 Because most microbes have large pangenomes and flexible gene content [20], there is 75 significant interest in elucidating the differential functional capabilities of microbial ecotypes 76 and mapping their biogeographical distributions. Associating geographic distributions of 77 microbial ecotypes with environmental conditions could illuminate the links between microbial 78 community structure, function, and ecosystem processes, enabling predictions of biological and 79 chemical shifts in the world's oceans as environmental conditions change. However, there have 80 been very few efforts to predict biogeographic patterns of genetic and functional diversity of key 81 microbial taxa at large spatial scales in the ocean [17, 21]. 82 SAR86 is a ubiquitous marine heterotroph frequently found in surface waters, classified

by their 16S rRNA gene similarity as a clade within the Gammaproteobacteria [22–24]. SAR86 is a very diverse group with at least three subclades [23, 24]. Despite its ubiquity in marine systems, SAR86 eludes cultivation, and therefore knowledge of the ecological role of SAR86 in marine microbial communities is limited to evidence from genomes curated from single-cell

sequencing or metagenomic assembly [25–27]. These genomes suggest that SAR86 gene sets,
and hence functional capabilities, vary greatly across locations, even though the clade is very
commonly detected in marine environments. However, little is known about the manner in which
the distribution of subspecies and the vast genetic diversity within the SAR86 pangenome may
vary across large spatial extents, and what environmental factors may affect the geographic
distributions of different SAR86 gene families.

93 In this study, we build a custom pangenome of SAR86 genes from metagenomic co-94 assemblies and five available reference genomes. We then quantify the presence of each gene in 95 the pangenome across diverse marine epipelagic waters using hundreds of publicly available, 96 globally-distributed shotgun metagenomes. We find that geographic distributions of SAR86 97 genes are strongly associated with environmental variables, and we leverage these associations to 98 build machine learning models that accurately predict the presence of SAR86 genes from 99 environmental data. Using global-scale environmental measurements from satellite and 100 shipboard sources, we use our models to predict the global distribution of each geographically 101 variable gene in the SAR86 pangenome at a 9km² resolution. Our machine learning approach 102 enables patterns in the environmental variables that best predict the distributions of SAR86 genes 103 to emerge from the global metagenomic dataset without explicitly assuming a priori 104 relationships between inputs and outputs. Analysis of the resultant models reveals five clusters of 105 genes with unique environmental and geographic distributions, defining five ecotypes within the 106 SAR86 clade. We conclude that patterns of taxonomic and functional enrichment across these 107 ecotypes reveal previously underappreciated complexity in the geographic distributions 108 underlying the pangenome of this otherwise ubiquitous marine heterotroph, with great potential 109 to illuminate structure-function relationships across the marine environment.

110 Materials & Methods

111

112 Creation of the SAR86 pangenome and global SAR86 gene presence/absence dataset

- 113 A custom pangenome of 51 711 nonredundant SAR86 genes was created with the
- 114 MIDAS tool [20], from a combination of genomic sources [23, 24, 25] as well as a massive co-
- assembly of metagenomic sequences (Supplemental Text 1.1-1.2).
- 116 A global dataset of SAR86 gene presence/absence for each gene in the SAR86
- 117 pangenome was then created. Shotgun metagenomic sequencing reads from the TARA project
- 118 [9] were mapped to the SAR86 pangenome, and the resulting normalized read coverage for each
- gene was used to determine SAR86 gene presence or absence for all SAR86 genes at 198 TARA
- 120 sites (Supplemental Text 1.3).
- 121

122 Environmental data curation and processing

In order to build models predicting SAR86 gene presence from environmental variables, environmental data available at resolution between 9km to 1-degree and at global scale were curated from a combination of contemporary satellite data and historical averages of satellite and interpolated in situ measurements. A total of 51 environmental features were compiled (SI Table 1, Supplemental Text 1.4). Normalized environmental feature values closest to each TARA site's latitude, longitude, and, where relevant, sampling depth and/or sampling date (SI Table 2) served as the input feature vectors for each TARA site during model training.

131 Gene presence/absence models & predictions

132 Classification models predicting SAR86 gene presence or absence as a function of the 133 environmental feature vectors across TARA sites were built for each of 24 317 geographically 134 variable SAR86 genes, using logistic regression with L1 regularization (Supplemental Text 1.5). 135 Geographically variable genes were defined as genes present at between 20-80% of TARA sites. 136 155 TARA sites for which SAR86 was present and environmental data was available were split 137 into training, validation, and test sets of 111, 13, and 31 sites respectively. The final models 138 trained independently for each of the 24 317 geographically variable genes can be reproduced 139 with code available on the associated Github repository [29].

140

141 Clustering, global maps of ecotypes, & enrichment analysis

142 To identify groups of SAR86 genes whose geographic distributions are best predicted by 143 similar environmental variables, we clustered genes into 5 clusters on the logistic regression 144 model coefficients for each environmental feature using a k-means algorithm (Supplemental 145 Text 1.6). Clustering on environmental features associated with gene models enabled us to 146 identify the environmental variables underlying geographic distributions of genes, and also 147 enabled the projection of predicted cluster distributions at global scales. To produce global 148 projections (i.e., maps) of each SAR86 gene cluster, we predicted the presence or absence of 149 each cluster at 9km2 resolution and global scale from the available satellite and historical 150 environmental data ([29], Supplemental Text 1.6). A Jupyter notebook and a python script for 151 reproducing clusters and cluster projections are available ([29]). 152 The distribution and enrichment across clusters were evaluated at the genome, contig, and

153 functional level for two SAR86 reference genomes SAR86A and SAR86E, for the contigs of the

154	SAR86 co-assembly, and for the functional annotations to Pfam [30] for the SAR86 pangenome			
155	(Supplemental Text 1.7). This produced a vector of taxonomic/functional enrichment values			
156	associated with each contig/annotation for each cluster, with which the statistical significance of			
157	cluster enrichment could be tested (Supplemental Text 1.7).			
158				
159	Results			
160	This study first modeled the relationships between SAR86 gene distributions and			
161	environmental variables. We used a regularized logistic regression approach to identify the			
162	subset of environmental variables that are most important for predicting the geographical			
163	distributions of each gene and to estimate the strength of these gene-environmental variable			
164	relationships. Using unsupervised clustering of these association profiles, we then identified			
165	clusters of genes with similar environmental distributions. Clustering enabled us to identify the			
166	structure underlying the environmental gene distributions without explicit prior knowledge of			
167	expected SAR86 ecotypes. By using environmental variables available at global scale, we			
168	leveraged our gene models to predict the geographic distribution of these emergent ecotypes in			
169	regions far beyond the sampling locations specific to the TARA study.			
170				
171	Accurate prediction of SAR86 gene distributions from environmental variables			
172	SAR86 gene content in TARA Oceans metagenomes is associated with environmental			
173	characteristics of the sampling locations. We built a regularized logistic regression model for			
174	each gene that accurately predicts the probability of the gene being present at a given location as			
175	a function of the most predictive subset of environmental variables (Methods, Supplemental Text			

176 1.5).

177	The resulting 24 317 gene models predict SAR86 gene presence/absence with an average
178	of 79.4% accuracy in the test set, and a median test accuracy of 80.6%. Precision and recall
179	measures are roughly even (0.85 and 0.81, respectively; SI Fig 3a), with an F1 score of 0.83. For
180	21 264 out of 24 317 genes (87.4%), the models have accuracies in the test set that are an
181	improvement over the majority class accuracy – the accuracy of the model if it predicts 'always
182	absent' or 'always present', whichever is in the majority (SI Fig 3b).
183	As an additional test of the robustness of the models, the accuracy of predictions at those
184	TARA sites that were not included in model development, where SAR86 was not present or were
185	in very low abundance, was also examined. There were 20 such sites for which environmental
186	data was available for all features. These 20 sites were primarily mesopelagic samples,
187	distributed across all ocean basins (Supplemental Text 1.5). Across these 20 sites, the average
188	accuracy of the gene models is 68.5%, while the median accuracy is 70.0%. While this
189	performance is below that achieved at sites where SAR86 was present, it suggests that our
190	models are able to make fairly accurate predictions even when extrapolating outside of the
191	distribution of gene presence used in training.
192	An average of 17 of 51 environmental features is significantly associated with each
193	gene's distribution across TARA Oceans sites. Across multiple gene models, the same
194	environmental feature was frequently selected during model training (SI Fig 4). These frequently
195	associated variables include latitude, longitude, distance from land, ocean depth, and other
196	features that might describe the general ocean basin or region of a sample; as well as pH, sea
197	surface temperature, pycnocline depth, nitrogen:phosphorous ratio, cloud fraction, and other
198	environmental factors that describe regions of the ocean that experience particular environmental
199	conditions.

200	While the environmental features that best predict gene presence/absence vary by the
201	individual gene model, and many of the 51 environmental variables covary with one another,
202	training logistic regression multiple times on the same data with different random seeds resulted
203	in the same sets of environmental features being chosen as the most predictive for each gene
204	model (see Jupyter notebook in [29]). This consistency suggests that the environmental features
205	selected in each model reflect a true difference in predictive power between the selected features
206	and those that were not selected, rather than a random choice among features that are roughly
207	equally predictive.

208

Clustering of SAR86 genes into common environmental distributions & global projections of their biogeographic distributions

The environmental features that best predict individual genes, and the strength of the coefficients associated with any particular environmental feature, vary by the individual gene model. However, there are apparent patterns among genes, with some groups of genes appearing to be predicted by similar environmental variables, as well as similar magnitudes and signs of the coefficients associated with those variables. These patterns suggest that genes that are predicted by similar environmental features occupy similar geographic distributions characterized by unique environmental conditions.

K-means clustering of genes by their logistic regression environmental feature
coefficients identified five clusters within the SAR86 pangenome characterized by similar
environmental distributions (Fig 1). The average environmental feature coefficient across all
genes in each cluster (the "centroid") demonstrates the distinct pattern of association with
environmental features of each cluster (SI Table 3).

223	Each TARA site contains genes from a mixture of clusters, but the dominant clusters and
224	the evenness of the proportion of each cluster is variable across sites (Fig 2, SI Fig 5, SI Table 4).
225	For example, cluster 2 is strongly associated with longitudes in the western hemisphere, and this
226	is also reflected across TARA samples, for which cluster 2 is present in highest proportions for
227	those TARA sites sampled in the Pacific Ocean (Fig 2, SI Fig 5b). In contrast, cluster 3 genes are
228	found in higher proportions at TARA sites sampled in the eastern hemisphere, reflecting their
229	predicted geographic distributions (Fig 2, SI Fig 5c).
230	A Shannon diversity metric was used to measure the relative evenness and proportion of
231	the five clusters at each TARA site (SI Table 4, Supplemental Text 1.7). The TARA sites with
232	the lowest Shannon diversity include TARA station 93 at 34°S and 73°W off the coast of Chile,
233	which is dominated by cluster 5 genes, and TARA stations 38, 42, 45, and 36 in the Indian
234	Ocean, which are dominated by cluster 4 genes. The TARA sites with the highest Shannon
235	diversity include many of the mesopelagic depth samples in the Pacific Ocean, as well as station
236	70 in the South Atlantic basin at 20.4°S and 3.2°W.
237	We next used the cluster centroids and global-scale environmental data to predict the geographic
238	distribution of each cluster beyond the TARA sampling locations (Fig 3). These global
239	projections reveal the differential distributions of SAR86 gene clusters. These differential
240	distributions are reflected in variation across longitude (e.g. cluster 2 versus clusters 3 and 4),
241	latitude (e.g. clusters 1 and 5 versus clusters 2, 3, and 4), and season (e.g. cluster 1, Fig 3). In
242	each case, the highest magnitude coefficients for each cluster are suggestive of their predicted
243	geographic distributions (SI Table 3, Supplemental Text 2.1).
244	

245 Taxonomic enrichment & functional differentiation across clusters define SAR86 ecotypes

246 The cluster assignments of genes from the SAR86 reference genomes SAR86A and 247 SAR86E show clear partitioning on taxonomic lines. Genes from each genome are assigned 248 primarily to two clusters, and each cluster is dominated by one genome. SAR86A genes are 249 partitioned primarily into clusters 4 and 3, with 493 and 118 out of the 622 SAR86A genes 250 assigned to cluster 4 and 3 respectively, while only 4 and 7 genes were assigned to clusters 2 and 251 5, and 0 genes to cluster 1. The 157 SAR86E genes were partitioned into clusters 1 and 5, with 252 76 and 78 genes respectively, while only 2 and 1 genes were assigned to clusters 2 and 4, 253 respectively, and 0 genes to cluster 3. 254 Clusters also show clear taxonomic differentiation at the contig level. Those genes that do 255 not originate from one of the five SAR86 genomes constitute a total length of 22 Mbp 256 originating from 732 contigs from the SAR86 co-assembly. All clusters are significantly 257 enriched in specific contigs (p<0.001, Fig 4c), with a unique set of contigs enriched on each 258 cluster. Genes from the same contig are generally assigned to the same cluster, such that gene 259 assignments of almost all contigs, 540 out of 732 contigs, are enriched on only one cluster, 183 260 contigs are enriched on only two clusters, and the remaining 9 contigs are enriched on 3 clusters 261 (Fig 4). Where a contig is enriched, the enrichment is strong, with an average enrichment of 3.03 262 and a standard deviation of 0.43, and ranging from 1.41 in cluster 4 to 5.25 in cluster 2. 263 The taxonomic partitioning of clusters is also evident in their distribution across TARA 264 sites. First, the cluster proportions and the relative abundances of SAR86 genomes at TARA sites 265 reflect the taxonomic differentiation of genomes across clusters. The clusters associated with 266 SAR86A (clusters 3 and 4) are in higher proportions relative to the clusters associated with 267 SAR86E (clusters 1 and 5) at TARA sites where SAR86A abundances are higher relative to

268 SAR86E (SI Fig 6, Pearson R2 = 0.70, $P=1.56 \times 10^{-26}$). In addition to this genomic evidence, the 269 normalized read coverage across TARA sites for genes from the same cluster are more highly 270 correlated with one another than genes from different clusters (SI Fig 7), as would be expected if 271 genes belonging to the same cluster share a common taxonomic origin. This indicates that genes 272 from the same genome are assigned to the same cluster, although a single cluster may be made 273 up of genes from multiple genomes. Indeed, the 22Mbp of genomic material in the SAR86 co-274 assembly is enough for at least 11 genomes of size similar to that of known SAR86 reference 275 genomes, so multiple genomes are expected to be contained within the 5 identified clusters. 276 These clusters are thus composed of genes that co-occur with one another across similar 277 environmental contexts, and are taxonomically differentiated, but do not necessarily represent 278 individual SAR86 genomes.

279 In addition to taxonomic enrichment across clusters, there is also significant partitioning 280 of genes at the functional level, with differential enrichment of Pfam annotated genes across 281 clusters (Fig 5). Pfams are enriched by an average value of 0.25 and a standard deviation of 0.10, 282 ranging from 0.13 in cluster 4 to 0.32 in cluster 2. This enrichment is significant (p<0.01) for 283 most of the clusters (Fig 5c). This result suggests that clusters 1, 2, and 4 have significant 284 functional enrichment, while functional enrichment on cluster 3 is marginally significant. Genes 285 from a particular Pfam are most often assigned to only two or three clusters (Fig 5b). While 286 functional enrichment in general is less strong than taxonomic enrichment, this may be due to the 287 relative coarseness of functional annotation compared to taxonomic assignments, and our 288 inability to annotate many genes with confidence.

Enrichment of specific Pfams corresponding to some ecologically important functions
indicate possible differentiation in ecological function between clusters. For example, glycosyl

291	hydrolase family 3 (Pfams PF00933, PF01915), which corresponds to exo-acting glucosidases, is
292	enriched across clusters 3, 4, and 5, and depleted in clusters 1 and 2, while glycosyl hydrolase
293	family 16 (Pfam PF00722), which corresponds to endo-acting glucanases, is enriched strongly on
294	cluster 3, depleted in clusters 1 and 2, and near the null value for clusters 4 and 5 (SI Fig 8).
295	Proteorhodopsin, a photoactive transmembrane proton pump first identified in bacteria in SAR86
296	[31] and used by SAR86 for photoheterotrophic ATP generation, is enriched in clusters 3 and 4,
297	and depleted in clusters 1, 2, and 5 (SI Fig 9).
298	
299	Discussion
300	While SAR86 is generally considered to be a ubiquitous heterotroph in the ocean, this
301	study demonstrates that SAR86 harbors immense within-species genetic diversity that is strongly
302	associated with environmental variables. These distinct environmental distributions of gene
303	clusters define a deeper geographic variability within the SAR86 clade than previously
304	appreciated. The three near-complete and two partial genomes available for SAR86 [25, 26]
305	show high diversity within this clade; average nucleotide identity between genomes is between

306 70-80% (SI Table 5). In light of this high diversity, it is perhaps not surprising that the

307 geographically variable genes in the SAR86 pangenome can be decomposed into five distinct

308 clusters with different geographic distributions associated with unique environmental variables.

309 These clusters are differentiated at the taxonomic and functional level, which has implications

310 for our understanding of the biogeography of SAR86, as well as its ecological role within

311 microbial communities in the marine environment.

Using a data intensive approach to build machine learning models of the relationship
between SAR86 genes and environmental variables at a global scale, we demonstrate how such

an approach can be used to better understand the factors shaping the biogeography of microbial
clades. This approach can reveal patterns that would likely be missed at the 16S OTU or
community level, or using data from a smaller scale. Particularly as metagenomics data become
increasingly available in the future, such an approach holds promise for illuminating the
relationship between microbial community structure and ecological function across broad
taxonomic and spatial scales.

320 The results of this study identify clusters of genes that, while their phylogenetic 321 relatedness is unknown, are taxonomically and functionally differentiated and occupy distinct 322 environmental distributions. While the functional traits that confer niche restriction within these 323 distributions is not obvious from our results, functional differentiation across clusters of glycosyl 324 hydrolases (SI Fig 8) – an important class of enzymes for heterotrophic metabolism of 325 polysaccharides – and proteorhodopsin (SI Fig 9) – a light-driven means of energy generation 326 and enhanced nutrient and organic carbon uptake - suggest that genes associated with different 327 clusters define distinct functional roles filled by each cluster. Glycosyl hydrolase families 3 and 328 16 target many of the same substrates – β -linked glucans, including the abundant marine 329 plankton storage glucan laminarin – but using different enzymatic mechanisms [32]. The strong 330 enrichment in cluster 3, and strong depletion in clusters 1 and 2, of both families, compared to 331 the enrichment of only family 16 in clusters 4 and 5, may indicate distinct ecological functions of 332 SAR86 across clusters that utilize differing metabolic strategies and have disparate impacts on 333 carbon remineralization. Proteorhodopsin genes are only enriched in clusters 3 and 4, the two 334 clusters associated with lower latitudes and more abundant sunlight, and are depleted in clusters 335 1 and 5, which are associated with temperate latitudes. This latitudinal pattern may also indicate 336 distinct energy generation and metabolic strategies that correspond with the environmental

distributions of the clusters. Given the clear taxonomic and functional partitioning of the SAR86
pangenome across clusters with distinct geographic distributions associated with unique
environmental conditions, we conclude that the clusters described here define previously
unidentified ecotypes within the SAR86 clade.

341 The geographic distributions of SAR86 ecotypes are consistent with previous studies. An 342 investigation of temporal and geographic patterns in SAR86 noted that while the phylogenetic 343 substructure of the SAR86 clade implies that it may be made up of multiple ecotypes, these 344 could not be identified at the limited geographic resolution of the study [24]. The potential 345 existence of SAR86 ecotypes was also noted in the apparent geographic distributions of 346 SAR86A, B, C, and D genomes [25], which differed in their distributions across coastal versus 347 open ocean sampling sites and along temperature gradients. This general observation is 348 supported by the predicted distributions of the clusters identified in our study, for which three 349 clusters (clusters 2, 3, and 4) are partially defined by their warmer, open ocean distributions, and 350 two (clusters 1 and 5) are associated with cooler temperatures. The difficulty of identifying 351 ecotypes in SAR86 contrasts with SAR11, for which distinct ecotypes have been identified 352 within a constrained geographic sample because they were strongly associated with differences 353 in depth and salinity distributions [15]. This study was able to identify SAR86 ecotypes, despite 354 their partially sympatric distributions that cause single sampling sites to be composed of genes 355 from multiple clusters, because of the larger data size and geographic distribution of the TARA 356 dataset, and our unique approach to defining ecotypes based on quantitative models of 357 environmental associations with geographically variable genes. Whereas ecotypes are typically 358 identified by building a phylogeny based on core genes and observing whether environmental 359 variables map over the phylogeny [e.g. 23, 33], our approach is quantitative, objective and

independent of a priori knowledge of phylogeny, and results in sets of genes and functionalfeatures that define the ecotype.

362	The taxonomic and functional differentiation of genes across SAR86 ecotype clusters is
363	significant in the context of interactions between microbial community structure, function, and
364	ecology. Both community composition [16–18, 34] and functional traits [3, 4, 8, 19] vary
365	geographically and can be predicted to some extent by environmental variables [8, 17].
366	Taxonomic variation can lead to functional differentiation of microbial communities [4, 35, 36],
367	which ultimately shapes biogeochemical cycling and ecosystem function; conversely, functional
368	redundancy across microbial taxa can complicate the relationship between structure and function
369	[37], with taxonomically variable communities playing similar functional roles [38].
370	Disentangling the relationship between environment, biogeography, structure, and function is
371	therefore a significant ongoing challenge in microbial ecology [5, 7, 8, 10]. By focusing on
372	patterns at the individual gene level within a single clade, we are able to uncover patterns in
373	environmental distributions of genetic diversity at a scale that would normally be obscured by
374	the complexity inherent to microbial communities. For example, previous studies have found that
375	functional classifications of taxa are better predicted by environmental parameters than
376	taxonomic 16S-based classifications [8]; however, these functional classifications are broad – all
377	of the SAR86 pangenome would be classified as 'aerobic chemoheterotroph' – in order to
378	control for the vast genetic diversity of traits in mixed microbial communities. It is likely that
379	within the SAR86 pangenome there is ecological differentiation within this category that, for
380	example, could lead closely related phylotypes of SAR86 that belong to different ecotypes to
381	utilize different substrates [33, 39, 40]. This hypothesis is supported by the functional enrichment
382	across our clusters and the differential enrichment of carbohydrate utilizing enzymes (SI Fig 8).

Previous analyses of the genomic context of SAR86 genomes also suggest that much of the diversity among SAR86 genomes may be driven by fine scale diversification of catabolic enzymes on loci associated with TonB dependent receptors [25], which are responsible for transporting carbon compounds (as well as metals) into the cell [41].

387 The accuracies of our gene models are better on average than previous studies (0.79 vs)388 0.48, [8]), which may similarly be due in part to our focus on modeling individual genes rather 389 than whole communities. This difference in model accuracy may also be due to our consideration 390 of different, and a larger number, of input environmental features. Here, the environmental 391 features were chosen for their availability at global resolution rather than their human-predicted 392 importance in regulating microbial function. These environmental features may be more 393 predictive of the distributions of SAR86 genes, even if they are less relevant to biological 394 function. The environmental factors that influence whether an organism grows in a particular 395 location or community may be different from those that drive their function within that 396 community: for example, an organism may only grow in fresh or saline waters, while the 397 maintenance of a nitrogen fixation pathway depends on nutrients or other factors. It is important 398 to note that those environmental features that are selected as most predictive for each gene model 399 do not necessarily drive the growth of SAR86 in a causal manner, but implies only that these 400 environmental features are good predictive proxies for the presence of that gene. The 401 interpretation of the most predictive environmental features may vary depending on the feature; 402 some features may be a proxy for biological phenomena, while others simply define 403 oceanographic regions, or are proxies for other factors that cannot be measured that are true 404 causal drivers of variation. The features chosen by the L1 regularization procedure are also likely 405 biased by the scope of the samples used as inputs to the model. For example, the cluster

406 associated with western hemisphere longitudes is overrepresented in sites from the Pacific Ocean 407 in the TARA expedition dataset. However, there are longitudes both east and west of the 408 antemeridian in the Pacific, represented as negative and positive longitudes in the models, and it 409 is a limitation of the TARA dataset that only samples from the eastern part of the basin, in the 410 western hemisphere, are represented. This limitation results in an unnaturally sharp transition in 411 cluster projections on the antemeridian in the Pacific Ocean for those clusters for which 412 longitude is a strong predictor. This observation also serves as a note of caution for the 413 interpretation of the global projections, whose predicted distributions will likely break down 414 most in locations for which representation of samples is most sparse, e.g. in polar regions. We are able to make accurate predictions of geographic distributions of SAR86 genes, 415 416 identifying previously unknown biogeographical complexity within an otherwise ubiquitous 417 heterotrophic clade and making global projections of the distributions of SAR86 ecotypes 418 associated with distinct environmental distributions. Our modeling approach leverages a large 419 dataset across broad geographic regions, demonstrating the potential of machine learning and the 420 use of broader scale integrated datasets for marine microbial ecology. The five global ecotypes 421 underlying the highly diverse SAR86 clade, the taxonomic and functional differentiation across 422 ecotypes, and the distinct environmental distributions of SAR86 genetic diversity highlight the 423 importance of SAR86 within marine microbial communities and broadens the context for 424 interpreting their ecological impact across the world's oceans. 425

426 Acknowledgements

427 This work was supported by a Deep Carbon Observatory Deep Life Modeling &
428 Visualization Fellowship to AH; OCE-1736772 to CA; Gordon and Betty Moore Foundation

- 429 grant #3300 to KSP; and grants from the Beyster Family Fund of the San Diego Foundation and
- 430 Life Technologies Foundation to JCVI.
- 431

432 **Conflict of Interest**

- 433 The authors declare no conflict of interest.
- 434

435 Author Contributions

- 436 CD and SY created the SAR86 co-assembly of SAR86 genes from the Global Ocean
- 437 Sampling sequences, and CD annotated the SAR86 pangenome. SN created the pangenome and

438 mapped TARA samples to the SAR86 pangenome. AH gathered satellite environmental data,

- 439 created the models, did clustering, identified ecotypes and analyzed data. JL gathered historical
- 440 environmental data. All authors contributed to discussion of data and writing of the manuscript.
- 441

442 **References**

- Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive Earth's biogeochemical cycles. *Science* 2008; **320**: 1034–9.
 Azam F. Microbial control of oceanic carbon flux: The plot thickens. *Science* (80-) 1998;
 280: 694–696.
- Shi Y, Tyson GW, Eppley JM, Delong EF. Integrated metatranscriptomic and
 metagenomic analyses of stratified microbial assemblages in the open ocean. *ISME J*2011; 5: 999–1013.
- 450 4. Delong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N, et al. Community
 451 Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. *Science (80-)*452 2006; **311**: 496–503.
- 453 5. Raes J, Letunic I, Yamada T, Jensen LJ, Bork P. Toward molecular trait-based ecology
 454 through integration of biogeochemical, geographical and metagenomic data. *Mol Syst Biol*455 2014; 7.
- Guidi L, Chaffron S, Bittner L, Eveillard D, Larhlimi A, Roux S, et al. Plankton networks
 driving carbon export in the oligotrophic ocean. *Nature* 2015; **532**: in review.
- 458 7. Morales SE, Holben WE. Linking bacterial identities and ecosystem processes: Can
 459 'omic' analyses be more than the sum of their parts? *FEMS Microbiol Ecol* 2011; **75**: 2–
 460 16.
- 8. Louca S, Parfrey LW, Doebeli M. Decoupling function and taxonomy in the global ocean

462		microbiome. Science (80-) 2016; 353: 1272–1277.
463	9.	Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure
464		and function of the global ocean microbiome. <i>Science (80-)</i> 2015; 348 : 1–10.
465	10.	Widder S. Allen RJ. Pfeiffer T. Curtis TP. Wiuf C. Sloan WT. et al. Challenges in
466	-	microbial ecology: building predictive understanding of community function and
467		dynamics. ISME J 2016.
468	11	Treseder KK Balser TC Bradford MA Brodie EL Dubinsky EA Eviner VT et al
469		Integrating microbial ecology into ecosystem models: Challenges and priorities
470		Riogeochemistry 2012: 109: 7–18
471	12	Wieder WR Allison SD Davidson FA Georgiou K Hararuk O He Y et al Explicitly
472	12.	representing soil microbial processes in Earth system models <i>Global Riogeochem Cycles</i>
473		$2015 \cdot 29 \cdot 1782 = 1800$
173 474	13	Cohan FM. Towards a concentual and operational union of bacterial systematics ecology
474 1.75	15.	and evolution <i>Philos Trans R Soc R Riol Sci</i> 2006: 361 : 1085–1006
476	14	Begon M. Townsend C. Harner I. Ecology: From individuals to access tems. 4th ed. 2006.
470 1.77	17.	Blackwell Publishing
478	15	Carlson CA Morris P. Parsons P. Trausch AH Giovannoni SI Vergin K. Seasonal
470	15.	dynamics of SAD11 nonulations in the suphotic and mesonalogic zones of the
479		northwastern Sargassa Saa, ISME 12000: 3: 282, 205
400	16	Mortiny IDH Dohannan DIM Brown IH Colwell DK Fuhrman IA Groon II at al
401	10.	Martiny JDH, Bonannan BJW, BIOWH JH, COlwell KK, Fullman JA, Gleen JL, et al. Microbiol biogeography: putting microorganisms on the man. Nat Pay Microbiol 2006; 4:
402		Microbial biogeography: putting microorganisms on the map. <i>Nat Kev Microbiol</i> 2000; 4 :
483	17	102-12. Leden I. Sherriten TI. Einneene MM. Jeanin C. Kernhel SW. O'Dunnen I. et al. Clabel
484	1/.	Ladau J, Sharpton IJ, Finucane MM, Jospin G, Kembel SW, O Dwyer J, et al. Global
485	10	marine bacterial diversity peaks at high latitudes in winter. <i>ISME J</i> 2013; 7: 1669–77.
486	18.	Zinger L, Amaral-Zettler L a, Fuhrman JA, Horner-Devine MC, Huse S, Weich DBM, et
487		al. Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. <i>PLoS</i>
488	10	<i>One</i> 2011; 6 : e24570.
489	19.	Jiang X, Langille MGI, Neches RY, Elliot M, Levin S a., Eisen J a, et al. Functional
490		Biogeography of Ocean Microbes Revealed through Non-Negative Matrix Factorization.
491	•	PLoS One 2012; 7: 1–9.
492	20.	Naytach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics
493		pipeline for strain profiling reveals novel patterns of bacterial transmission and
494		biogeography. Genome Res 2016; 26: 1612–1625.
495	21.	Kent AG, Dupont CL, Yooseph S, Martiny AC. Global biogeography of Prochlorococcus
496		genome diversity in the surface ocean. <i>ISME J</i> 2016; 10 : 1856–1865.
497	22.	Britschgi TB, Giovannoni SJ. Phylogenetic analysis of a natural marine bacterioplankton
498		population by rRNA gene cloning and sequencing. <i>Appl Environ Microbiol</i> 1991; 57:
499		1707–1713.
500	23.	Suzuki MT, Beja O, Taylor LT, Delong EF. Phylogenetic analysis of ribosomal RNA
501		operons from uncultivated coastal marine bacterioplankton. <i>Environ Microbiol</i> 2001; 3:
502		323–331.
503	24.	Treusch AH, Vergin KL, Finlay LA, Donatz MG, Burton RM, Carlson CA, et al.
504		Seasonality and vertical structure of microbial communities in an ocean gyre. $ISMEJ$
505		2009; 3 : 1148–1163.
506	25.	Dupont CL, Rusch DB, Yooseph S, Lombardo MJ, Alexander Richter R, Valas R, et al.
507		Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. ISME

508		<i>L</i> 2012: 6 : 1186–1199
500	26	Rusch DB Lombardo M-L Yee-Greenbaum I Novotny M Brinkac I M Lasken RS et al.
510	20.	Draft genome sequence of a single cell of SAR86 clade subgroup IIIa Genome Announc
510		2013: 1: e00030-12
512	27	Swan BK Tupper B. Sczyrba A. Lauro FM. Martinez-Garcia M. González IM. et al.
512	21.	Bravalant ganama straamlining and latitudinal divergence of planktonic basteria in the
515		rievalent genome streamming and fattudinar divergence of planktome bacteria in the
514	20	Wettern A.D. Abraham D. Dalaw O. Dian TI. Duissell T. Calibard II. et al. DATDIC the
515	28.	wallam AK, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the
516		bacterial bioinformatics database and analysis resource. <i>Nucleic Acids Res</i> 2014; 42: 581–
51/	20	
518	29.	Hoartrost A. SAR86. Github repository.
519	30.	Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam
520		protein families database: Towards a more sustainable future. <i>Nucleic Acids Res</i> 2016; 44:
521		D279–D285.
522	31.	Béjà O, Aravind L, Koonin E V, Suzuki MT, Hadd A, Nguyen LP, et al. Bacterial
523		Rhodopsin : Evidence for a New Type of Phototrophy in the Sea Linked references are
524		available on JSTOR for this article : Bacterial Bacterial Rhodopsin : Rhodopsin : Evidence
525		for for a a New New Type Type of of Phototrophy Phototrophy in the Sea. 2017; 289:
526		1902–1906.
527	32.	Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The
528		carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 2014; 42: 490-
529		495.
530	33.	Aguilar D. Aviles FX. Ouerol E. Sternberg MJE. Analysis of phenetic trees based on
531		metabolic capabilites across the three domains of life. J Mol Biol 2004: 340 : 491–512.
532	34	Pommier T. Canbäck B. Riemann L. Boström KH. Simu K. Lundberg P. et al. Global
533	5	patterns of diversity and community structure in marine bacterioplankton. <i>Mol Ecol</i> 2007:
534		16 : 867–80
535	35	Galand PE Pereira O Hochart C Auguet IC Debroas D A strong link between marine
536	55.	microbial community composition and function challenges the idea of functional
530		redundancy ISME 12018. 1
538	36	Strickland MS Lauber C Fierer N Bradford MA Testing the functional significance of
530	50.	microbial community composition Ecology 2000: 00: 441 451
539	27	Louce S. Dolz MF. Mozol F. Albright MDN. Huber IA. O'Conner MI at al Eurotion and
540	57.	Louca S, Foiz Mir, Mazer F, Albright MDN, Huber JA, O Comior Mi, et al. Function and
541	20	Leave C. Leaves CMC. Diver ADE Least IC. Spinstere DC. Developed W. et al. U. a.
542	38.	Louca S, Jacques SINS, Pires APF, Leai JS, Srivastava DS, Parirey LW, et al. High
543		taxonomic variability despite stable functional structure across microbial communities.
544	20	Nat Ecol Evol 2016; 1: 0015.
545	39.	Martiny JBH, Jones SE, Lennon JT, Martiny AC. Microbiomes in light of traits: A
546		phylogenetic perspective. Science (80-) 2015; 350.
547	40.	Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. Resource partitioning and
548		sympatric differentiation among closely related bacterioplanktn. <i>Science (80-)</i> 2008; 320 :
549		1081–1085.
550	41.	Noinaj N, Guiller M, Barnard TJ, Buchanan SK. TonB-dependent transporters: regulation,
551		structure, and function. Annu Rev Microbiol 2010; 64: 43-60.
552		
553		

554 **Figure Legends**

555

556 Fig. 1 – Heatmap of model coefficients for each environmental feature (rows) and gene 557 (columns), ordered by cluster (x axis).

558

559 Fig. 2 – Relative proportion of clusters at each TARA site (vertical bars). TARA sites are sorted 560 by longitude (x axis; negative numbers correspond to longitude west of the prime meridian).

- 561 Blue, cluster 1; green, cluster 2; yellow, cluster 3; purple, cluster 4; pink, cluster 5.
- 562

563 **Fig. 3** – Global predictions of SAR86 gene cluster distributions for each cluster (rows) in 564 January, April, July, and October of 2009 (columns). Red indicates a high confidence of a gene 565 cluster being present, blue a high confidence of a gene cluster being absent, and white a low 566 confidence prediction.

567

568 Fig. 4 – Contig enrichment in clusters. (a) Heatmap of enrichment (red) or depletion (blue) of

569 each contig (columns) across each cluster (rows). (b) Pie chart of the number of clusters in which

570 SAR86 contigs are enriched. (c) Mean positive enrichment value, standard deviation of positive

571 enrichment values, and the Mann-Whitney P value for significance of cluster enrichment, for each cluster.

572 573

574 Fig. 5 – Functional enrichment in clusters. (a) Heatmap of enrichment (red) or depletion (blue) 575 of the 405 most abundant Pfam families (columns) across each cluster (rows). Pfams are ordered 576 left to right by the number of genes annotated to it, from the most abundant Pfams to the Pfams

577 with as few as 20 genes annotated to it. (b) Pie chart of the number of clusters in which Pfams

578 are enriched. (c) Mean positive enrichment value, standard deviation of positive enrichment

- 579 values, and the Mann-Whitney P value for significance of cluster enrichment, for each cluster.
- 580 581







cluster 2

cluster 3

cluster 4

cluster 5







(c)

	mean enrichment	std. dev. enrichment	MW P-value
cluster 1	2.67	1.22	1.13 x 10 ⁻⁷³
cluster 2	5.25	3.33	7.36 x 10 ⁻¹⁵⁹
cluster 3	3.97	2.00	2.23 x 10 ⁻¹³⁵
cluster 4	1.41	0.80	8.86 x 10 ⁻⁵⁷
cluster 5	2.16	0.96	5.72 x 10-4





(c)



(b)

	mean enrichment	std. dev. enrichment	MW P-value
cluster 1	0.22	0.17	5.61 x 10 ⁻³
cluster 2	0.33	0.24	1.63 x 10 ⁻⁰⁹
cluster 3	0.26	0.24	0.016
cluster 4	0.14	0.13	1.06 x 10 ⁻¹¹
cluster 5	0.17	0.13	0.108

bioRxiv preprint first posted online May. 10, 2019: doi: http://dx.doi.org/10.1101/635185. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.