Title

Functional Unknomics of the SAR11 clade using bioinformatics approaches

Authors

Satoshi Nishino[1,2], Kento Tominaga[3*], Kimiho Omae[4], Teppei Deguchi[5], Koji Hamasaki[1,2], Susumu Yoshizawa[2,3] and Yuki Nishimura[1*]

Affiliations

1) Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8562, Japan
2) Atmosphere and Ocean Research Institute, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8564, Japan
3) Department of Natural Environmental Studies, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8563, Japan
4) RIKEN Pioneering Research Institute, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan
5) Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa, Chiba 277-0882, Japan

Keywords

SAR11, Pelagibacterales, Unknome, Functional unknown genes, Antiviral defense system

*Correspondence should be addressed to K.T. (tominaga-kento@edu.k.u-tokyo.ac.jp) and Y.N. (ynishimura@edu.k.u-tokyo.ac.jp).

Abstract (249 words)

A substantial fraction of the genes in marine bacteria lack detectable sequence similarity to genes with known functions. These functionally uncharacterized genes—collectively referred to as the "unknome"—represent a largely unexplored genetic repertoire harboring insights into marine bacterial ecology. In this study, we explored the function of the unknome of SAR11 clade, the most abundant bacterial lineage in the ocean, with a particular focus on genes that provide insight into their ecology. Based on the COG classification, approximately 58% of SAR11 ortholog groups were classified as unknome. Among the SAR11 unknome, we successfully inferred the functions of 69 ortholog groups that are conserved in SAR11 clade by protein structure similarity searches and genomic context analyses. These ortholog groups include putative transporter components, supporting the current ecological understanding that SAR11 clade is specialized in substrate uptake to adapt to oligotrophic marine environments. Furthermore, structural analysis indicated that the DUF2237-containing protein, enriched in marine environments, has potential interactions with purine nucleotide-containing compounds. This may suggest the existence of unique nucleotide utilization mechanisms in marine bacteria. In addition, we found candidates of virus defense systems within the unknome, which demonstrates that diverse defense systems are present at least in one-third of the cultured SAR11 strains. The conservation of these viral defense systems, even within streamlined SAR11 genomes, suggests that they confer significant ecological advantages. Our exploration provided insights into the genetic basis of bottom-up processes (adaptation to oligotrophic environments) and top-down processes (antiviral defense strategy) contributing to ecological success of SAR11.

Introduction (607 words)

The SAR11 clade, also known as *Candidatus* Pelagibacterales, is the most abundant group of heterotrophic bacteria in marine environments, accounting for approximately 25% of all bacterioplankton cells in the ocean photic zone (1). Due to its high abundance, members of SAR11 clade play crucial roles in marine ecosystems and material cycles (2,3). The notable characteristics of SAR11 clade are their slow growth rate and highly streamlined genome, which typically lacks accessory genes such as antiviral defense systems and, occasionally, essential biosynthetic pathways (3–6). Due to their poor growth under laboratory conditions, culture-based studies of SAR11 clade are limited in achieving comprehensiveness and throughput, and the reasons for their ecological dominance are still not fully resolved.

The ecological success of SAR11 has been explored from both bottom-up and top-down perspectives. On the bottom-up side, their extremely resource-efficient lifestyle is thought to provide a strong advantage in oligotrophic marine environments. Several observations supports this view, including the genome streamlining (5,7), presence of multifunctional transporters and highly expressed solute binding-proteins (8–10), loss of motility and transcriptional regulation (5,7,11), and light-utilization by proteorhodopsin (12,13). In contrast, from a top-down perspective, SAR11 are considered to be less affected by predation and lysis, in line with the “Kill-the-Winner” hypothesis. According

to the hypothesis, fast-growing species are expected to be more susceptible to viral infections, implying that the slow-growing SAR11 strains are less affected by viral infections (14,15). However, the isolation of Pelagiphages (16) and evidence of their frequent infections to SAR11 (17–20) indicate that virus-host interactions are more complex than initially assumed. The "King of the Mountain" (KoM) hypothesis offers one possible explanation, proposing that the high abundance of SAR11 leads to increased contact frequency, promoting frequent genetic recombination, which could influence viral susceptibility (21). The KoM hypothesis is supported by the high intraspecific recombination rate and the presence of highly diverse genomic regions that encode membrane proteins that could act as viral receptors, variation that may help SAR11 populations evade viral infection (22–24). However, there is still a lack of direct evidence for the presence of antiviral defense systems and their frequent horizontal gene transfer among SAR11 (3,20).

Currently, genome-based approaches play indispensable roles in microbiology, particularly for lineages that are difficult to cultivate, such as the SAR11 clade. However, a growing challenge in the field is the discovery of a large number of genes with no similarity to any functionally characterized sequences (25,26). The issue is particularly acute in marine metagenomes (25,27), suggesting that ecologically important genes may

remain overlooked and annotated as hypothetical proteins. For example, the recent discoveries of diverse antiviral defense systems originating form previously uncharacterized genes (28) suggest that many important functions await discovery within this uncharacterized gene space. To overcome the limitation of sequence-based annotation, an increasing number of studies are adopting bioinformatics approaches that go beyond sequence similarity, including protein structure prediction and genomic context analysis (29–31). However, these cutting-edge approaches that could substantially enhance functional annotation have not been systematically applied to SAR11 clade.

Here, we focused on the functionally unknown genes, referred to as the 'functional unknome' (26), within SAR11 genomes and sought to predict their functions by combining bioinformatics approaches. We placed particular emphasis on genes that may be ecologically important for the SAR11 clade, including those related to substrate uptake, phototrophy, and virus-host interactions. By combining structure-based predictions, genomic context information, and environmental distribution patterns, we inferred potential functions for conserved components of the SAR11 unknome. Further, we systematically explored viral defense systems within the SAR11 unknome. Our findings

demonstrate the importance of re-annotating genes with unknown functions to uncover hidden aspects of non-model marine lineages.

Materials and Methods (1212 words)

Collection and evaluation of the SAR11 genomes

We compiled the SAR11 genomes that were analyzed by Haro-Moreno et al. (2020) (32) and expanded the dataset with genomes from freshwater, brackish water, Eastern Pacific, and polar environments to better represent the under-sampled clades (**Table S1**) (33–36). The final dataset includes 321 genomes, mainly from cultured strains and single-amplified genomes (SAGs), as metagenome-assembled genomes (MAGs) are difficult to reconstruct for SAR11 due to high genomic diversity (Fig. S1) (37). Genome quality was assessed by using CheckM2 v1.0.2 (**Fig. S1A**) (38). Estimated genome size was calculated using Completeness and Contamination scores following the previous study (36), and GC content was calculated using Seqkit v2.4.0 (**Fig. S1B**) (39). Among the 253 genomes with >50% completeness and <10% contamination, we identified at least 184 species based on <95% ANI, as calculated using FastANI v1.34 (40).

Phylogenomic classification of the SAR11 genomes

The concatenated alignment of 120 bacterial single-copy genes (5,036 sites;) was constructed using GTDB-Tk v2.1.0 classify-wf (41). Eight genomes were excluded because they had <10% of the alignment sites covered. The *Rickettsia bellii* genome

128 (GCA_002078315.1) served as the outgroup following a previous study (32). The
129 maximum-likelihood tree was inferred with IQ-TREE2 v2.2.0.3 using ModelFinder-
130 selected Q.insect+F+I+R9 model, with 1,000 ultrafast bootstrap and 1,000 SH-aLRT
131 replicates (42,43). The tree was visualized in R using ggplot2, ggtree, and ggtreeExtra
132 (44–46). Subclades were assigned based on tree topology following the previous
133 classifications (6,3,32,36); genomes falling outside known subclades remained
134 unassigned. Although the phylogenetic positions of subclades IV and V are debated and
135 are not classified within the *Ca.* Pelagibacterales in several previous studies (32,47–49),
136 we treated them as part of SAR11 clade, considering their shared genomic and ecological
137 characteristics (3).
138

139 **Functional annotation of ortholog groups and definition of the unknome**

140 A total of 306,416 protein-coding genes in all 321 SAR11 genomes were
141 predicted with prokka v1.14.6 (50), and clustered into ortholog groups (OGs) using
142 SonicParanoid2 v2.0.3 (51). Transmembrane regions were predicted using TMHMM 2.0
143 (52), and proteins were classified as transmembrane (TM) if they met either of the
144 following criteria: (i) ExpAA >10 aa and PredHel >1, or (ii) ExpAA >10 aa, PredHel = 1,
145 and First60 <10. All coding sequences were annotated with COG-2020 and KEGG

146 Orthology (KO) using COGclassifier v1.0.5 (https://github.com/moshi4/COGclassifier)
147 and KofamScan v1.3.0 (53), respectively. The unknome genes were defined as those were
148 not annotated by COGclassifier or assigned COGs belonging to category S (Function
149 unknown), category R (General function prediction only); unknome OGs were those with
150 >50% unknome members. The OGs distributions among each subclade were visualized
151 with the UpSetR (54). Environmental distribution data were retrieved from Global
152 Microbial Gene Catalog (GMGC) v1.0 for bacteriorhodopsin and core unknome OGs
153 based on their COG IDs (55). The core unknome was defined as unknome OGs occurring
154 in more than half of the SAR11 genomes. Only GMGC entries annotated with both habitat
155 categories (e.g., marine, freshwater and human gut) and COG IDs were analyzed. Habitat
156 proportions were calculated per OG, counting multiple habitat annotations as independent
157 occurrences. To simplify classification, "human skin," "human gut," "human nose,"
158 "human vagina", and "human oral" were grouped as "human-associated".
159

160 **Neighborhood OG analysis**

161 Genomic contexts of all coding sequences were analyzed using GFF files
162 generated by the prokka pipeline. Two genes were considered neighbors if they were
163 located within five genes of each other on the same contig. Neighboring pairs that showed

164 a reciprocal co-localization probability of ≥75% across at least ten genomes were retained
165 and visualized as a network using igraph (56) and ggnetwork (57). Since our analysis
166 focused only on a phylogenetically close lineage, we adopted a more conservative
167 threshold than the previous study (29) to avoid false positives due to phylogenetic
168 constraints. Note that pairs with highly unequal occurrence frequencies may not meet this
169 criterion, even if the rarer OG always co-occurs with the more prevalent one. Neighboring
170 pairs containing unknome OGs are manually checked using predicted structures by
171 AlphaFold2, and the complex formation was evaluated by pDockQ and pDockQ2 scores
172 (58–60).
173

174 **Collection and comparison of predicted protein structures**

175 All protein structures of *Ca.* Pelagibacter ubique HTCC1062 strain and *Ca.*
176 Pelagibacter sp. IMCC9063 strain were downloaded from the AlphaFold Protein
177 Structure Database (61). To improve the structural prediction of the conserved 108 core
178 unknome OGs, the structures were predicted with AlphaFold2 based on the core unknome
179 proteins of HTCC1062 excluding the signal peptide regions predicted by DeepTMHMM
180 ver 1.0.24 (62). Structural similarity search for the unknome proteins were performed

181 using Foldseek version 7.04e0ec8 (63) against the Protein Data Bank (PDB), which was
182 downloaded on September 4, 2023 (64).
183

184 **Meta-transcriptome data analysis of SAR11 genes**

185 Sequence reads of the 509 meta-transcriptomes obtained in the Tara Ocean
186 samples were mapped onto the 321 SAR11 genomes (**Table S2**) (65). We downloaded
187 paired-end sequence data from NCBI SRA by SRA toolkit (https://github.com/ncbi/sra-
188 tools), and then quality control was performed by using fastp with options -q 20 -n 10 -l
189 60 (66). The quality-controlled reads were mapped via bwa-mem2 (67). Transcripts per
190 million (TPM) were calculated using a script from a previous study
191 (https://github.com/yosuken/CountMappedReads2) (68).
192

193 **Analysis of DUF2237 domain-containing proteins**

194 A total of 7,968 DUF2237 domain-containing proteins were retrieved from the
195 Pfam database on April 21, 2023. These sequences were annotated with domain
196 information using pfam_scan (69) and aligned using MAFFT v7.508 (70) in the auto
197 mode. The alignment was visualized on the predicted monomer structure of the
198 DUF2237-containing protein of *Ca.* P. ubique HTCC1062 strain using ConSurf-DB (71)

199 and PyMOL v2.5.4 (72). The Predicted Aligned Error plot was visualized using a Python
200 package (https://github.com/cddlab/alphafold3_tools).

201 The OceanDNA MAG catalog v1 (73) and the Pfam database were utilized to assess
202 the phylogenetic distribution of DUF2237 genes. Genes annotated as COG3651 and
203 COG5524 were identified as DUF2237-containing proteins and proteorhodopsin,
204 respectively. A multiple sequence alignment of 120 marker genes from 7,696
205 representative MAGs was generated with the GTDB-Tk classify-wf. The resulting
206 alignment contained 5,036 sites. Subsequently, a maximum likelihood phylogenetic tree
207 was inferred with IQ-TREE2 (v2.2.0.3) using ModelFinder-selected Q.yeast+R10 model,
208 with 1,000 SH-aLRT replicates. The phylogenetic distribution of DUF2237 and
209 proteorhodopsin was visualized on the phylogenetic tree using ggplot2, ggtree, and
210 ggtreeExtra (44–46).

211 The structure of the *Ca. P.* ubique HTCC1062 strain's DUF2237 protein, predicted
212 by AlphaFold2, was prepared in PDBQT format using MGL Tools v1.5.7 (74) for
213 docking simulations. A ligand library was constructed by converting 258 compounds with
214 biological roles from the KEGG Compound database (br8001) and seven nucleotide-
215 containing molecules from the Protein Data Bank Japan (NAI, PO4, NDP, FAD, COA,
216 C2E, NAD) into PDBQT format using Open Babel v3.1.0 (75) and MGL Tools **(Table**

S3). The grid box was defined as 25 × 20 × 20 Å to the pocket site of the DUF2237 protein predicted by PrankWeb 3 (76). All docking simulations were performed using Autodock vina v1.2.5 (77) with an exhaustiveness value of 100. The possibility of binding to ATP was evaluated using ATPbind (78).

Defense system annotation

DefenseFinder (updated on June 5, 2024) was used to identify potential antiviral defense genes (79). To check for the completeness of defense systems, neighboring genes of each defense gene found by DefenseFinder were manually annotated and compared to known defense genes registered to DefenseFinder wiki (80) by genomic context, sequence alignment using MAFFT, and structure similarity search using AlphaFold2 and Foldseek.

Result (2024 words)

Collection and phylogenomic analysis of SAR11 genomes

We collected 321 SAR11 genomes representing five subclades (**Table S1**). Subclade classification based on a maximum-likelihood tree inferred from GTDB marker genes was largely consistent with those inferred from 16S rRNA gene and 232 conserved single-copy genes in previous studies (3,6,32) **(Fig. 1A)**. Subclade V exhibited larger genome size than other subclades **(Fig. S1B)**, consistent with its higher metabolic potential (49). The 306,416 protein-coding genes were clustered into 5,603 OGs using SonicParanoid2 (81). To estimate the unknome fraction, all genes were annotated using the KO and the COG databases. As a result, 173,549 and 257,412 genes were assigned to KO and COG, respectively **(Fig. 1B)**. At the OG level, 1,286 OGs (23%) and 2,683 OGs (48%) were assigned KO and COG annotation, respectively **(Fig. 1CD and Table S4)**.

We next classified these OGs according to their conservation within the five SAR11 subclades. The largest set combination corresponded to OGs conserved across all subclades (968 OGs), exceeding those unique to any single subclade **(Fig. S2A)**. This pattern is consistent with the streamlined SAR11 genomes, in which the core genome constitutes a large proportion (5,6). In contrast, most OGs with unknown functions were

248 specific to clade I or clade V **(**790 and 557 OGs; **Fig. S2B)**, reflecting diversity of clade
249 I and the large genome size of clade V (49). Additionaly, 192 unknown-function OGs
250 specific to clade III, a brackish and freshwater clade, may contribute to the low-salinity
251 adaptation **(Fig. S2B)**.
252
253 Among the COG-assigned genes, categories E (amino acid transport and metabolism,
254 271 OGs), M (cell wall and plasma membrane biosynthesis, 252 OGs), C (energy
255 production, 236 OGs), and J (translation and ribosome-related, 236 OGs) and were the
256 most frequently observed in the SAR11 clade **(Fig. 1D)**. The ordering of these categories,
257 such as category M exceeding category K (transcription), contrasts with that of *E. coli*
258 (82). This pattern reflects their reduced transcriptional regulation, a feature of genome
259 streamlining (5–7). KO-assigned genes were fewer than COG-assigned genes **(Fig. 1BC)**
260 and included functionally unknown genes, such as those containing domains of unknown
261 function (DUF).
262
263 **Definition and characteristics of the 'unknome'**
264 Because some KOs lack functional annotation, we defined the SAR11 unknome
265 based on COG classifications. It comprised category S, R, and OGs without COG

assignment **(Fig. 1B)**, totaling 3,255 OGs (58.1%) **(Fig. 1D)**. The unknome genes are typically short and AT-rich **(Fig. S3AB)**. As in *E. coli* (83), these unknome genes show a slightly higher proportion of TM proteins than functionally known genes **(Fig. 1C, Fisher's exact test p < 2.2e-16)**. Tara Oceans metatranscriptomes confirmed the unknome OGs are expressed in situ, with only slightly lower expression than functionally known genes **(Fig. S3C and Fig. S4, Wilcoxon rank sum test p < 2.2e-16)**.

Functional prediction of the unknome based on neighboring gene functions and predicted protein structures.

Next, we predict the functions of the unknome using two bioinformatics approaches. The first was neighborhood gene analysis, which predicts gene function from the genomic context because functionally related genes in prokaryotes often occur in close proximity in prokaryotic genomes (e.g., in operons or syntenic regions) (84,85). We defined neighboring OG pairs as those located within five genes of each other (details in Methods). We then constructed a neighboring OG network containing 474 OGs and 620 edges **(Fig. 2A)**. Functionally related OGs formed connected components (**Fig. 2A**), the largest of which corresponded to the ribosomal protein operon.

284 Focusing on the connected components that include the unknome OGs (gray nodes),
285 OG1814, a DUF4340-containing OG without a COG assignment, was conserved near the
286 genes annotated as ABC transporter components, such as OG 1815, OG1772, and OG688.
287 In HIMB59 (GCA_000299115.1), proteins from these four OGs were predicted to form
288 a transporter complex with a pDcokQ2 score of 0.268, greater than the confidence
289 threshold of 0.23 **(Fig. 2BC and Fig. S5AB, pDockQ2 > 0.268)**, suggesting that OG1814
290 function as a novel transporter subunit. Similarly, OG1249 (COG4696, a DUF2139-
291 containing gene) was conserved within the type IV pilus operon, indicating a potential
292 role as a minor pilus component (**Fig. 2D**). In total, 38 components containing both
293 unknome and characterized OGs were identified. Although detailed functions remain
294 unresolved for many, these patterns reveal functional links within the SAR11 clade—
295 between unknome and known OGs, and among the unknome themselves.
296
297 We next applied protein structure prediction and structure similarity searches. Using
298 *Ca.* P. ubique HTCC1062 (subclade Ia.3) and LD12 IMCC9063 (subclade IIIa.2) as the
299 representatives, we compared predicted protein structures to those in the PDB. In
300 HTCC1062, 1,212/1,328 proteins (91%) showed >90% probability of being homologous
301 to a PDB entry, and 237 of these shared <20% sequence identity with their best hits (**Fig.**

302 **2E**). In IMCC9063 strain, 1,216/1,382 proteins (88%) showed high probability of
303 structural homologs, 221 of them having <20% sequence identity **(Fig. S6**). Among
304 unknome proteins, structurally similar proteins were found for 154/242 (67%) in
305 HTCC1062 and 164/293 (56%) in IMCC9063. These results demonstrate that structural
306 similarity searches effectively detect remote homologs within the SAR11 genomes.
307

308 **Function prediction of the core unknome OGs**

309 Given that OGs widely conserved throughout the SAR11 clade likely play critical
310 roles for their ecology, we further focused on the 108 core unknome OGs that are
311 conserved across all five clades and present in >162 genomes (>50% of genomes in the
312 dataset). By integrating information from neighboring genes, structurally similar proteins,
313 and KO assignments, 87 of 108 OGs had at least one line of supporting information (**Fig.**
314 **S7**). However, in many cases, the assigned KO showed no functional information, or the
315 functions of the similar structures or neighboring genes were also unknome
316 (**Supplemental Text**). Then, we performed an extensive literature search using gene and
317 domain names registered in the COG, KO, and STRING databases to capture functional
318 information not yet incorporated into current annotations. This search identified
319 functional insights for 10 OGs, including several characterize in *Caulobacter crescentus*,

320 a model alpha-proteobacterium (**Supplemental Text**). Overall, we found that 69 of 108
321 core unknome OGs have additional information or putative functions not registered in
322 COG database **(Table S5 and Supplemental Text)**. These include predicted components
323 of transporters, channels, or metabolic enzymes (e.g., OG228, OG305, OG365, OG448,
324 OG497, OG814, and OG824), which may support the SAR11 clade's adaptation to
325 nutrient-limited environments (**Fig S8 and Supplemental Text**).
326

327 **Functional analysis of DUF2237, the unknown gene with a skewed distribution in**
328 **the ocean**

329 We next examined ecological distribution to infer their potential functions of core
330 unknome OGs lacking functional information from any of the above analyses. We focused
331 on the OGs enriched in marine environment, as such genes often underpin ecological
332 success. A representative example is microbial rhodopsin (COG5524: bacteriorhodopsin)
333 (**Fig. 3A**), which function as a photoreceptor, enabling the capture and utilization of light
334 enagy in oligotrophic waters (86). The most marine-biased functionally unknown OG was
335 COG3908, which contains the DUF2093 domain **(Fig. 3A)**. Although no further
336 functional information was obtained, this OG was consistently located near the lipid
337 synthesis operon in SAR11 clade (**Fig. 2A**).

338

339 The second most marine-biased OG was annotated as COG3651, which contains
340 uncharacterized DUF2237 domain. DUF2237-containing genes have been reported to co-
341 occur with the proteorhodopsin, a light-driven proton pumps, in marine *Flavobacteriia*,
342 and a DUF2237 knockout in *Synechocystis* sp. PCC6803-P showed reduced phototaxis
343 (87). DUF2237 has been suggested to have a function related to light reception or
344 utilization. Although, three different structures of the DUF2237 protein had been
345 experimentally solved (PDB 2LQ3, 3USH, and 4IOH), the mechanistic link to phototaxis
346 remain unclear. Because light utilization is also ecologically important for SAR11 clade
347 (13,88,11), we further analyzed this unknome protein.

348

349 Initially, we re-evaluated the association between proteorhodopsin and DUF2237-
350 containing genes using OceanDNA MAG catalog (73) to validate the previous
351 observation in marine *Flavobacteriia*. As in the previous study, DUF2237 genes
352 significantly co-occurred with the proteorhodopsin **(Fisher's exact test, p < 2.2e-16).**
353 The high conservation of DUF2237 genes across most marine bacterial lineages, except
354 *Patescibacteria* and *Chloroflexota* **(Fig. 3B)**, indicates their broad ecological importance.

The predicted structure of the DUF2237-containing protein (SAR11_1247) indicates dimeric assembly **(Fig. S9AB, pDockQ score = 0.314)**. Analysis of 7,968 DUF2237 domain-containing proteins in the Pfam database showed that this domain typically occurs as a single-domain protein **(Fig. S10AB)**. Mapping sequence conservation from these Pfam entries onto the structure revealed a highly conserved pocket region, likely a concave ligand-binding site **(Fig. 3C)**. This suggests a potential enzymatic role for DUF2237 in interacting with a specific ligand.

To identify potential ligands for the DUF2237 pocket, we performed docking simulations using KEGG "Compounds with biological roles", excluding polysaccharides to avoid including macromolecules. Nucleotide-containing ligands, including NADH, c-di-GMP, and cAMP, showed the highest affinities for the DUF2237 protein from HTCC1062 strain (**Fig. 3D, Table S3**). ATP-binding prediction also supported nucleotide interaction (78) (Probability = 0.85). Because NADH and ATP participate in key metabolic pathways, and c-di-GMP and cAMP act as motility-related second messengers (89–92), these results suggest that the reduced phototaxis observed in the DUF2237 knockout mutant may arise not from a defective light response but rather from altered levels of small nucleotide-containing molecules that could mediate phototactic signaling.

373 **Antiviral defense genes on the SAR11 genomes**

374 Given the numerous recent discoveries of defense genes within previously
375 uncharacterized genes (28), we explored putative antiviral defense genes within the
376 SAR11 genomes. Using DefenseFinder (80), we identified 43 defense systems, including
377 restriction-modification (RM) and toxin-antitoxin systems **(Table S6)**. The RM systems
378 were most common as previously reported (3,20). Excluding the well-characterized RM
379 systems, 11 unknome genes showed putative defense roles **(Fig. S11)**. Furthermore, we
380 examined genomic regions with only partial components of known systems and manually
381 curated neighboring genes, as defense genes often cluster within specific loci. This led
382 to the identification of 8 additional putative defense systems, bringing the total to 51
383 putative defense systems across 46 SAR11 genomes **(Fig. 4A)**.

384

385 A previous study showed that only a few deep sea SAR11 genomes possess part of
386 the CRISPR-Cas-like gene set (93). As in the previous study, Cas-like genes without
387 CRISPR array were detected in two deep sea SAGs (AG-337-I02 and AG-410-B18).
388 Three cultured-strains genomes and one MAG possess *dnd*, a dsDNA phosphorothioation
389 (PT)-based defense system (94). In addition, we identified Ssp-like operons on two
390 SAR11 genomes **(Fig. 4A)**. Same as the *dnd*, Ssp is a PT-dependent system, modifing
391 DNA by SspBCD and cleaving foreign ssDNA by SspE (95,96). The SAR11 Ssp-like
392 operons encode proteins structurally similar to the known SspBCD and an
393 uncharacterized protein shorter than the known SspE. Although experimental validation

394 is needed, this finding indicates that PT-based modifications are used to distinguish self
395 from foreign DNA in the SAR11 clade.

396 We also identified Mza, DISARM, and ShosTA-like operons **(Fig. 4A)**. Mza
397 defends against ssDNA phages, although its mechanism remains unknown (97). The
398 DISARM-like operon in SAR11 encodes DrmAB, DrMII-type methyltransferase, and
399 helicases **(Fig. 4B)**, whereas DrmC, nonessential for defense, appears absent (98). The
400 distinct gene order suggests a divergent DISARM-like configuration. The ShosTA toxin-
401 antitoxin system was detected in three SAR11 genomes. The SAR11 ShosT protein is
402 approximately 100 amino acids shorter than the *E. coli* ShosT but retains the
403 phosphoribosyl transferase domain **(Fig. 4C)**. This suggests that the SAR11 clade may
404 have acquired a truncated but functional toxin, likely through genome reduction. AbiRb-
405 and AbiRc-like genes were also present as operons, but the AbiRc-like gene in SAR11
406 lacks its terminal domain **(Fig. 4D)**. Moreover, AbiRa is a sigma70 factor, yet the only
407 sigma70-type RNA polymerase subunit is present in the SAR11 genome, it is not located
408 within this operon. Therefore, it remains unclear whether this AbiR-like genes functions
409 as a known abortive infection system, and we did not include it as a defense system found
410 in this study.

411 Previous limited identifications of viral defense genes in SAR11 may be partly
412 due to the presence of these overlooked putative defense systems that slightly differ from
413 experimentally validated defense system components (**Fig. 4BCD**). When we focused on
414 high-quality genomes, defense systems were found in at least one-third of the cultured
415 strains and one-quarter of the SAGs. In contrast, defense systems were rarely detected
416 from the MAGs **(Fig. 4E)**. This result suggests that the SAR11 clade harbors more diverse
417 and abundant defense systems than previously inferred from the MAG-based analyses
418 (20).
419

420 **Discussion (989 words)**

421 As genomic data continue to expand exponentially, functionally unknown genes are
422 becoming more prevalent, and their study—termed Functional Unknomics (26)—is
423 gaining attention. Previous research on genes of unknown function has primarily focused
424 on model organisms (83,26,99) or large-scale sequence analyses of whole ecosystems,
425 involving MAGs and SAGs (25,29,30). This study represents the first systematic
426 investigation focused on genes of unknown function in the SAR11 clade as one of the
427 non-model organism lineages.

428

429 In the SAR11 clade, one of the most abundant marine bacterial lineages, a large
430 fraction of OGs remains functionally uncharacterized despite their crucial roles in marine
431 microbial ecology. To address this issue, we clustered 306,416 protein-coding genes in
432 321 SAR11 genomes into 5,603 OGs and classified 3,255 of them (58.1%) as unknome
433 based on the COG database **(Fig. 1CD)**. Utilizing genome context and predicted protein
434 structures, we successfully obtained functional information for 69 out of the 108 core
435 unknomes **(Table S5 and Supplemental Text).** These included many transporter genes,
436 supporting the strategy of SAR11 that specializes in substrate uptake, as proposed in
437 previous studies (3,9,10). Furthermore, our analysis confirmed that the utilization of

438 genomic context and predicted protein structures is valuable even for non-model
439 organisms. Especially, structural similarity search proved particularly effective in
440 functional annotation (**Fig. 2D, S7, S8 and Supplemental Text**). This may reflect both
441 the accumulation of neutral mutations during the divergence of SAR11 clade from well-
442 studied model organisms and shifts in genomic properties, including reduced GC content,
443 leading to the presence of many remote homologs (5). Although structural similarity
444 search can find distant relationships between SAR11 proteins and those in the PDB, the
445 latter also lack functional information, such as the DUF2237-containing protein. In this
446 study, we addressed this problem by manually examining the individual results. This
447 involved cross-referencing them with databases such as PDB, Pfam, COG, KO, Uniprot,
448 and STRING, and conducting literature searches for any reported functional insights
449 **(Supplemental Text)**. This careful curation allowed us to focus on previously
450 uncharacterized genes potentially support ecological success of SAR11 clade
451 (**Supplementary Text**).
452
453 We found functionally unknown genes specifically related to the marine environment,
454 which are probably crucial for the ecology of the SAR11 clade **(Fig. 3A)**. Among them,
455 we found the DUF2237-containing proteins, which were previously thought to be related

456 to proteorhodopsin (87). The predicted protein structure and docking simulation results
457 suggested that DUF2237 is likely to be an enzyme associated with purine nucleotide-
458 containing compounds **(Fig. 3CDE)**. This result may indicate that the DUF2237 protein
459 binds to second messengers, such as c-di-GMP and cAMP, and is thus involved in
460 phototaxis. Alternatively, it might interact with NADH, indicating involvement in redox
461 processes. This finding collectively suggests that DUF2237 participates in fundamental
462 cellular processes, extending beyond a direct link to phototaxis. Although definitive
463 conclusions require further experimental validation, the possible binding of DUF2237 to
464 these compounds suggests that it represents a novel nucleotide-binding motif and it may
465 play a key role in marine bacterial ecosystems.
466
467 We explored the unknome genes to identify viral defense systems that may offer
468 insights into top-down factors contributing to ecological success of the SAR11 clade.
469 Previous studies suggested that SAR11 possesses only limited viral defense capacity
470 (3,20). However, our comprehensive genomic analysis revealed a scattered distribution
471 of diverse viral defense systems **(Fig. 4A)**. This pattern is consistent with the high
472 intraspecific recombination rates of SAR11 genes (22). While the previous hypothesis
473 proposed that a slow growth rate of SAR11 made it less vulnerable to phage lysis (15),

474 the discovery of antiviral defense systems indicates that SAR11 face viral pressure,
475 requiring more active resistance to viral infection than previously appreciated. The recent
476 discovery of anti-RM systems in the Pelagiphage genome further supports the existence
477 of the arms race between *Ca.* Pelagibacter and the Pelagiphage (80). The presence of
478 defense systems might potentially explain the unique ribosome-deprived phenomenon
479 known as 'zombie cells' in SAR11 upon viral infection, possibly resulting from abortive
480 infection (20). Moreover, the discovery of the Ssp and Mza systems in the SAR11
481 genomes, both of which are known to target ssDNA viruses, suggests the potential
482 existence of undiscovered ssDNA Pelagiphages. Given the high intraspecific
483 recombination rate of SAR11 genes (22), the sporadic distribution of defense systems
484 likely reflects population-level sharing of this repertoire, rather than frequent transfer
485 across distantly related taxa. Such population-level sharing may help SAR11 balance the
486 trade-off between genome streamlining and antiviral defense. This interpretation is
487 consistent with the concept of a bacterial pan-immune system, in which defense genes
488 are frequently gained and lost among closely related strains (100), and also consistent
489 with the KoM hypothesis, which proposes SAR11 maintains its dominance through high
490 diversity and efficient horizontal gene transfer driven by high contact frequency of closely
491 related strains (16,21). Collectively, these observations suggest a complex evolutionary

492 interplay between SAR11 and its phages, likely shaped by genome streamlining,
493 horizontal gene transfer, and diverse defense strategies.
494
495 Our results also underscore the importance of genome reconstruction methods in
496 evaluating the distribution of viral defense systems. Compared with SAGs and genomes
497 from cultured strains, viral defense genes were underrepresented in MAGs **(Fig. 4D)**,
498 which may have led to the previous observation that SAR11 possesses little viral defense
499 capacity (3,20). This underrepresentation likely stems from limitations in MAG assembly
500 and binning, reflecting the high variability of specific genomic regions (e.g., hyper
501 variable regions and metagenomic island) (6,23,24). Consequently, SAG analysis
502 provides a powerful approach to capture diversity of defense systems, as it circumvents
503 the challenges associated with MAG construction (37).
504
505 Looking forward, further advances in automated gene function annotation will be
506 essential to keep pace with the rapid expansion of microbial genomic data and to
507 complement meticulous but low-throughput manual annotation. By developing robust
508 analytical approaches and moving beyond model organisms, we can uncover novel
509 biological and ecological functions of microbes, providing deeper insight into the

diversity and ecological roles of microbial life in natural environments. Bioinformatics-driven Functional Unknomics is poised to become a major frontier in microbial ecology, particularly for lineages that remain difficult to culture.

Acknowledgments

The authors would like to acknowledge Wataru Iwasaki for providing computational resources. Computation time was provided by the SuperComputer System, Institute for Chemical Research, Kyoto University.

Author contributions

K.T., S.N., Y.N., and K.O. conceived and designed this study. S.N. collected data and performed the bioinformatic analysis and data visualization throughout this study. K.T. performed the meta-transcriptomic read mapping analysis. K.T., Y.N., K.O., T.D., K.H., and S.Y. provided critical interpretations of the data. S.N. drafted the manuscript. All authors revised and approved the final version.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Funding

This work was supported by JST GteX Grant Number JPMJGX23B2 and JSPS KAKENHI Grant Number JP22H04925 and JP24KJ0979.

Data availability

The code and datasets generated and analyzed in this study are available in our Zenodo repository, https://doi.org/10.5281/zenodo.16318636.

538 **Reference**

539 1. Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, et
540 al. SAR11 clade dominates ocean surface bacterioplankton communities. Nature. 2002
541 Dec;420(6917):806–10.
542 2. Sun J, Steindler L, Thrash JC, Halsey KH, Smith DP, Carter AE, et al. One
543 Carbon Metabolism in SAR11 Pelagic Marine Bacteria. PLOS ONE. 2011 Aug
544 23;6(8):e23973.
545 3. Giovannoni SJ. SAR11 Bacteria: The Most Abundant Plankton in the Oceans.
546 Annu Rev Mar Sci. 2017;9(1):231–55.
547 4. Rappé MS, Connon SA, Vergin KL, Giovannoni SJ. Cultivation of the
548 ubiquitous SAR11 marine bacterioplankton clade. Nature. 2002 Aug;418(6898):630–3.
549 5. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, et al.
550 Genome Streamlining in a Cosmopolitan Oceanic Bacterium. Science. 2005 Aug
551 19;309(5738):1242–5.
552 6. Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ, et al.
553 Streamlining and Core Genome Conservation among Highly Divergent Members of the
554 SAR11 Clade. mBio. 2012 Sep 18;3(5):10.1128/mbio.00252-12.
555 7. Noell SE, Hellweger FL, Temperton B, Giovannoni SJ. A Reduction of
556 Transcriptional Regulation in Aquatic Oligotrophic Microorganisms Enhances Fitness in
557 Nutrient-Poor Environments. Microbiol Mol Biol Rev. 2023 Mar 30;87(2):e00124-22.
558 8. Sowell SM, Wilhelm LJ, Norbeck AD, Lipton MS, Nicora CD, Barofsky DF, et
559 al. Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in
560 the Sargasso Sea. ISME J. 2009 Jan;3(1):93–105.
561 9. Noell SE, Giovannoni SJ. SAR11 bacteria have a high affinity and

562 multifunctional glycine betaine transporter. Environ Microbiol. 2019;21(7):2559–75.

563 10. Clifton BE, Alcolombri U, Uechi GI, Jackson CJ, Laurino P. The ultra-high
564 affinity transport proteins of ubiquitous marine bacteria. Nature. 2024
565 Oct;634(8034):721–8.

566 11. Noell SE, Brennan E, Washburn Q, Davis II EW, Hellweger FL, Giovannoni SJ.
567 Differences in the regulatory strategies of marine oligotrophs and copiotrophs reflect
568 differences in motility. Environ Microbiol. 2023;25(7):1265–80.

569 12. Giovannoni SJ, Bibbs L, Cho JC, Stapels MD, Desiderio R, Vergin KL, et al.
570 Proteorhodopsin in the ubiquitous marine bacterium SAR11. Nature. 2005
571 Nov;438(7064):82–5.

572 13. Steindler L, Schwalbach MS, Smith DP, Chan F, Giovannoni SJ. Energy Starved
573 Candidatus Pelagibacter Ubique Substitutes Light-Mediated ATP Production for
574 Endogenous Carbon Respiration. PLOS ONE. 2011 May 9;6(5):e19725.

575 14. Thingstad TF. Elements of a theory for the mechanisms controlling abundance,
576 diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. Limnol
577 Oceanogr. 2000;45(6):1320–8.

578 15. Suttle CA. Marine viruses — major players in the global ecosystem. Nat Rev
579 Microbiol. 2007 Oct;5(10):801–12.

580 16. Zhao Y, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC, et
581 al. Abundant SAR11 viruses in the ocean. Nature. 2013 Feb;494(7437):357–60.

582 17. Yoshida T, Nishimura Y, Watai H, Haruki N, Morimoto D, Kaneko H, et al.
583 Locality and diel cycling of viral production revealed by a 24 h time course cross-omics
584 analysis in a coastal region of Japan. ISME J. 2018 May;12(5):1287–95.

585 18. Martinez-Hernandez F, Fornas Ò, Lluesma Gomez M, Garcia-Heredia I,

586 Maestre-Carballa L, López-Pérez M, et al. Single-cell genomics uncover Pelagibacter as
587 the putative host of the extremely abundant uncultured 37-F6 viral population in the ocean.
588 ISME J. 2019 Jan 1;13(1):232–6.
589 19. Tominaga K, Ogawa-Haruki N, Nishimura Y, Watai H, Yamamoto K, Ogata H,
590 et al. Prevalence of Viral Frequency-Dependent Infection in Coastal Marine Prokaryotes
591 Revealed Using Monthly Time Series Virome Analysis. mSystems. 2023
592 Feb;8(1):e00931-22.
593 20. Brüwer JD, Sidhu C, Zhao Y, Eich A, Rößler L, Orellana LH, et al. Globally
594 occurring pelagiphage infections create ribosome-deprived cells. Nat Commun. 2024
595 May 2;15(1):3715.
596 21. Giovannoni S, Temperton B, Zhao Y. Giovannoni et al. reply. Nature. 2013
597 Jul;499(7459):E4–5.
598 22. Vergin KL, Tripp HJ, Wilhelm LJ, Denver DR, Rappé MS, Giovannoni SJ. High
599 intraspecific recombination rate in a native population of Candidatus Pelagibacter ubique
600 (SAR11). Environ Microbiol. 2007;9(10):2430–40.
601 23. Gilbert JA, Mühling M, Joint I. A rare SAR11 fosmid clone confirming genetic
602 variability in the 'Candidatus Pelagibacter ubique' genome. ISME J. 2008 Jul;2(7):790–
603 3.
604 24. Rodriguez-Valera F, Martín-Cuadrado AB, Rodriguez-Brito B, Pasic L,
605 Thingstad TF, Rohwer F, et al. Explaining microbial population genomics through phage
606 predation. Nat Preced. 2009 Jul 29;1–1.
607 25. Vanni C, Schechter MS, Acinas SG, Barberán A, Buttigieg PL, Casamayor EO,
608 et al. Unifying the known and unknown microbial coding sequence space. Brown CT,
609 Storz G, Brown CT, Smith B, editors. eLife. 2022 Mar 31;11:e67667.

610 26. Rocha JJ, Jayaram SA, Stevens TJ, Muschalik N, Shah RD, Emran S, et al.
611 Functional unknomics: Systematic screening of conserved genes of unknown function.
612 PLOS Biol. 2023 Aug 8;21(8):e3002222.
613 27. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al.
614 Structure and function of the global ocean microbiome. Science. 2015 May
615 22;348(6237):1261359.
616 28. Georjon H, Bernheim A. The highly diverse antiphage defence systems of
617 bacteria. Nat Rev Microbiol. 2023 Oct;21(10):686–700.
618 29. Pavlopoulos GA, Baltoumas FA, Liu S, Selvitopi O, Camargo AP, Nayfach S,
619 et al. Unraveling the functional dark matter through global metagenomics. Nature. 2023
620 Oct;622(7983):594–602.
621 30. Rodríguez del Río Á, Giner-Lamia J, Cantalapiedra CP, Botas J, Deng Z,
622 Hernández-Plaza A, et al. Functional and evolutionary significance of unknown genes
623 from uncultivated taxa. Nature. 2023 Dec 18;1–3.
624 31. Jha N, Kravitz J, West-Roberts J, Lu C, Camargo AP, Roux S, et al. Gaia: An
625 AI-enabled genomic context–aware platform for protein sequence annotation. Sci Adv.
626 2025 Jun 20;11(25):eadv5109.
627 32. Haro-Moreno JM, Rodriguez-Valera F, Rosselli R, Martinez-Hernandez F,
628 Roda-Garcia JJ, Gomez ML, et al. Ecogenomics of the SAR11 clade. Environ Microbiol.
629 2020 May;22(5):1748–63.
630 33. Hiraoka S, Okazaki Y, Anda M, Toyoda A, Nakano S ichi, Iwasaki W.
631 Metaepigenomic analysis reveals the unexplored diversity of DNA methylation in an
632 environmental prokaryotic community. Nat Commun. 2019 Jan 11;10(1):159.
633 34. Hiraoka S, Sumida T, Hirai M, Toyoda A, Kawagucci S, Yokokawa T, et al.

634 Diverse DNA modification in marine prokaryotic and viral communities. Nucleic Acids
635 Res. 2022 Feb 22;50(3):1531–50.
636 35. Martínez-Pérez C, Greening C, Bay SK, Lappan RJ, Zhao Z, De Corte D, et al.
637 Phylogenetically and functionally diverse microorganisms reside under the Ross Ice Shelf.
638 Nat Commun. 2022 Jan 10;13(1):117.
639 36. Lanclos VC, Rasmussen AN, Kojima CY, Cheng C, Henson MW, Faircloth BC,
640 et al. Ecophysiology and genomics of the brackish water adapted SAR11 subclade IIIa.
641 ISME J. 2023 Apr;17(4):620–9.
642 37. Chang T, Gavelis GS, Brown JM, Stepanauskas R. Genomic representativeness
643 and chimerism in large collections of SAGs and MAGs of marine prokaryoplankton.
644 Microbiome. 2024 Jul 15;12(1):126.
645 38. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: a rapid, scalable
646 and accurate tool for assessing microbial genome quality using machine learning. Nat
647 Methods. 2023 Aug;20(8):1203–12.
648 39. Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for
649 FASTA/Q File Manipulation. PLOS ONE. 2016 Oct 5;11(10):e0163962.
650 40. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High
651 throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries.
652 Nat Commun. 2018 Nov 30;9(1):5114.
653 41. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory
654 friendly classification with the genome taxonomy database. Bioinforma Oxf Engl. 2022
655 Nov;38(23):5315–6.
656 42. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS.
657 ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods.

2017 Jun;14(6):587–9.

43. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020 May 1;37(5):1530–4.

44. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2nd ed. 2016版. Switzerland: Springer; 2016. 276 p.

45. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8(1):28–36.

46. Xu S, Dai Z, Guo P, Fu X, Liu S, Zhou L, et al. ggtreeExtra: Compact Visualization of Richly Annotated Phylogenetic Data. Mol Biol Evol. 2021 Aug;38(9):4039–42.

47. Viklund J, Martijn J, Ettema TJG, Andersson SGE. Comparative and Phylogenomic Evidence That the Alphaproteobacterium HIMB59 Is Not a Member of the Oceanic SAR11 Clade. PLOS ONE. 2013 Nov 1;8(11):e78858.

48. Muñoz-Gómez SA, Hess S, Burger G, Lang BF, Susko E, Slamovits CH, et al. An updated phylogeny of the Alphaproteobacteria reveals that the parasitic Rickettsiales and Holosporales have independent origins. Rokas A, Wittkopp PJ, Irisarri I, editors. eLife. 2019 Feb 21;8:e42535.

49. Getz EW, Lanclos VC, Kojima CY, Cheng C, Henson MW, Schön ME, et al. The AEGEAN-169 clade of bacterioplankton is synonymous with SAR11 subclade V (HIMB59) and metabolically distinct. mSystems. 2023 May 18;8(3):e00179-23.

50. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068–9.

682 51. Cosentino S, Iwasaki W. SonicParanoid2: fast, accurate, and comprehensive
683 orthology inference with machine learning and language models [Internet]. bioRxiv; 2023
684 [cited 2023 Nov 22]. p. 2023.05.14.540736. Available from:
685 https://www.biorxiv.org/content/10.1101/2023.05.14.540736v2
686 52. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting
687 transmembrane protein topology with a hidden markov model: application to complete
688 genomes1. J Mol Biol. 2001 Jan 19;305(3):567–80.
689 53. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al.
690 KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score
691 threshold. Bioinformatics. 2020 Apr 1;36(7):2251–2.
692 54. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization
693 of intersecting sets and their properties. Bioinformatics. 2017 Sep 15;33(18):2938–40.
694 55. Coelho LP, Alves R, del Río ÁR, Myers PN, Cantalapiedra CP, Giner-Lamia J,
695 et al. Towards the biogeography of prokaryotic genes. Nature. 2022 Jan
696 1;601(7892):252–6.
697 56. Csardi G, Nepusz T. The igraph software package for complex network research.
698 InterJournal. 2006;Complex Systems:1695.
699 57. Briatte F. ggnetwork: Geometries to Plot Networks with 'ggplot2' [Internet].
700 2023. Available from: https://github.com/briatte/ggnetwork
701 58. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly
702 accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583–
703 9.
704 59. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein
705 complex prediction with AlphaFold-Multimer [Internet]. bioRxiv; 2022 [cited 2023 Dec

706 30]. p. 2021.10.04.463034. Available from:
707 https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2
708 60. Zhu W, Shenoy A, Kundrotas P, Elofsson A. Evaluation of AlphaFold-Multimer
709 prediction on multi-chain protein complexes. Bioinformatics. 2023 Jul 1;39(7):btad424.
710 61. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al.
711 AlphaFold Protein Structure Database: massively expanding the structural coverage of
712 protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022 Jan
713 7;50(D1):D439–44.
714 62. Hallgren J, Tsirigos KD, Pedersen MD, Armenteros JJA, Marcatili P, Nielsen H,
715 et al. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural
716 networks [Internet]. bioRxiv; 2022 [cited 2023 Dec 12]. p. 2022.04.08.487609. Available
717 from: https://www.biorxiv.org/content/10.1101/2022.04.08.487609v1
718 63. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al.
719 Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2023 May 8;1–
720 4.
721 64. Bekker GJ, Yokochi M, Suzuki H, Ikegawa Y, Iwata T, Kudou T, et al. Protein
722 Data Bank Japan: Celebrating our 20th anniversary during a global pandemic as the Asian
723 hub of three dimensional macromolecular structural data. Protein Sci. 2022;31(1):173–
724 86.
725 65. Salazar G, Paoli L, Alberti A, Huerta-Cepas J, Ruscheweyh HJ, Cuenca M, et al.
726 Gene Expression Changes and Community Turnover Differentially Shape the Global
727 Ocean Metatranscriptome. Cell. 2019 Nov 14;179(5):1068-1083.e21.
728 66. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ
729 preprocessor. Bioinformatics. 2018 Sep 1;34(17):i884–90.

730 67. Vasimuddin Md, Misra S, Li H, Aluru S. Efficient Architecture-Aware
731 Acceleration of BWA-MEM for Multicore Systems. In: 2019 IEEE International Parallel
732 and Distributed Processing Symposium (IPDPS) [Internet]. 2019 [cited 2025 Jan 15]. p.
733 314–24. Available from: https://ieeexplore.ieee.org/document/8820962
734 68. Nishimura Y, Watai H, Honda T, Mihara T, Omae K, Roux S, et al.
735 Environmental Viral Genomes Shed New Light on Virus-Host Interactions in the Ocean.
736 mSphere. 2017 Mar;2(2):10.1128/msphere.00359-16.
737 69. Madeira F, Pearce M, Tivey ARN, Basutkar P, Lee J, Edbali O, et al. Search and
738 sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res. 2022 Jul
739 5;50(W1):W276–9.
740 70. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software
741 Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013 Apr
742 1;30(4):772–80.
743 71. Chorin AB, Masrati G, Kessel A, Narunsky A, Sprinzak J, Lahav S, et al.
744 ConSurf-DB: An accessible repository for the evolutionary conservation patterns of the
745 majority of PDB proteins. Protein Sci Publ Protein Soc. 2020 Jan;29(1):258.
746 72. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. 2015.
747 73. Nishimura Y, Yoshizawa S. The OceanDNA MAG catalog contains over 50,000
748 prokaryotic genomes originated from various marine environments. Sci Data. 2022 Jun
749 17;9(1):305.
750 74. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, et al.
751 AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.
752 J Comput Chem. 2009;30(16):2785–91.
753 75. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR.

754 Open Babel: An open chemical toolbox. J Cheminformatics. 2011 Oct;3:33.

755 76. Jakubec D, Skoda P, Krivak R, Novotny M, Hoksza D. PrankWeb 3: accelerated
756 ligand-binding site predictions for experimental and modelled protein structures. Nucleic
757 Acids Res. 2022 Jul 5;50(W1):W593–7.

758 77. Eberhardt J, Santos-Martins D, Tillack AF, Forli S. AutoDock Vina 1.2.0: New
759 Docking Methods, Expanded Force Field, and Python Bindings. J Chem Inf Model. 2021
760 Aug 23;61(8):3891–8.

761 78. Hu J, Li Y, Zhang Y, Yu DJ. ATPbind: Accurate Protein–ATP Binding Site
762 Prediction by Combining Sequence-Profiling and Structure-Based Comparisons. J Chem
763 Inf Model. 2018 Feb 26;58(2):501–10.

764 79. Tesson F, Hervé A, Mordret E, Touchon M, d'Humières C, Cury J, et al.
765 Systematic and quantitative view of the antiviral arsenal of prokaryotes. Nat Commun.
766 2022 May 10;13(1):2561.

767 80. Tesson F, Huiting E, Wei L, Ren J, Johnson M, Planel R, et al. Exploring the
768 diversity of anti-defense systems across prokaryotes, phages, and mobile genetic
769 elements [Internet]. bioRxiv; 2024 [cited 2024 Nov 8]. p. 2024.08.21.608784. Available
770 from: https://www.biorxiv.org/content/10.1101/2024.08.21.608784v1

771 81. Cosentino S, Sriswasdi S, Iwasaki W. SonicParanoid2: fast, accurate, and
772 comprehensive orthology inference with machine learning and language models. Genome
773 Biol. 2024 Jul 25;25(1):195.

774 82. Shimoyama Y. COGclassifier: A tool for classifying prokaryote protein
775 sequences into COG functional category [Internet]. 2022 [cited 2023 Nov 22]. Available
776 from: https://github.com/moshi4/COGclassifier

777 83. Ghatak S, King ZA, Sastry A, Palsson BO. The y-ome defines the 35% of

Escherichia coli genes that lack experimental evidence of function. Nucleic Acids Res. 2019 Mar 18;47(5):2446–54.

84. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci. 1999 Mar 16;96(6):2896–901.

85. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021 Jan 8;49(D1):D605–12.

86. Gómez-Consarnau L, Raven JA, Levine NM, Cutter LS, Wang D, Seegers B, et al. Microbial rhodopsins are major contributors to the solar energy captured in the sea. Sci Adv. 2019 Aug 7;5(8):eaaw8855.

87. Kumagai Y, Yoshizawa S, Nakajima Y, Watanabe M, Fukunaga T, Ogura Y, et al. Solar-panel and parasol strategies shape the proteorhodopsin distribution pattern in marine Flavobacteriia. ISME J. 2018 May;12(5):1329–43.

88. Munson-McGee JH, Lindsay MR, Sintes E, Brown JM, D'Angelo T, Brown J, et al. Decoupling of respiration rates and abundance in marine prokaryoplankton. Nature. 2022 Dec;612(7941):764–70.

89. Bhaya D, Nakasugi K, Fazeli F, Burriesci MS. Phototaxis and Impaired Motility in Adenylyl Cyclase and Cyclase Receptor Protein Mutants of Synechocystis sp. Strain PCC 6803. J Bacteriol. 2006 Oct;188(20):7306–10.

90. Hengge R. Principles of c-di-GMP signalling in bacteria. Nat Rev Microbiol. 2009 Apr;7(4):263–73.

91. McDonough KA, Rodriguez A. The myriad roles of cyclic AMP in microbial pathogens: from signal to sword. Nat Rev Microbiol. 2012 Jan;10(1):27–38.

802 92. Hengge R, Gründling A, Jenal U, Ryan R, Yildiz F. Bacterial Signal
803 Transduction by Cyclic Di-GMP and Other Nucleotide Second Messengers. J Bacteriol.
804 2015 Dec 14;198(1):15–26.
805 93. Thrash CJ, Temperton B, Swan BK, Landry ZC, Woyke T, DeLong EF, et al.
806 Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. ISME J.
807 2014 Jul;8(7):1440–51.
808 94. Wang L, Chen S, Vergin KL, Giovannoni SJ, Chan SW, DeMott MS, et al. DNA
809 phosphorothioation is widespread and quantized in bacterial genomes. Proc Natl Acad
810 Sci. 2011 Feb 15;108(7):2963–8.
811 95. Xiong X, Wu G, Wei Y, Liu L, Zhang Y, Su R, et al. SspABCD–SspE is a
812 phosphorothioation-sensing bacterial defence system with broad anti-phage activities.
813 Nat Microbiol. 2020 Jul;5(7):917–28.
814 96. Jiang S, Chen K, Wang Y, Zhang Y, Tang Y, Huang W, et al. A DNA
815 phosphorothioation-based Dnd defense system provides resistance against various phages
816 and is compatible with the Ssp defense system. mBio. 2023 Jun;14(4):e00933-23.
817 97. Gao L, Altae-Tran H, Böhning F, Makarova KS, Segel M, Schmid-Burgk JL, et
818 al. Diverse enzymatic activities mediate antiviral immunity in prokaryotes. Science. 2020
819 Aug 28;369(6507):1077–84.
820 98. Ofir G, Melamed S, Sberro H, Mukamel Z, Silverman S, Yaakov G, et al.
821 DISARM is a widespread bacterial defence system with broad anti-phage activities. Nat
822 Microbiol. 2018 Jan;3(1):90–8.
823 99. Moore LR, Caspi R, Boyd D, Berkmen M, Mackie A, Paley S, et al. Revisiting
824 the y-ome of Escherichia coli. Nucleic Acids Res. 2024 Nov 11;52(20):12201–7.
825 100. Bernheim A, Sorek R. The pan-immune system of bacteria: antiviral defence as

a community resource. Nat Rev Microbiol. 2020 Feb;18(2):113–9.

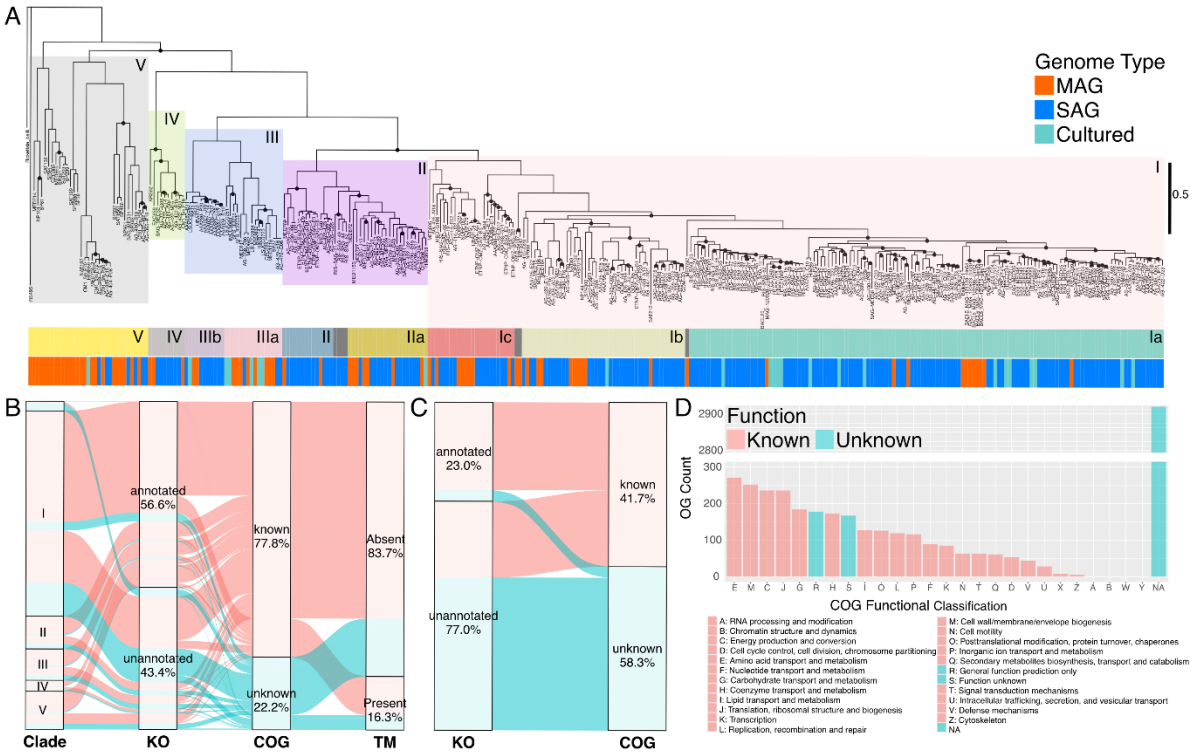

Figure 1. Characterization of the Unknome from a phylogenetically comprehensive SAR11 genome dataset.

(A) Maximum likelihood phylogenetic tree of SAR11 genomes. The color coding in the phylogenetic tree indicates five subclade classifications. The colors below provides more detailed subclade classifications and types of the genomes, that is MAG (red), SAG (blue), or pure culture (light blue). The dots on the tree show UltraFast Bootstrap values > 95%. **(B,C)** Sankey plots summarize the annotation profiles of all genes/OGs and transmembrane region prediction results. Blue shows the unknome genes/OGs. TM indicates transmembrane proteins predicted using TMHMM 2.0 (see Methods for details). **(D)** COG annotation results for all OGs. NA denotes sequences not annotated in the COG database. Blue bars show unknome genes in this study.

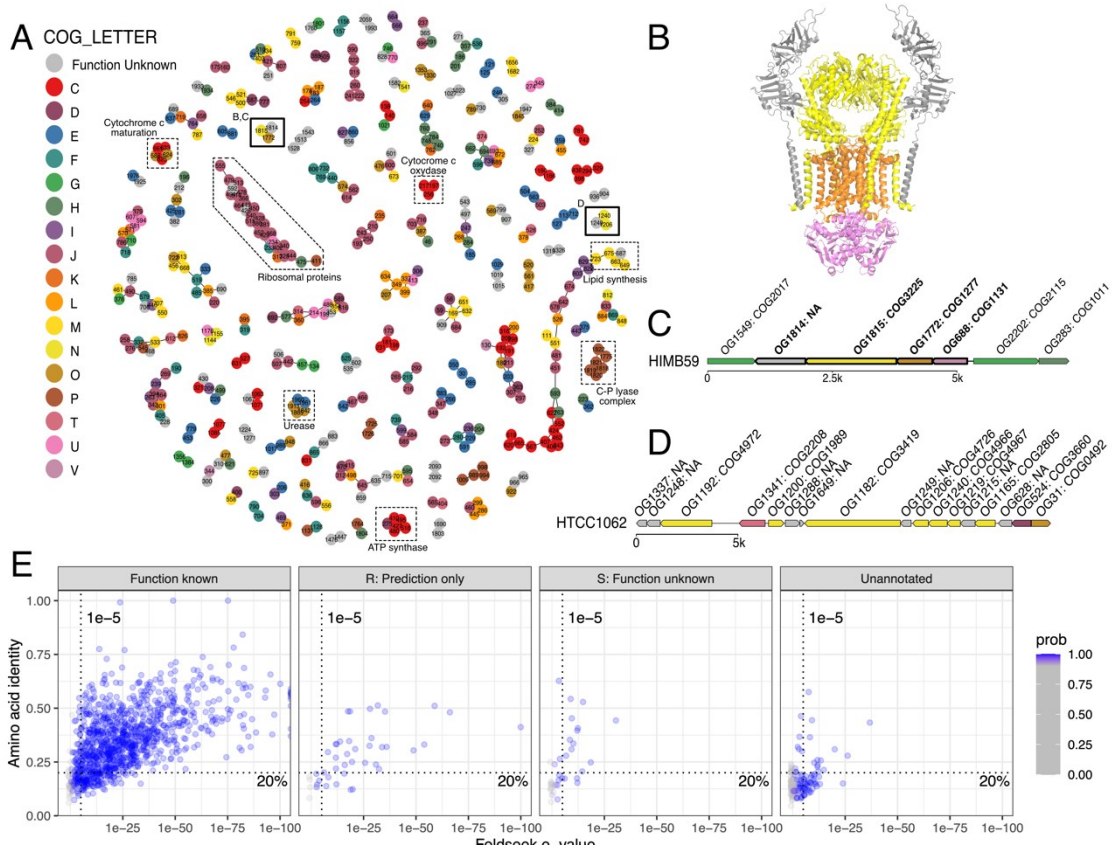

Figure 2. Function prediction based on genomic context and protein structural similarity search.

(A) Summary of neighboring gene analysis. Network nodes represent each OGs, and colors indicate the COG functional categories. Pairs with a co-localization probability ≧75% within five genes upstream and downstream on the same contig are connected by edges. Several known operons are highlighted with dashed outlines as examples. **(B)** Predicted protein complex structure of OG688, OG1772, OG1815, and OG1549 using AlphaFold-multimer. The complex forms an octameric assembly composed of four homodimers. The colors show the COG functional category. **(C, D)** Genome context around each network component in panel A. The genes outlined in bold were used for protein structure prediction in panel B. **(E)** Scatter plots of all *Ca.* P. ubique HTCC1062 strain proteins compared with known protein structures in the PDB. The dot lines show e-value = 1e-5 and amino acid identity = 20%, highlighting potential remote homologs

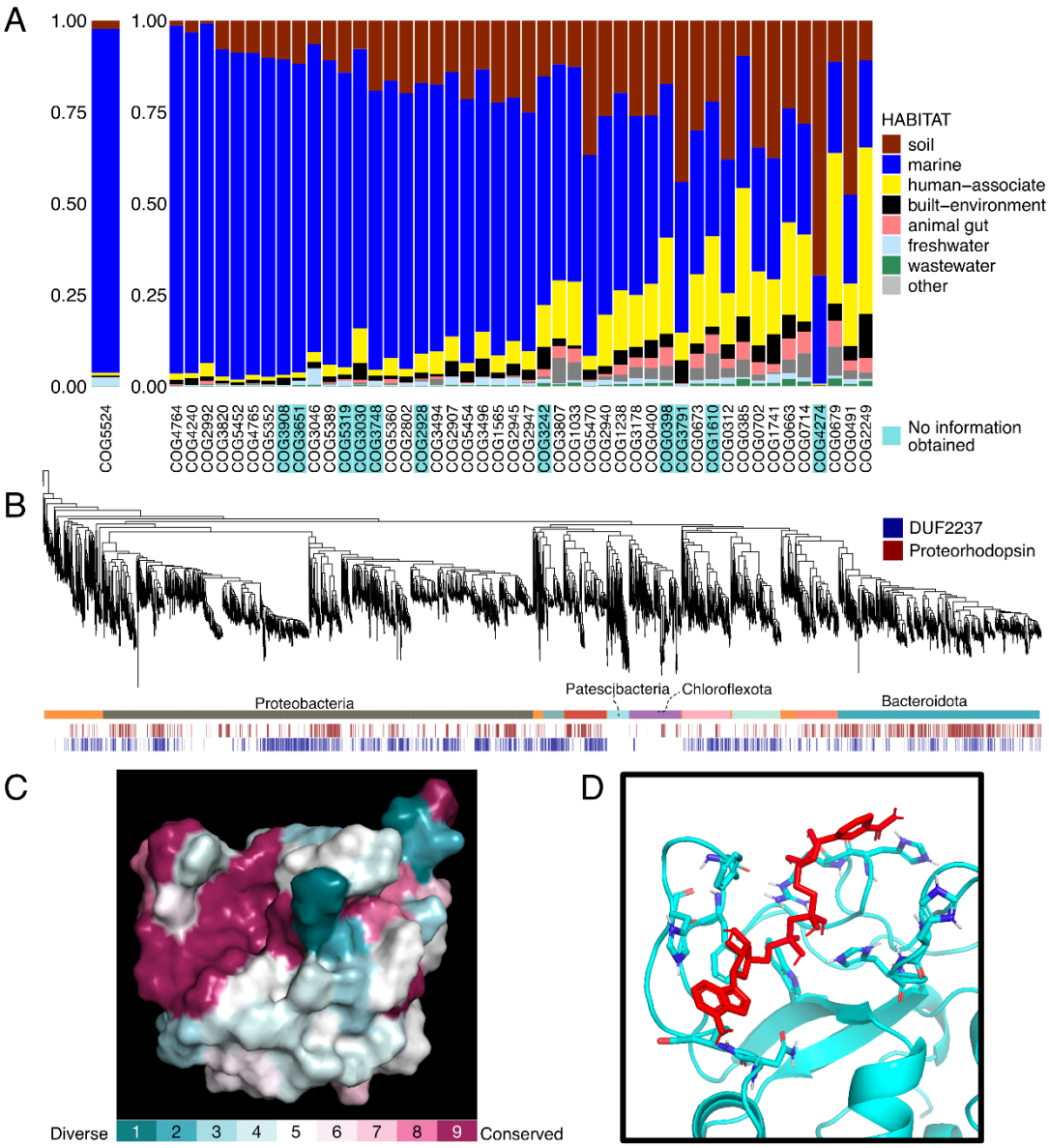

Figure 3. The core unknome distribution and functional analysis of DUF2237-containing protein.

(A) Environmental distribution of bacteriorhodopsin (COG5524) and COG-assigned core unknomes based on GMGC. Bar plot colors indicate habitat, and the COG ID color indicate whether function could be inferred using the bioinformatics approaches. **(B)** Phylogenetic distribution of the DUF2237-containing gene and proteorhodopsin genes on representative bacterial genomes in the OceanDNA MAG catalog. The heatmap below indicates phylum, and the presence of DUF2237-containing gene and proteorhodopsin gene. **(C)** The amino acid conservation of 7,963 DUF2237-containing genes in the Pfam database, mapped onto the predicted structure of the DUF2237

protein from *Ca.* P. ubique HTCC1062. Conservation is color-coded from blue (<10%), to white (50%), to purple (>90%) based on the multiple sequence alignment. **(D)** Top docking pose of *P. ubique* HTCC1062 DUF2237 protein (light blue) and NADH (red).

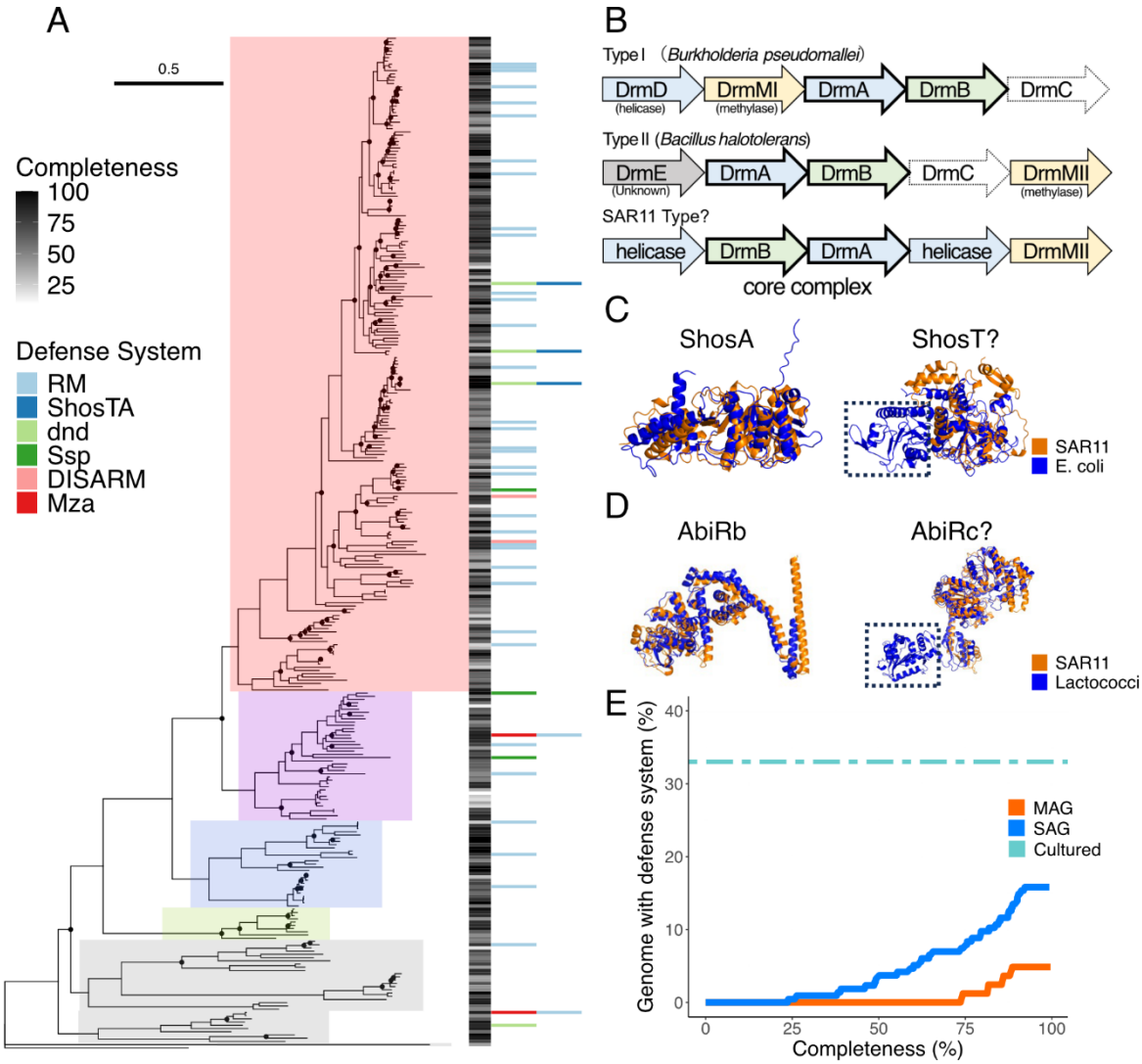

Figure 4. Defense genes identified in SAR11 genomes.
(A) Sporadic distribution of defense systems. Heatmap colors indicate genome completeness and presence of defense systems. **(B)** Operon structure of known DISARM systems and AG-447-N21 coding DISARM-like system. The colors indicate function as follows: helicase (blue), methylase (yellow), DUF1998-containing protein (green), gray (unknown), and PLD domain (white; not required for defense). **(C)** Predicted structure comparison of the known ShosTA proteins from DefenseFinder wiki (blue) and the ShosTA-like proteins in HTCC1002 (orange). Domain loss is highlighted by the dotted box. **(D)** Predicted structure comparison of known AbiRbc proteins from DefenseFinder wiki (blue) and AbiRbc-like proteins in AG-313-A04 (orange). **(E)** The cumulative proportion of genomes with detected viral defense systems is plotted against genome completeness. The cumulative proportion at each completeness value

58 represents the fraction of genomes at or below that completeness in which at least one
59 defense system (from panel A) was detected. Solid blue and red lines show the
60 cumulative proportion for SAGs and MAGs, respectively. The light blue dashed line
61 indicates the detection rate in 24 cultured genomes, serving as a reference.
62