Increasing transposase abundance with ocean depth correlates with a particle-associated lifestyle

Juntao Zhong[1], Thais del Rosario Hernández[1,2], Oleksandr Kyrysyuk[1,3], Benjamin J. Tully[4], Rika E. Anderson[1*]

Author affiliations

[1]Department of Biology, Carleton College, Northfield, MN, USA

[2]Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence, RI, USA

[3]Department of Cancer Immunology and Virology, Dana Farber Cancer Institute, Boston, MA, USA

[4]Marine & Environmental Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA USA

*Corresponding author: randerson@carleton.edu

Importance

Genes can move within and between microbial genomes via mobile genetic elements, which include transposases and transposons. In the oceans, there is a puzzling increase in transposase abundance in microbial genomes as depth increases. To gain insight into this trend, we conducted an extensive analysis of marine microbial metagenomes and metatranscriptomes. We found a significant correlation between transposase abundance and a particle-associated lifestyle among marine microbes. We also observed a link between transposase abundance and genes related to defense mechanisms. These results suggest that as microbes become densely packed into crowded particles, mobile genes are more likely to spread and carry genetic material that provides a competitive advantage in crowded habitats. This may enable deep sea microbes to effectively compete in such environments.

Abstract

Transposases are mobile genetic elements (MGEs) that move within and between genomes, promoting genomic plasticity in microorganisms. In marine microbial communities, the abundance of transposases increases with depth, but the reasons behind this trend remain unclear. Our analysis of metagenomes from the *Tara* Oceans and Malaspina Expeditions suggests that a particle-associated lifestyle is the main covariate for the high occurrence of transposases in the deep ocean. We observed a strong and depth-independent correlation between transposase abundance and the presence of biofilm-associated genes, as well as the prevalence of secretory enzymes. This suggests that mobile genetic elements readily propagate among microbial communities within crowded biofilms. Furthermore, we show that particle association positively correlates with larger genome size, which is in turn is associated with higher transposase abundance. Cassette sequences associated with transposons are enriched with genes related to defense mechanisms, which are more highly expressed in the deep sea. Thus, while transposons spread at the expense of their microbial hosts, they also introduce novel genes and potentially benefit the hosts in helping to compete for limited

41 resources. Overall, our results suggest a new understanding of deep ocean particles as highways for
42 gene sharing among defensively oriented microbial genomes.
43
44 **Introduction**
45 Mobile genetic elements (MGEs) are segments of DNA that facilitate the movement of
46 genetic sequences within and between bacterial and archaeal genomes (1). MGEs generally encode
47 enzymes to mediate this process (1): in transposons, the mediating enzymes are transposases, and
48 the mechanism of transfer results in inverted repeats (2, 3). During migration, some MGEs carry
49 cassette sequences that often contain functional genes, including those for antibiotic resistance (4),
50 metal uptake (5, 6), and regulatory genes influencing gene expression in host cells (7, 8).
51 Transposases are the most abundant and ubiquitous genes in nature (9), although their
52 distribution among taxa is uneven (10, 11). In marine systems, one of the most striking – and as of
53 yet unexplained – trends in microbial genomics is the distinct increase in transposase abundance
54 with depth. Genomic analyses of samples collected at station ALOHA in the North Pacific revealed
55 a substantial increase in transposase abundance from 500 m to their observed maximum at 4000 m
56 (12, 13). Transposases were one of the most overrepresented COG-categories in ALOHA deep
57 waters, accounting for 1.2% of all fosmid sequences from 4000 m (12, 14). Similarly, a study of
58 hydrothermal chimneys demonstrated a high abundance of transposases in biofilms, comprising 8%
59 of all metagenomic reads, which is ten times higher than observed in metagenomes from in other
60 habitats (15). Based on these observations, our analyses focused on three central questions: 1) why
61 does transposase abundance increase with depth in marine systems, 2) are transposases selfish genes
62 with neutral or deleterious effects on host genomes, and 3) do transposases provide useful functions
63 to the microbial hosts that harbor them?

64 To better understand the high transposase abundance in the deep sea and to gain insights
65 into the role of MGEs in marine microbial communities, we analyzed 138 microbial metagenomes
66 and 152 microbial metatranscriptomes from the *Tara* Oceans Expedition (16, 17). The *Tara* Oceans
67 samples spanned depths from 5m to 1000 m. In order to represent bathypelagic microbial
68 communities, we also incorporated 58 metagenomes from the 2010 Malaspina Expedition (18),
69 collected between 2400m and 4000m. To explore the role of transposases at a genome-resolved
70 level, we analyzed 1402 metagenome assembled genomes (MAGs) previously generated from the
71 *Tara* Oceans (19) and Malaspina (18) metagenomes. Previous studies have shown that deep-sea
72 prokaryotes have a predominantly particle-associated lifestyle (20). Here, we show that the increasing
73 abundance of transposases with depth in ocean microbial communities is associated with a shift
74 towards a particle-associated lifestyle. We identified taxonomy, particle association, and genome size
75 as key covariates with transposase abundance in MAGs. Additionally, we observed a high abundance
76 of ORFs in the functional category "defense mechanisms" among the open reading frames (ORFs)
77 in cassette sequences associated with transposases, suggesting that transposons introduce beneficial
78 genes to their microbial hosts inhabiting highly competitive habitats.
79
80 **Results and Discussion**
81 Previous studies have revealed that transposase abundance in microbial genomes is
82 associated with increased ocean depth, lower dissolved oxygen (DO) concentrations, and a particle-
83 associated lifestyle (as opposed to a planktonic lifestyle) (10, 13, 21). However, the reason and nature
84 of these correlations has not been further explored. To confirm these findings on a broader scale,
85 we screened for significant covariates for transposase abundance in the *Tara* Oceans and Malaspina
86 metagenomes. We plotted the transposase abundance against depth and DO. In our metagenome
87 analysis, transposase abundance was defined as the proportion of reads mapped back to all

88 transposase ORFs (the database used to identify putative transposase ORFs is shown in **Table S1**).
89 We confirmed that transposase abundance increases steadily as depth increases, despite large
90 variations across samples (**Fig. 1**). In every ocean, the transposase abundances of the deep water
91 samples were higher than those of the shallow water samples (**Table S2**). Here, "deep water" was
92 defined as mesopelagic (depth 250-1000 m) and bathypelagic (2500-4000 m) waters, and "shallow
93 water" was defined as surface (<10 m) and deep chlorophyll maximum (DCM, depth 17-120 m)
94 waters.

95 The depth of a sample (e.g., surface, DCM) was a significant predictor for transposase
96 abundance (Table S3). Adding temperature, however, did not significantly improve the prediction of
97 transposase abundance in a sample (ANOVA *F*-test p =0.48, **Table S3** and **Fig. S1A**). Similarly,
98 adding dissolved oxygen also did not improve the prediction of transposase abundance (ANOVA *F*-
99 test p = 0.09, **Table S3**). This was partially due to the inconsistent relationship between DO and
100 transposase abundance across depths. In the surface and bathypelagic samples, DO positively
101 correlated with transposase abundance, but this correlation was reversed in DCM and mesopelagic
102 samples (**Fig. S1B**).
103

104 **Transposase abundance is positively correlated with a particle-associated lifestyle in all**
105 **depths.**

106 To further investigate the relationship between transposase abundance and particle-
107 associated lifestyles, we examined the correlation between transposase abundance and the
108 percentage of predicted secretory carbohydrate-active enzymes (CAZymes) and peptidases among all
109 CAZyme and peptidase ORFs (see Methods). The percentage of secretory enzymes among all
110 CAZymes and peptidases has previously been applied to quantify the degree to which microbial
111 communities rely on a particle-associated lifestyle (20). Microorganisms rely on extracellular enzymes

112 to degrade large particulate organic carbon into compounds of smaller molecular weight to
113 incorporate into the cell (20, 22) and CAZymes and peptidases are key enzymes for carbohydrate
114 and protein degradation, respectively (23, 24). To account for differences in biological processes at
115 different depths, we normalized the gene and transcript abundance of predicted secretory CAZymes
116 and peptidases by those of all CAZyme and peptidase ORFs in metatranscriptomic and
117 metagenomic samples (the bathypelagic zone was excluded for all transcript analysis, because no
118 metatranscriptomes were available for Malaspina samples).The large number of CAZymes and
119 peptidases in the database make it challenging to identify their exact binding substrates, which might
120 negatively impact our ability to quantify the particle association of a microbial community. However,
121 microorganisms often use biofilms to attach to the surfaces of particles (22), and thus we paired the
122 CAZyme/peptidase analysis with a secondary analysis quantifying the abundance of biofilm-
123 associated genes using a manually curated database of biofilm-associated ORFs (see Table S3 for the
124 database of biofilm-associated genes).

125 From the *Tara* Oceans metagenomes, we identified 52,890 secretory CAZymes out of
126 421,080 CAZyme ORFs, and 179,932 secretory peptidases out of 1,294,210 peptidase ORFs; from
127 the Malaspina metagenomes, we identified 7,002 secretory CAZymes in 40,415 CAZyme ORFs, and
128 18,014 secretory peptidases out of 67,617 peptidases (see Methods). We observed an increasing
129 percentage of secretory CAZymes and peptidases with depth in both metagenomic and
130 metatranscriptomic samples (**Fig. S2A-D**), which was consistent with previous findings (20). The
131 biofilm-associated ORFs showed a similar increase in gene and transcript abundance with depth
132 (**Fig. S2E, F**). The increase in the percentage of secretory CAZymes and peptidases with depth
133 indicates increased extracellular enzyme activity, suggesting a shift toward a particle-associated
134 lifestyle toward the deep ocean.

135 Marine particles are thought to be hotspots for horizontal gene transfer and transposase
136 propagation (24), and the increased importance of a particle-associated lifestyle with depth in the
137 ocean offers a promising explanation for the elevated transposase abundance in the deep ocean. In
138 both metagenomic and metatranscriptomic samples, the abundance of transposases and the
139 percentage of secretory CAZymes and peptidases were strongly correlated (**Fig. 2**). Given a model
140 with only depth as predictor of the transposase abundance in a sample, the addition of secretory
141 CAZymes and peptidases each significantly improved the accuracy of prediction (both ANOVA *F*-
142 test $p = 3x10^{-7}$, Table S3). In addition, the percentage of secretory CAZymes and peptidases each
143 had consistent positive correlations with transposase gene and transcript abundance in each depth
144 (**Fig. S3**). The depth independent association between particle-associated lifestyle and transposase
145 abundance was also supported by persistent correlations between the abundance of transposase
146 genes/transcripts with those of biofilm-associated ORFs in each depth (**Fig. S4**). Correlations in
147 metagenomes and metatranscriptomes suggested that transposases are more abundant and more
148 frequently transcribed in microbial communities with a greater reliance on a particle-associated
149 lifestyle. After observing such trends at the community level, we sought to determine if these trends
150 hold true at a genome-resolved level to support our hypothesis that the particle-associated lifestyle is
151 a main driver for the high transposase abundance in the deep ocean.
152
153 **Genome size and taxonomy are key factors to transposase abundance in populations.**
154 To identify factors affecting transposase abundance at a genome-resolved level, we analyzed
155 1147 MAGs recovered from the *Tara* Oceans metagenomes (19) and 255 MAGs from the Malaspina
156 metagenomes (18). All MAGs in our analysis had percent completeness >70%, redundancy <10%,
157 and contained <5% of eukaryotic sequences (see Methods). The transposase abundance in MAGs
158 was quantified by the percentage of transposase ORFs among all ORFs in the MAG (abbreviated as

159 *%-transposase*). Similar to metagenomes, MAGs from deep waters had higher transposase abundance
160 than MAGs from shallow waters. Specifically, MAGs from the mesopelagic and bathypelagic zones
161 had a mean %-transposase 6 times (95% CI: 5.38 to 6.57) that of MAGs from the surface and DCM.
162 Given the enrichment of transposase ORFs in deep water MAGs, we ask 3 questions: 1)
163 whether the correlation between transposase abundance and the particle-associated lifestyle persists
164 in MAGs; 2) since larger genomes have an elevated rate of HGT , whether genome size correlates to
165 transposase abundance in MAGs; and 3) whether specific taxa encode more transposases than
166 others, and thus whether community composition plays a role in determining transposase
167 abundance.
168 As with the metagenomes, we quantified the degree to which individual genomes were
169 particle-associated by calculating the percentage of secretory CAZymes and peptidases out of all
170 CAZyme and peptidase ORFs (a ratio of counts) in a MAG. This method has been verified in
171 metagenomes separated through serial filtering techniques – MAGs assembled from samples with
172 larger filter sizes contain a higher percentage of secretory CAZymes and peptidases in the genome
173 compared to those filtered with smaller size fractions (25). We also independently validated this
174 method by comparing the percentage of secretory CAZyme and peptidase ORFs in taxa generally
175 known to be particle-associated or planktonic (**Fig. S5**). Taxonomy was analyzed at a class level, and
176 the "complete genome size" was estimated using the genome length of a MAG divided by its
177 estimated percent completeness. Although MAG assembly might bias against MGEs (26) and the
178 completeness of MAGs might be underestimated for rare taxonomy groups (27, 28), we can still gain
179 valuable insights by analyzing the distribution of transposase ORFs in MAGs of various genome
180 sizes. To address the first question, we observed a positive correlation between the percentage of
181 transposase ORFs and the percentage of secretory CAZymes and peptidases in MAGs (Spearman ρ

182 = 0.208 and ϱ = 0.210, respectively; both $p < 10^{-14}$), thus confirming that the correlation between
183 transposases and particle association persists on a genome-by-genome basis.
184 To examine the relationship between genome size and transposase abundance, we compared
185 the estimated genome size of each MAG with the percentage of ORFs characterized as transposases
186 within each MAG. In every depth, we observed a positive correlation between the estimated genome
187 size and the percentage of transposase ORFs within a MAG (Spearman ϱ was 0.56, 0.50, 0.46, 0.60
188 for surface, DCM, mesopelagic, and bathypelagic samples, respectively. All $p < 10^{-20}$), confirming
189 previous speculations that larger genomes were linked to high transposase abundance (10, 13). While
190 deep water genomes were bigger (**Fig. 3A**), the particle-associated lifestyle correlated with greater
191 complete genome size in every depth (**Fig. 3B**). Previous work has shown that large genomes are
192 associated with a high fraction of transposase ORFs in isolated NCBI microbial genomes (11, 29,
193 30), and in a study of metagenomes from the Baltic Sea (10). The correlation between transposase
194 abundance and genome size was previously attributed to a higher frequency of HGT in larger
195 genomes (23, 29). The correlation between the particle-associated lifestyle and genome size has also
196 been previously observed and may be linked to the expanded metabolic versatility of particle-
197 associated microbial populations (25). Thus, the particle-associated lifestyle might facilitate
198 transposase propagation by reducing the distance between microorganisms, and may lead to larger
199 genome size through increased rates of HGT. It is worth noting that the transposase abundance in
200 MAGs peaked when the complete genome size reached 5 Mbp, but it stabilized or declined beyond
201 the peak (**Fig. 3C**). A previous study has also reported a decreased proportion of transposase ORFs
202 after bacterial genomes surpass 6 Mbp in size (11). However, most (88.9%) MAGs recovered from
203 the *Tara* Oceans and Malaspina metagenomes were under 5 Mbp, so the association between large
204 genomes and high transposase abundance generally held true.

205 To address the third question, we found that taxonomic class was also a key predictor of
206 transposase abundance in MAGs; information on the taxonomy and depth of MAGs together
207 explained 53% of the variance in transposase abundance (**Table 1**). MAGs of the taxonomic classes
208 *Alpha-*, *Beta-*, *Gammaproteobacteria*, and *Actinobacteria* were enriched in transposases compared to other
209 MAGs (for each class with ≥10 MAGs, The %-transposase of MAGs in that class were compared
210 against those of all other MAGs; Wilcox-test p cutoff: 0.05). MAGs of *Flavobacteria*, *Acidimicrobidae*,
211 *novelClass_E*, and *SAR202-2* had low transposase abundance. Taxa that were enriched/low in
212 transposases mostly matched with previous studies (10, 11), except that *Actinobacteria* had previously
213 been reported to be low in transposase abundance.
214 The link between taxonomy and transposase abundance could partially be attributed to
215 depth. MAGs of transposase-enriched classes were more abundant in deep waters, and MAGs from
216 low-transposase classes were more abundant in shallow waters (2 χ^2-tests, both $p < 1\text{x}10^{-6}$; **Fig. S6**).
217 Thus, although community composition is a key covariate to transposase abundance, the
218 distribution of taxa is nonetheless linked to the depth of a microbial community.
219
220 **Relaxed selection does not explain transposase abundance in genomes.**
221 It is possible that transposases accumulate in deep ocean microbial populations as a result of
222 genetic drift, because those populations tend to be smaller in size (31) and experience slow growth
223 rates (32). To examine this possibility, we tested whether the strength of selection experienced by
224 populations correlated with their transposase abundance. For each ORF in each MAG, we
225 calculated its pN/pS ratio in each sample from its designated depth (pN/pS was only calculated for
226 ORFs with ≥20x coverage). pN/pS represents the proportion of nonsynonymous mutations to the
227 proportion of synonymous mutations, and characterizes selection at the level of the population, in
228 contrast to dN/dS, which characterizes selection between individual species (33–35).

229 We observed that MAGs from the mesopelagic zone had a higher median *pN/pS* than
230 MAGs from the surface and DCM, and MAGs from the bathypelagic zone had a higher median
231 *pN/pS* than MAGs from the mesopelagic zone (both $p < 5 \times 10^{-6}$, **Fig. 4A**). The high median *pN/pS*
232 of populations from the deeper layers suggested relaxed selection pressure relative to populations
233 from shallower layers, which was in agreement with previous results (13). The general explanation
234 for such a trend is a slower replication rate, which leads to smaller effective population sizes and a
235 reduced selection effect in the deep ocean (12, 13). However, further work is needed to substantiate
236 this hypothesis.
237 Since transposases and deep-sea ORFs were both under relaxed selection, we tested whether
238 high transposase abundance correlated with relaxed selection in a population. However, we did not
239 observe a correlation between the median *pN/pS* of MAGs and their transposase abundance
240 (Spearman $\rho = -0.075$, $p = 0.08$). Furthermore, the relationship between the selective pressure
241 experienced by a population and its transposase abundance was inconsistent across different depths
242 (**Fig. 4B**). It has been previously suggested that small populations would experience relaxed
243 selection against mobile genetic elements (36, 37). However, our analyses show that although relaxed
244 selection and transposase enrichment co-occur in deep oceans, we did not observe this trend on a
245 genome-by-genome basis, suggesting that there is not a direct connection between transposase
246 abundance and relaxed selection at a genome-resolved scale, and thus within individual microbial
247 populations.
248
249 **Transposons/integrons carry a high proportion of ORFs related to defense mechanisms**
250 One important question regarding high transposase abundance in the deep sea is whether
251 these transposases are neutral or deleterious, or whether they perform important functions in the
252 genomes encoding them. In some cases, transposases are deleterious and proliferate as a result of

253 genetic drift (38), but transposases can also benefit the host by introducing functional and regulatory
254 genes (7, 24, 39, 40).

255 Transposases mediate the migration and integration of cassette sequences into host
256 genomes; a transposon or integron is the combination of a transposase and its cassettes. Thus,
257 cassette sequences are functionally similar to auxiliary metabolic genes (AMGs) (41), in that they are
258 carried by a selfish MGE (or viruses in the case of AMGs) to increase the fitness of the host and
259 therefore increase the fitness of the MGE.

260 To determine whether transposons in marine habitats carry advantageous genes, we began
261 by querying the functional categories of cassette sequences. The software package Integron Finder
262 (11) was used to locate cassettes on contigs; the program searches for the two palindromic flanking
263 sites necessary for transposase-mediated recombination as well as a nearby transposase/integrase.
264 We treated transposons and integrons as equivalent, due to the sequence and function similarity
265 between integrases and transposases (11). We found 8,519 cassette ORFs from the co-assembled
266 Malaspina (18) and *Tara* Oceans (17, 19) metagenomes (see Methods). Only 27% of the cassette
267 sequences were assigned with clusters of orthologous genes (COG) annotations, which was a lower
268 annotation rate compared to other ORFs in the *Tara* Oceans (55-60%) and Malaspina (68%)
269 metagenomes.

270 Compared to the rest of the metagenome, cassettes were enriched in ORFs of the COG
271 categories "replication, recombination, and repair", "defense mechanisms", and "mobilome:
272 prophages, transposons" (**Fig. 5A**). Transposons/integrons are expected to be enriched in ORFs
273 related to the mobilome and replication/recombination. The prevalence of defense mechanism
274 genes in cassettes was noteworthy: defense mechanisms accounted for 14.3% of known function
275 calls in cassettes, but only accounted for 2.47% of known function calls in non-cassette ORFs
276 (*known* was defined as all functional groups except "function unknown" and "general function

277 prediction"). Out of the 276 defense mechanism cassettes, 123 of them encoded toxin/antitoxin
278 genes. Since transposases are more pervasive on particles, toxin and defense genes they carry would
279 be useful in competition and protection on crowded particles (42–45). Moreover, the percentage of
280 secretory CAZymes and peptidases in metagenomes both correlated with the abundance of defense
281 mechanism ORFs (**Fig. 5B**). Similarly, we also found a strong correlation between the gene
282 abundance of defense mechanism ORFs and that of biofilm-associated ORFs (**Fig. S7**). Thus, if
283 defense mechanism genes in the deep ocean were more highly expressed, this would substantiate the
284 hypothesis that deep ocean microorganisms benefit from novel genes introduced by
285 integrons/transposons.
286
287 **ORFs related to defense mechanisms, secretory CAZymes, and biofilm-associated genes**
288 **have higher expression in the deep sea**
289 To determine whether defense mechanism genes are more highly expressed and how this
290 correlates to expression of particle-associated genes, we calculated their RNA/DNA ratios (defined
291 as the transcript abundance of a target gene divided by its gene abundance) of secretory CAZyme,
292 secretory peptidase, defense mechanism, and transposase ORFs in each *Tara* Oceans sample with a
293 paired metagenome and metatranscriptome (n = 94). Malaspina samples were excluded from this
294 analysis, because no metatranscriptomes were sequenced.
295 The ORFs related to secretory CAZymes and defense mechanisms had greater RNA/DNA
296 ratios in the mesopelagic zone than in the surface and DCM (**Fig. 6**), demonstrating a greater need
297 for these genes as deep ocean microbial communities switch toward a particle-associated lifestyle.
298 The RNA/DNA ratio of biofilm-associated ORFs also supported this switch to particle-associated
299 lifestyle (**Fig. S8**). In contrast, although transposases were more abundant in the mesopelagic zone
300 (**Fig. S9**), their RNA/DNA ratios were similar across all depths (**Fig. 6**). A possible explanation is

301 that although transposases are not upregulated, they sometimes carry beneficial cassette genes, which
302 are increasingly expressed in deep-sea microbial genomes.
303
304 **Conclusion**
305 Our analysis provides insights into the factors driving the increasing abundance of
306 transposases with ocean depth, highlighting the interplay between the particle-associated lifestyle,
307 expanded genome size, and selection for defense mechanism genes (**Fig. 7**). We hypothesize that as
308 microbial communities shift from a predominantly planktonic to particle-associated lifestyle from
309 the surface to the bathypelagic zone, microbial communities become more densely packed on
310 particles, leading to rampant transposase spread and more intense resource competition.
311 Additionally, particle-associated microorganisms tend to have larger genomes, which are associated
312 with high transposase abundance. These high abundances of mobile genetic elements actively shape
313 the ecology of particle-associated microbial communities by introducing novel genes, with cassette
314 sequences being particularly enriched in defense mechanism genes. These genes are highly expressed
315 in the deep ocean, offering competitive advantages in the competitive particle-associated
316 environment (42, 44).
317 When interpreting these results, it is important to consider certain caveats. Our analysis
318 focused on pelagic ocean water samples to minimize potential confounding variables, and thus these
319 observed trends may not apply to regions with unique characteristics, such as Arctic sea ice (47) and
320 deep-sea hydrothermal vents (15). Moreover, as our results are based on correlative analyses, further
321 experimental evidence is needed to establish causation and to identify mechanistic relationships
322 between these variables.
323 Understanding these trends in mobile genetic element abundance is important for
324 understanding gene flow, competition, and other ecological characteristics of the dark ocean, one of

325 the largest habitats on Earth. Our results show a strong association between transposase abundance
326 and a particle-associated lifestyle, suggesting that transposons may enable microbial lineages to
327 compete in crowded biofilm-associated habitats. We did not find any correlation between the
328 strength of selection and transposase abundance at the genome level, indicating that relaxed
329 selection pressure is not the primary driver for the high transposase abundance in the deep ocean.
330 Overall, our results suggest an emerging understanding of the ocean as a stratified system in which
331 the deep ocean acts as a gene sharing highway, fostering networks of gene exchange, particularly on
332 particles. This results in large genomes with many defense-oriented genes in the deep-sea microbial
333 communities, contrasting with the streamlined and specialized genomes that dominate the surface
334 oceans. Future experimental studies should establish mechanistic connections between transposase
335 gene cassette contents and microbial activity, particularly in planktonic and particle-associated
336 communities.
337
338 **Materials and Methods**
339 **Analysis of the Ocean Microbial Reference Catalog v2 (OM-RGC.v2) from the *Tara* Oceans**
340 **Project.** OM-RGC.v2 contained 47 million non-redundant open reading frames (ORFs) (16, 17),
341 and the coverage of each ORF in every metagenomic (n = 138) and metatranscriptomic (n = 152)
342 sample (website) (16). Samples from the Arctic Ocean were excluded. We identified transposases
343 and biofilm-associated ORFs in OM-RGC.v2 through TBLASTN (e-value < 10^{-5}). The transposase
344 database contained "transposase" and "integrase" genes from the Pfam database (48) (Table S1). We
345 identified ORFs of the "defense mechanisms" category through COG annotations in OM-RGC.v2.
346 Finally, the coverage data in OM-RGC.v2 was used to determine the relative abundance of a target
347 gene in a sample.
348

349 **Acquisition of global ocean metagenomes.** Metagenomic reads of the *Tara* Oceans Project and
350 Malaspina Expedition were downloaded from the European Nucleotide Archive (ENA) under study
351 accessions PRJEB402 and PRJEB44456, respectively. Methods for sample collection and Illumina
352 sequencing were described by Sunagawa *et al.* (49) for the *Tara* Oceans Project and Acinas *et al.* (18)
353 for the Malaspina Expedition. For the assembled contigs used as references for mapping of the *Tara*
354 Oceans metagenomes, we used 10 existing co-assembled metagenomes from the 10 oceanic
355 provinces (19). For the reference contigs of the Malaspina metagenomes, we downloaded the co-
356 assembled metagenome of all 58 samples (18). Samples from the *Tara* Oceans were filtered through
357 1.6- or 3-µm filters and collected on 0.22-µm filters (49), whereas samples from the Malaspina
358 expedition were collected from the 0.2 – 0.8 µm and 0.8 – 5 µm size fractions (18). We used
359 EukDetect (50) to confirm that samples collected with a larger size fraction did not have a higher
360 contribution from eukaryotic reads (see Supplementary Text).
361
362 **Mapping and calculating *pN/pS* ratios.** To assess the strength of selection of target ORFs, we
363 calculated the *pN/pS* ratio (the proportion of nonsynonymous mutations over the proportion of
364 synonymous mutations) for all ORFs. To do this, we mapped the metagenomic reads against the
365 assembled contigs for each metagenome. We mapped the raw reads of each of the *Tara* Oceans
366 metagenomes to the co-assembly from their corresponding provinces, and all raw reads of each of
367 the Malaspina metagenomes to one co-assembled metagenome using Bowtie2 (51) (v2.2.9; paired-
368 end alignment with default parameters). We calculated *pN/pS* using anvi'o (52) with the script "anvi-
369 script-calculate-pn-ps-ratio" with the "--min-coverage" flag set to 20. The *pN/pS* ratios were
370 calculated on a sample-per-sample basis, so an ORF from a co-assembled metagenome might have
371 multiple *pN/pS* ratios from different samples.
372

373 **Prediction of secretory CAZyme and peptidase ORFs.** We adapted the methods employed in
374 Zhao et al (20) to assess whether microbial communities were largely planktonic or particle-
375 associated by quantifying the relative abundance of secretory CAZymes and peptidases. Predicted
376 CAZymes and peptidases were annotated using DIAMOND (2.0.15) (53) BLASTp (e-value $<10^{-10}$)
377 to search against the dbCAN (54) and MEROPS (55) databases, respectively. SignalP (5.0) (56) was
378 used to identify signal peptides. We used the Gram-positive mode for ORFs affiliated with
379 *Actinobacteria* and *Firmicutes*, Gram-negative mode for other bacterial phylum, and Archaea mode for
380 ORFs affiliated with the *Archaea* domain. Only ORFs in the *Bacteria* or *Archaea* domains were
381 included in the analysis. For MAGs, counts of secretory CAZymes and peptidases were used instead
382 of abundance; validation based on predicted lifestyles of known taxa confirmed that this gave an
383 accurate prediction of lifestyle (see Supplementary Text).
384
385 **MAG Selection.** To conduct genome-resolved analyses, we selected 1147 out of 2631 MAGs that
386 had previously been recovered from the Tara Ocean metagenomes (19) and 255 out of 317 MAGs
387 recovered from the Malaspina metagenomes (18). All selected MAGs had <5% of eukaryotic
388 sequences (in base pairs, Tiara *(57)* was used to identify putative eukaryotic contigs), completeness >
389 70%, and redundancy < 10% (both criteria were calculated with CheckM (58)).
390
391 **Determination of MAG's depth and lifestyles.** MAGs that originated from co-assemblies in a
392 specific province in the *Tara* Oceans dataset (*e.g.*, Red Sea), as in Tully *et al.* (19), were used to recruit
393 reads from all *Tara* Oceans samples (n = 180) using Bowtie2 (default parameters). Each province
394 was performed separately. SAM files were converted to BAM files using samtools (59) and used to
395 determine RPKM (reads per kilobase pair MAG per Million pair metagenome) as in Graham et al.
396 (60).

397 To assign the depth origin of a MAG from the *Tara* Oceans metagenomes, we summed up
398 its RPKM in surface, DCM, and mesopelagic samples (3 groups, separately). The layer with the
399 highest sum RPKM was the depth of that MAG. All Malaspina MAGs belonged to the bathypelagic
400 layer.
401
402 **Integron and cassette sequences detection.** We used Integron Finder (v2) (11) to identify
403 cassette sequences from the *Tara* Ocean and Malaspina metagenomes. Integron Finder uses
404 HMMER to locate the integron-integrase *intI*, which is conserved for most integrons. Then, cassette
405 sequences are identified with the near-palindromic flaking regions. anvi'o (52) was used to identify
406 and annotate ORFs on metagenomic contigs using the script 'anvi-run-ncbi-cogs.' If the anvi'o ORF
407 calls from Prodigal (61) were within 100bps (start and stop) of the Integron Finder's cassette calls,
408 the anvi'o COG annotations were used for those cassette ORFs.
409
410 **Statistical analysis**
411 We used the Wilcoxon signed-rank test to compare the differences between means of two numeric
412 variables, and the Spearman's rank correlation was used to determine the association between two
413 numeric variables. Multiple hypothesis *p*-values were adjusted using the Benjamini–Hochberg
414 procedure. All linear regressions had a normal distribution of residuals (Shapiro–Wilk test).
415 Statistical significance was assumed if $p < 0.05$. All statistical analyses were performed using base R
416 packages (v4.1.3).
417
418 **Code and data availability.**
419 All Python, R scripts, code explanation, and raw data for analysis are publicly accessible on GitHub
420 at https://github.com/carleton-spacehogs/transposase-deep-ocean.

Acknowledgments

We would like to thank Dr. Shinichi Sunagawa for kindly providing information about the *Tara* Oceans datasets and the Ocean Microbial Reference Catalog, and Dr. Silvia G. Acinas for providing information about the Malaspina Deep Ocean dataset. Dr. Murat Eren provided help with anvi'o, and Mike Tie provided assistance with server administration. Funding for JZ and OK was provided by grants from the Towsley Endowment at Carleton College, and funding for TdRH was provided by the Summer Science Fellows program at Carleton College.

Competing Interest

The authors declare that they have no financial or non-financial competing interests.

432 **References**

433 1. Frost LS, Leplae R, Summers AO, Toussaint A. 2005. Mobile genetic elements: the agents of
434 open source evolution. Nat Rev Microbiol 3:722–732.
435 2. Transposons. http://www.nature.com/scitable/topicpage/transposons-the-jumping-genes-
436 518. Retrieved 30 November 2021.
437 3. Muñoz-López M, García-Pérez JL. 2010. DNA transposons: nature and applications in
438 genomics. Curr Genomics 11:115–128.
439 4. Collis C M, Hall R M. 1995. Expression of antibiotic resistance genes in the integrated
440 cassettes of integrons. Antimicrob Agents Chemother 39:155–162.
441 5. Rankin DJ, Rocha EPC, Brown SP. 2011. What traits are carried on mobile genetic elements,
442 and why? Heredity 106:1–10.
443 6. Hacker J, Carniel E. 2001. Ecological fitness, genomic islands and bacterial pathogenicity. A
444 Darwinian view of the evolution of microbes. EMBO Rep 2:376–381.
445 7. Jones JM, Grinberg I, Eldar A, Grossman AD. 2021. A mobile genetic element increases
446 bacterial host fitness by manipulating development. Elife 10:e65924.
447 8. Casacuberta E, González J. 2013. The impact of transposable elements in environmental
448 adaptation. Mol Ecol 22:1503–1517.
449 9. Aziz RK, Breitbart M, Edwards RA. 2010. Transposases are the most abundant, most
450 ubiquitous genes in nature. Nucleic Acids Res 38:4207–4217.
451 10. Vigil-Stenman T, Ininbergs K, Bergman B, Ekman M. 2017. High abundance and
452 expression of transposases in bacteria from the Baltic Sea. ISME J 11:2611–2623.
453 11. Cury J, Jové T, Touchon M, Néron B, Rocha EP. 2016. Identification and analysis of
454 integrons and cassette arrays in bacterial genomes. Nucleic Acids Res 44:4539–4550.
455 12. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U, Martinez A,
456 Sullivan MB, Edwards R, Brito BR, Chisholm SW, Karl DM. 2006. Community Genomics
457 Among Stratified Microbial Assemblages in the Ocean's Interior. Science 311:496–503.
458 13. Konstantinidis KT, Braff J, Karl DM, DeLong EF. 2009. Comparative metagenomic
459 analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the
460 North Pacific subtropical gyre. Appl Environ Microbiol 75:5345–5355.
461 14. Worden AZ, Cuvelier ML, Bartlett DH. 2006. In-depth analyses of marine microbial
462 community genomics. Trends Microbiol 14:331–336.

463 15. Brazelton WJ, Baross JA. 2009. Abundant transposases encoded by the metagenome
464 of a hydrothermal chimney biofilm. ISME J 3:1420–1424.
465 16. Salazar G, Paoli L, Alberti A, Huerta-Cepas J, Ruscheweyh H-J, Cuenca M, Field
466 CM, Coelho LP, Cruaud C, Engelen S, Gregory AC, Labadie K, Marec C, Pelletier E, Royo-
467 Llonch M, Roux S, Sánchez P, Uehara H, Zayed AA, Zeller G, Carmichael M, Dimier C,
468 Ferland J, Kandels S, Picheral M, Pisarev S, Poulain J, Tara Oceans Coordinators, Acinas SG,
469 Babin M, Bork P, Bowler C, de Vargas C, Guidi L, Hingamp P, Iudicone D, Karp-Boss L,
470 Karsenti E, Ogata H, Pesant S, Speich S, Sullivan MB, Wincker P, Sunagawa S. 2019. Gene
471 Expression Changes and Community Turnover Differentially Shape the Global Ocean
472 Metatranscriptome. Cell 179:1068–1083.e21.
473 17. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri
474 B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F,
475 Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez
476 G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M,
477 Searson S, Kandels-Lewis S, Null N, Bowler C, de Vargas C, Gorsky G, Grimsley N, Hingamp
478 P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB,
479 Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P, Boss E, Bowler C, Follows
480 M, Karp-Boss L, Krzic U, Reynaud EG, Sardet C, Sieracki M, Velayoudon D. 2015. Structure
481 and function of the global ocean microbiome. Science 348:1261359.
482 18. Acinas SG, Sánchez P, Salazar G, Cornejo-Castillo FM, Sebastián M, Logares R,
483 Royo-Llonch M, Paoli L, Sunagawa S, Hingamp P, Ogata H, Lima-Mendez G, Roux S,
484 González JM, Arrieta JM, Alam IS, Kamau A, Bowler C, Raes J, Pesant S, Bork P, Agustí S,
485 Gojobori T, Vaqué D, Sullivan MB, Pedrós-Alió C, Massana R, Duarte CM, Gasol JM. 2021.
486 Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic
487 microbial communities. Communications Biology 4:1–15.
488 19. Tully BJ, Graham ED, Heidelberg JF. 2018. The reconstruction of 2,631 draft
489 metagenome-assembled genomes from the global oceans. Scientific Data 5:1–8.
490 20. Zhao Z, Baltar F, Herndl GJ. 2020. Linking extracellular enzymes to phylogeny
491 indicates a predominantly particle-associated lifestyle of deep-sea prokaryotes. Sci Adv
492 6:eaaz4354.
493 21. Ganesh S, Parris DJ, DeLong EF, Stewart FJ. 2014. Metagenomic analysis of size-
494 fractionated picoplankton in a marine oxygen minimum zone. ISME J 8:187–211.

495 22. Tuson HH, Weibel DB. 2013. Bacteria-surface interactions. Soft Matter 9:4368–
496 4380.
497 23. Cordero OX, Hogeweg P. 2009. The impact of long-distance horizontal gene
498 transfer on prokaryotic genome size. Proc Natl Acad Sci U S A 106:21748–21753.
499 24. Stewart FJ. 2013. Where the genes flow. Nat Geosci 6:688–690.
500 25. Leu AO, Eppley JM, Burger A, DeLong EF. 2022. Diverse Genomic Traits
501 Differentiate Sinking-Particle-Associated versus Free-Living Microbes throughout the
502 Oligotrophic Open Ocean Water Column. MBio 13:e0156922.
503 26. Meziti A, Rodriguez-R LM, Hatt JK, Peña-Gonzalez A, Levy K, Konstantinidis KT.
504 2021. The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural
505 Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same
506 Fecal Sample. Appl Environ Microbiol 87.
507 27. Garcia SL, Buck M, McMahon KD, Grossart H-P, Eiler A, Warnecke F. 2015.
508 Auxotrophy and intrapopulation complementary in the "interactome" of a cultivated freshwater
509 model community. Mol Ecol 24:4449–4459.
510 28. Rodríguez-Gijón A, Nuy JK, Mehrshad M, Buck M, Schulz F, Woyke T, Garcia SL.
511 2021. A Genomic Perspective Across Earth's Microbiomes Reveals That Genome Size in
512 Archaea and Bacteria Is Linked to Ecosystem Type and Trophic Strategy. Front Microbiol
513 12:761869.
514 29. Iranzo J, Gómez MJ, López de Saro FJ, Manrubia S. 2014. Large-scale genomic
515 analysis suggests a neutral punctuated dynamics of transposable elements in bacterial genomes.
516 PLoS Comput Biol 10:e1003680.
517 30. Touchon M, Rocha EPC. 2007. Causes of Insertion Sequences Abundance in
518 Prokaryotic Genomes. Mol Biol Evol 24:969–981.
519 31. Costello MJ, Chaudhary C. 2017. Marine Biodiversity, Biogeography, Deep-Sea
520 Gradients, and Conservation. Curr Biol 27:R511–R527.
521 32. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM,
522 Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere."
523 Proc Natl Acad Sci U S A 103:12115–12120.
524 33. Simmons SL, Dibartolo G, Denef VJ, Goltsman DSA, Thelen MP, Banfield JF.
525 2008. Population genomic analysis of strain variation in Leptospirillum group II bacteria
526 involved in acid mine drainage formation. PLoS Biol 6:e177.

527 34. McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in
528 Drosophila. Nature 351:652–654.
529 35. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A,
530 Mende DR, Kultima JR, Martin J, Kota K, Sunyaev SR, Weinstock GM, Bork P. 2013.
531 Genomic variation landscape of the human gut microbiome. Nature 493:45–50.
532 36. Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, Holden
533 MTG, Churcher CM, Bentley SD, Mungall KL, Cerdeño-Tárraga AM, Temple L, James K,
534 Harris B, Quail MA, Achtman M, Atkin R, Baker S, Basham D, Bason N, Cherevach I,
535 Chillingworth T, Collins M, Cronin A, Davis P, Doggett J, Feltwell T, Goble A, Hamlin N,
536 Hauser H, Holroyd S, Jagels K, Leather S, Moule S, Norberczak H, O'Neil S, Ormond D, Price
537 C, Rabbinowitsch E, Rutter S, Sanders M, Saunders D, Seeger K, Sharp S, Simmonds M,
538 Skelton J, Squares R, Squares S, Stevens K, Unwin L, Whitehead S, Barrell BG, Maskell DJ.
539 2003. Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella
540 parapertussis and Bordetella bronchiseptica. Nat Genet 35:32–40.
541 37. Moran NA, Plague GR. 2004. Genomic changes following host restriction in
542 bacteria. Curr Opin Genet Dev 14:627–633.
543 38. Escobar-Páramo P, Ghosh S, DiRuggiero J. 2005. Evidence for Genetic Drift in the
544 Diversification of a Geographically Isolated Population of the Hyperthermophilic Archaeon
545 Pyrococcus. Molecular Biology and Evolution https://doi.org/10.1093/molbev/msi227.
546 39. Flynn KJ, Swanson MS. 2014. Integrative conjugative element ICE-βox confers
547 oxidative stress resistance to Legionella pneumophila in vitro and in macrophages. MBio
548 5:e01091–14.
549 40. Sullivan JT, Ronson CW. 1998. Evolution of rhizobia by acquisition of a 500-kb
550 symbiosis island that integrates into a phe-tRNA gene. Proc Natl Acad Sci U S A 95:5145–5149.
551 41. Breitbart M, Bonnain C, Malki K, Sawaya NA. 2018. Phage puppet masters of the
552 marine microbial realm. Nat Microbiol 3:754–766.
553 42. Nadell CD, Drescher K, Foster KR. 2016. Spatial structure, cooperation and
554 competition in biofilms. Nat Rev Microbiol 14:589–600.
555 43. Basler M, Ho BT, Mekalanos JJ. 2013. Tit-for-tat: type VI secretion system
556 counterattack during bacterial cell-cell interactions. Cell 152:884–894.
557 44. Hayes CS, Aoki SK, Low DA. 2010. Bacterial contact-dependent delivery systems.
558 Annu Rev Genet 44:71–90.

559 45. Ho BT, Dong TG, Mekalanos JJ. 2014. A view to a kill: the bacterial type VI
560 secretion system. Cell Host Microbe 15:9–21.
561 46. Mestre M, Ruiz-González C, Logares R, Duarte CM, Gasol JM, Sala MM. 2018.
562 Sinking particles promote vertical connectivity in the ocean microbiome. Proc Natl Acad Sci U
563 S A 115:E6799–E6807.
564 47. Rapp JZ, Sullivan MB, Deming JW. 2021. Divergent Genomic Adaptations in the
565 Microbiomes of Arctic Subzero Sea-Ice and Cryopeg Brines. Front Microbiol 12:701186.
566 48. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL,
567 Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: The protein
568 families database in 2021. Nucleic Acids Res 49:D412–D419.
569 49. Sunagawa S, Acinas SG, Bork P, Bowler C, Tara Oceans Coordinators, Eveillard D,
570 Gorsky G, Guidi L, Iudicone D, Karsenti E, Lombard F, Ogata H, Pesant S, Sullivan MB,
571 Wincker P, de Vargas C. 2020. Tara Oceans: towards global ocean ecosystems biology. Nat Rev
572 Microbiol 18:428–445.
573 50. Lind AL, Pollard KS. 2021. Accurate and sensitive detection of microbial eukaryotes
574 from whole metagenome shotgun sequencing. Microbiome.
575 51. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat
576 Methods 9:357–359.
577 52. Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, Fink I, Pan JN,
578 Yousef M, Fogarty EC, Trigodet F, Watson AR, Esen ÖC, Moore RM, Clayssen Q, Lee MD,
579 Kivenson V, Graham ED, Merrill BD, Karkman A, Blankenberg D, Eppley JM, Sjödin A, Scott
580 JJ, Vázquez-Campos X, McKay LJ, McDaniel EA, Stevens SLR, Anderson RE, Fuessel J,
581 Fernandez-Guerra A, Maignien L, Delmont TO, Willis AD. 2021. Community-led, integrated,
582 reproducible multi-omics with anvi'o. Nat Microbiol 6:3–6.
583 53. Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life
584 scale using DIAMOND. Nat Methods 18:366–368.
585 54. Huang L, Zhang H, Wu P, Entwistle S, Li X, Yohe T, Yi H, Yang Z, Yin Y. 2018.
586 dbCAN-seq: a database of carbohydrate-active enzyme (CAZyme) sequence and annotation.
587 Nucleic Acids Res 46:D516–D521.
588 55. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. 2018. The
589 MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a
590 comparison with peptidases in the PANTHER database. Nucleic Acids Res 46:D624–D632.

591 56. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O,
592 Brunak S, von Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using
593 deep neural networks. Nat Biotechnol 37:420–423.
594 57. Karlicki M, Antonowicz S, Karnkowska A. 2022. Tiara: deep learning-based
595 classification system for eukaryotic sequences. Bioinformatics 38:344–350.
596 58. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM:
597 assessing the quality of microbial genomes recovered from isolates, single cells, and
598 metagenomes. Genome Res 25:1043–1055.
599 59. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
600 Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence
601 Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.
602 60. Graham ED, Heidelberg JF, Tully BJ. 2018. Potential for primary productivity in a
603 globally-distributed bacterial phototroph. ISME J 12:1861–1866.
604 61. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal:
605 prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics
606 11:119.
607

608 **Figure Legends:**

609

610 **Fig. 1**: Samples from the deep ocean have higher transposase abundance. The abundance of
611 transposase ORFs in metagenomic samples from the *Tara* Oceans and Malaspina Expedition,
612 separated by depth. SRF: surface; DCM: deep chlorophyll maximum; MES: mesopelagic zone; BAT:
613 bathypelagic zone. (a), (b) and (c) represent Tukey's honestly significant differences, with each letter
614 representing groups that are statistically significant.

615

616 **Fig. 2:** The gene potential (DNA) and transcript abundance (RNA) of transposases correlates with
617 those of secretory CAZymes and peptidases. **A** The correlation between the abundance of
618 transposases and secretory CAZyme and peptidases in metagenomes. **B** The same correlation in
619 metatranscriptomes. Colors represent different ocean depths. Note that the percent transcript
620 abundances were log-transformed.

621

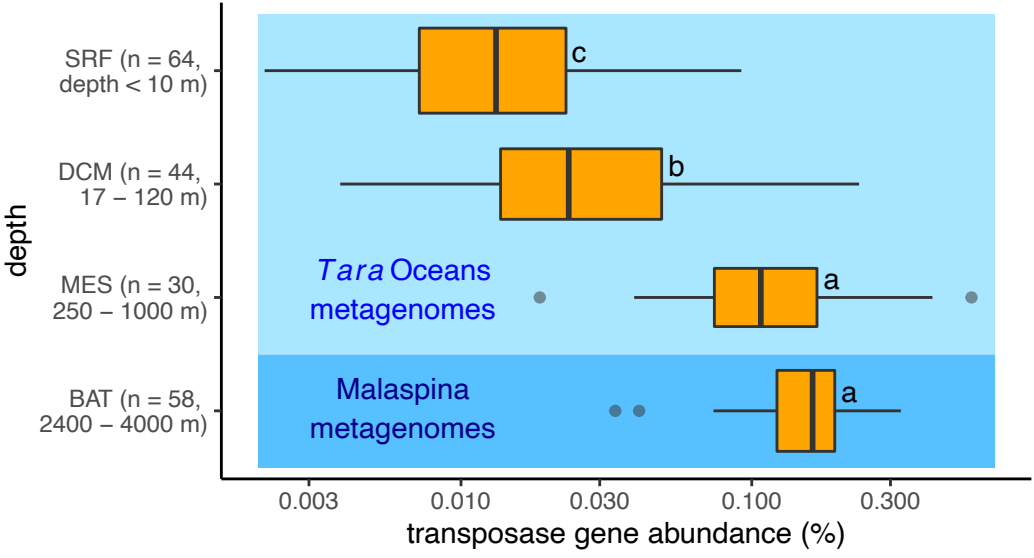
622 **Fig. 3:** Deep-sea and particle-associated microbial populations tend to have larger genomes, which
623 correlate with high transposase abundance in those populations. **A** The estimated complete genome
624 size of MAGs (complete genome size = number of base pairs in a MAG / % completeness),
625 grouped by depth. Letters on boxplots were generated from the Tukey HSD test. **B** Scatterplots
626 showing the correlation between secretory CAZymes/peptidases and the estimated complete
627 genome size of MAGs. **C** The transposase abundance in MAGs, grouped by depth and complete
628 genome size. See Methods for the determination of depth of a MAG.

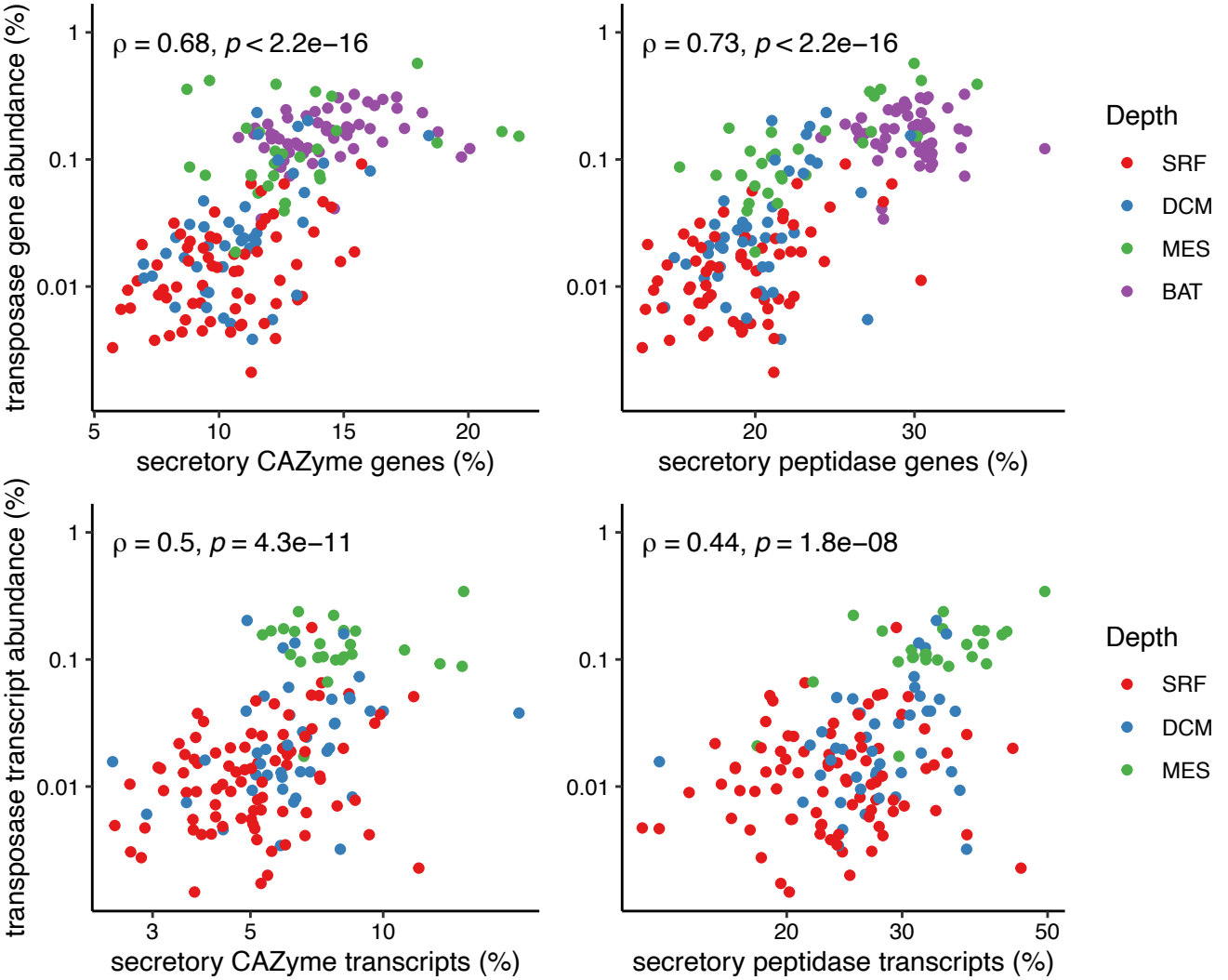
629

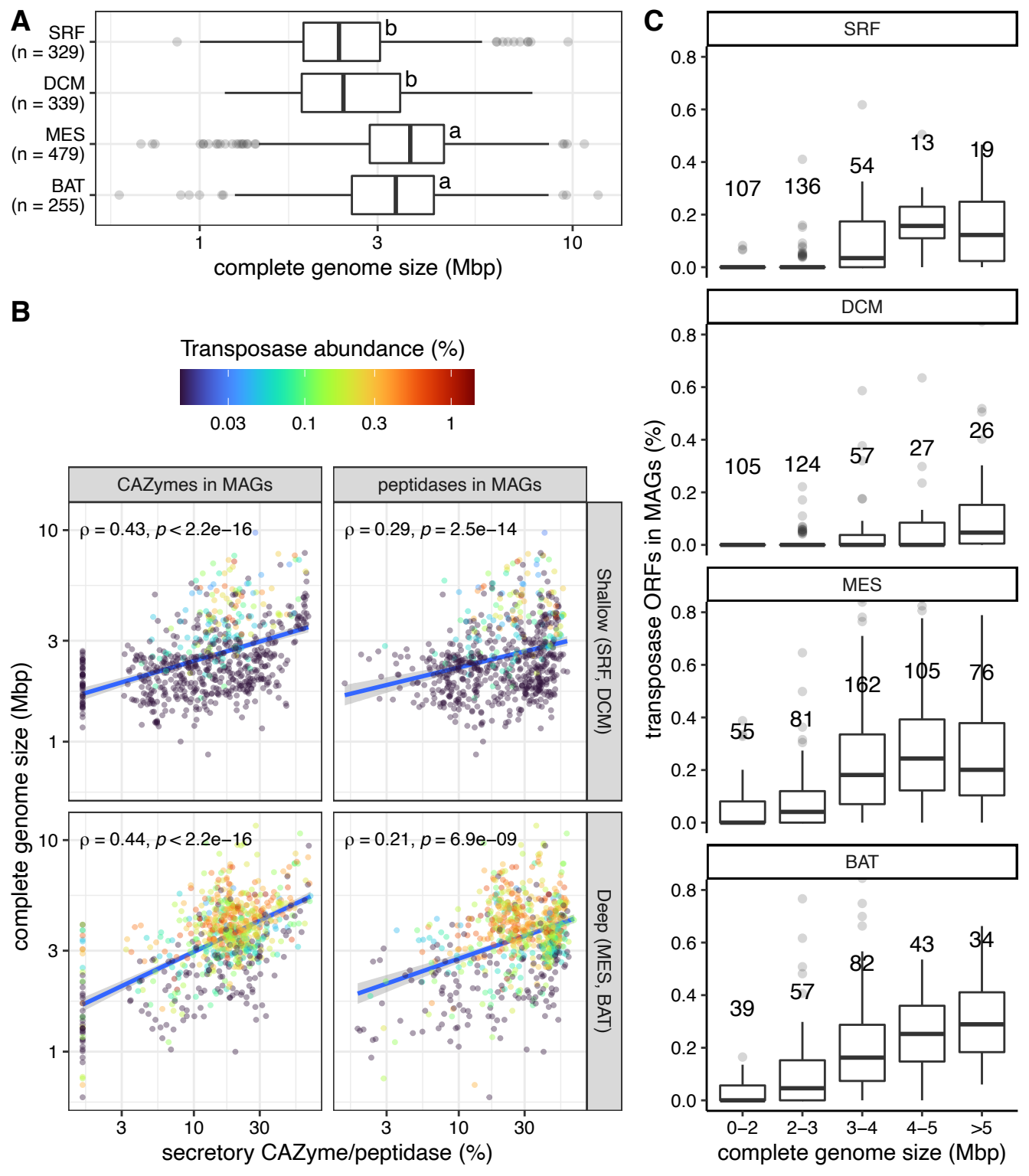
630 **Fig. 4:** Selection pressure and transposase abundance are not correlated on a genome-resolved scale.
631 **A** The median *pN*/*pS* ratio of MAGs, grouped by depth. Since *pN*/*pS* were calculated on a per-

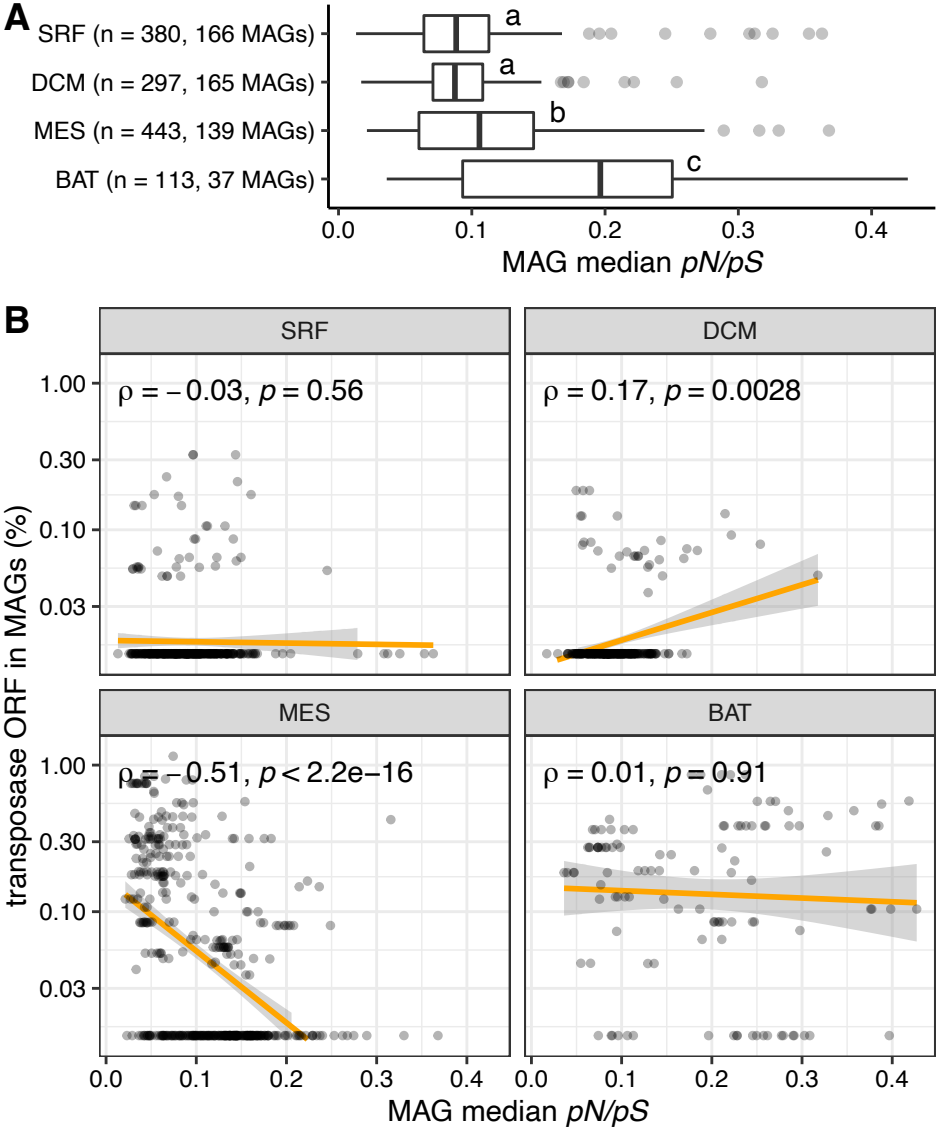
632 sample basis, a MAG might have several median *pN/pS* ratio from multiple samples. **B** The
633 relationship between the selection pressure (median *pN/pS*) and the transposase abundance of a
634 MAG, separated by depth. Only MAGs with ≥ 100 *pN/pS* were computed for median *pN/pS*.
635
636 **Fig. 5:** Cassette sequences include high proportions of defense mechanism genes, which are more
637 abundant in microbial communities that rely heavily on a particle-associated lifestyle. **A** The
638 distribution of COG functional categories in cassette and non-cassette metagenome ORFs. *TM:
639 transport and metabolism. 1951 cassette ORFs with COG function calls were identified from the 10
640 co-assembled *Tara* metagenomes, and 359 from the co-assembled Malaspina metagenome. 5 million
641 ORFs with COG function calls were then sampled from the *Tara* Oceans and Malaspina
642 metagenomes according to the ratio of identified cassettes described above. **B** The correlation
643 between the abundance of secretory CAZyme/peptidase ORFs and defense mechanism ORFs in
644 metagenomes. Samples from different depths are distinguished by different colors. The Spearman's
645 correlation coefficients, ϱ, are shown on top left. **** indicates that $p < 0.0001$.
646
647 **Fig. 6:** Particle-associated and defense mechanism ORFs are more highly expressed in deeper
648 waters, but transposases are not. Log-transformed RNA/DNA ratios of target genes in each sample,
649 separated by depth. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
650
651 **Fig. 7:** Linking a particle-associated lifestyle to the high abundance of transposons in the deep
652 ocean.
653
654 **Table 1.** Multiple Stepwise Regression for log-transformed transposase abundance (%) of 1402
655 MAGs. An ANOVA test was performed on each newly added covariate.

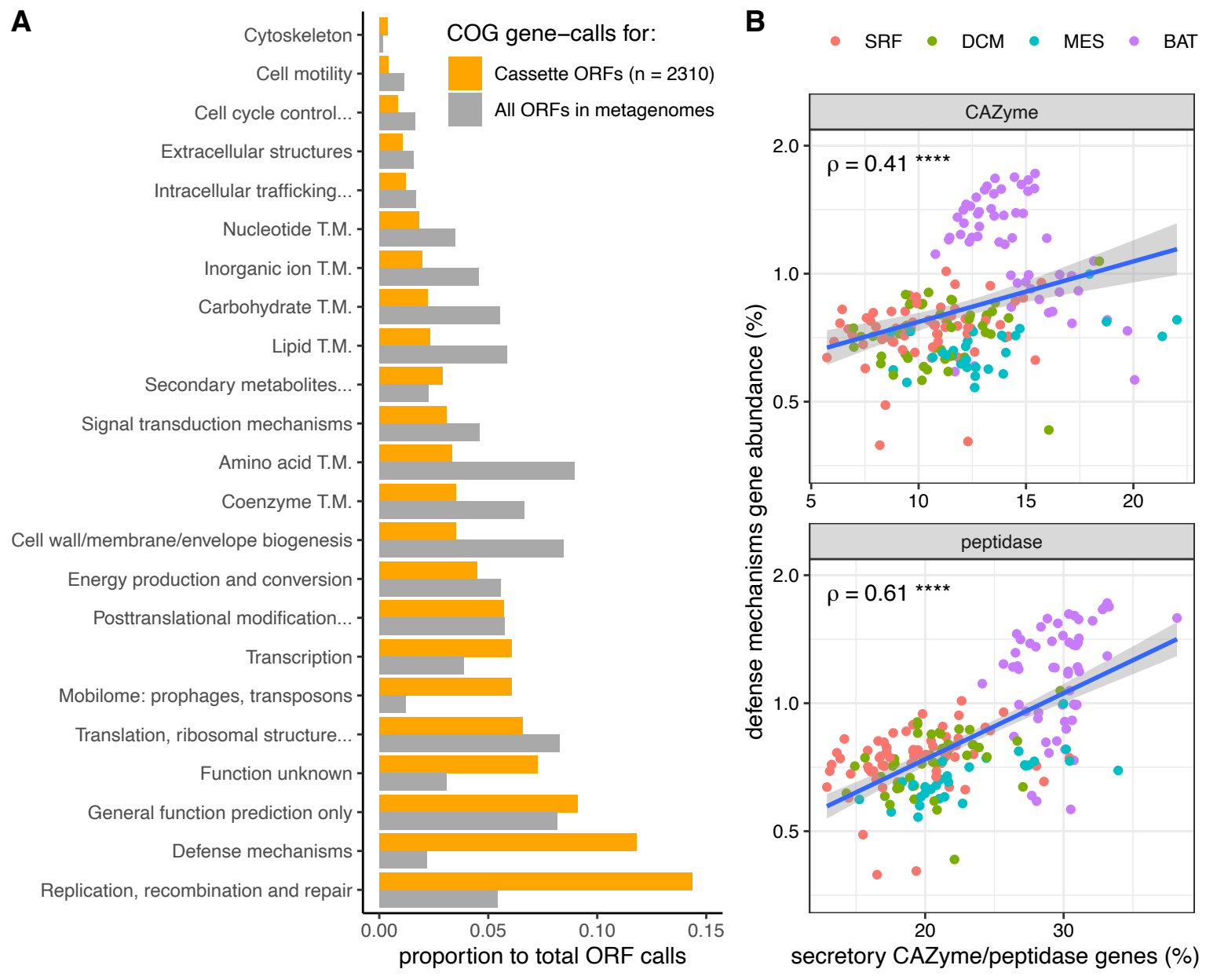
Explanatory Variable(s)	*p* for *F*-test	Cumulative R^2
depth	$< 10^{-10}$	0.34
depth + secretory CAZyme (%)	$< 10^{-10}$	0.37
depth + secretory peptidase (%)	$< 10^{-10}$	0.37
depth + genome size (Mbp)	$< 10^{-10}$	0.52
depth + taxon (class)	$< 10^{-10}$	0.53
depth + taxon + genome size + secretory CAZyme + secretory peptidase	NA	0.62

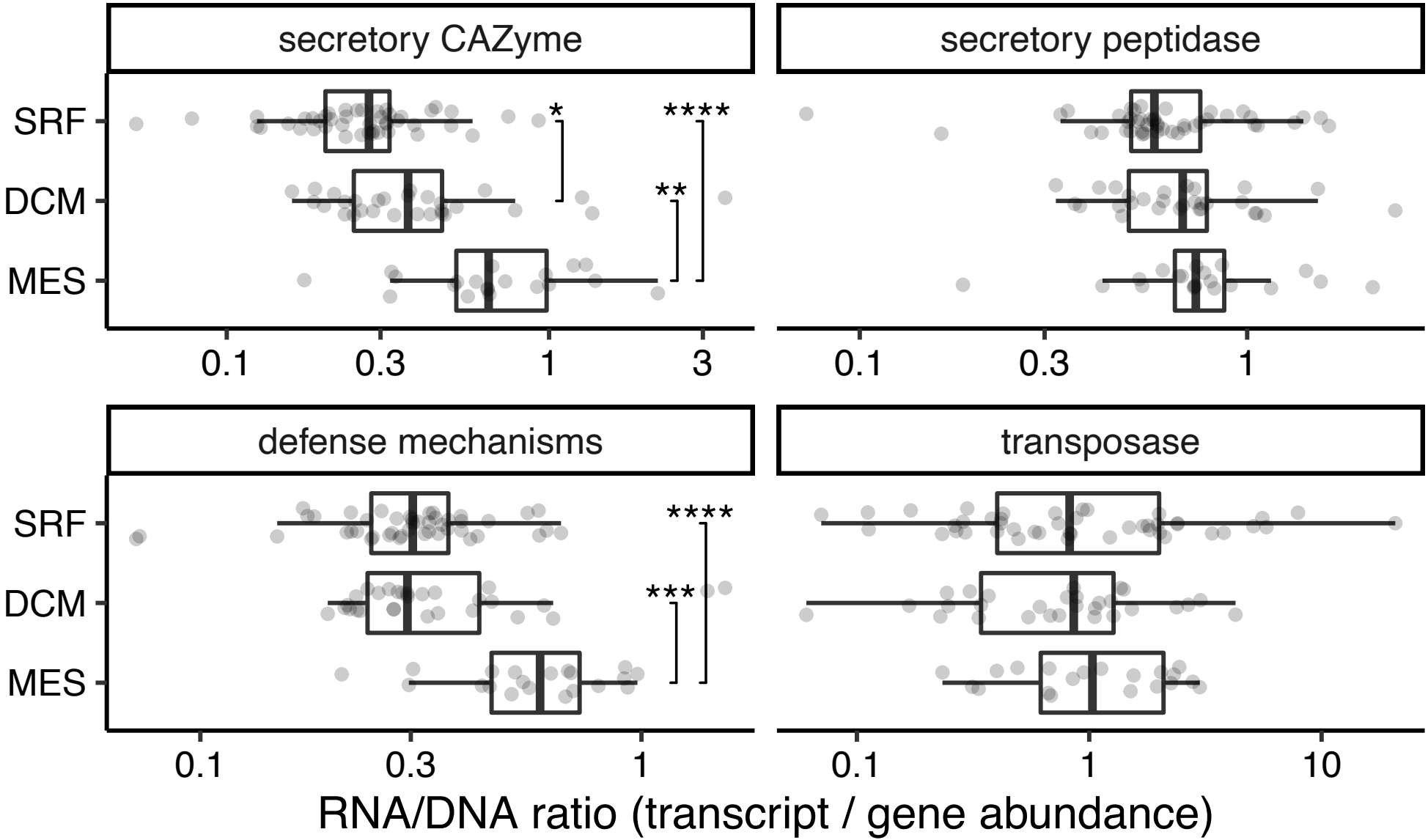

depth
SRF (n = 64, depth < 10 m)
DCM (n = 44, 17 – 120 m)
MES (n = 30, 250 – 1000 m)
BAT (n = 58, 2400 – 4000 m)
c
b
a
a
Tara Oceans metagenomes
Malaspina metagenomes
0.003
0.010
0.030
0.100
0.300
transposase gene abundance (%)

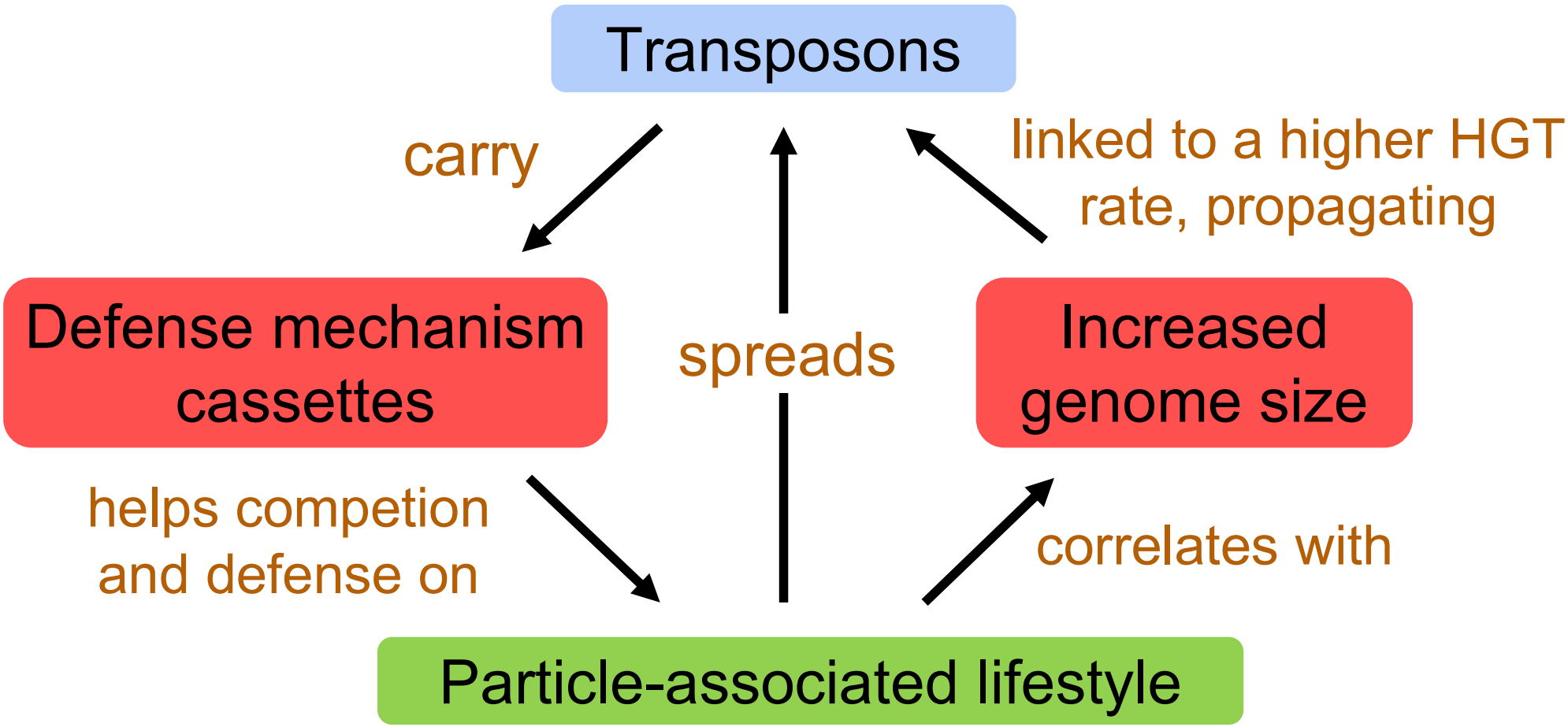

ρ = 0.68, p < 2.2e−16
ρ = 0.73, p < 2.2e−16
ρ = 0.5, p = 4.3e−11
ρ = 0.44, p = 1.8e−08
transposase gene abundance (%)
transposase transcript abundance (%)
secretory CAZyme genes (%)
secretory peptidase genes (%)
secretory CAZyme transcripts (%)
secretory peptidase transcripts (%)
Depth
SRF
DCM
MES
BAT


A
SRF
(n = 329)
DCM
(n = 339)
MES
(n = 479)
BAT
(n = 255)
b
b
a
a
1
3
10
complete genome size (Mbp)
B
Transposase abundance (%)
0.03
0.1
0.3
1
CAZymes in MAGs
peptidases in MAGs
ρ = 0.43, p < 2.2e−16
ρ = 0.29, p = 2.5e−14
ρ = 0.44, p < 2.2e−16
ρ = 0.21, p = 6.9e−09
Shallow (SRF, DCM)
Deep (MES, BAT)
complete genome size (Mbp)
secretory CAZyme/peptidase (%)
C
SRF
DCM
MES
BAT
transposase ORFs in MAGs (%)
107
136
54
13
19
105
124
57
27
26
55
81
162
105
76
39
57
82
43
34
0–2
2–3
3–4
4–5
>5
complete genome size (Mbp)


A
SRF (n = 380, 166 MAGs)
DCM (n = 297, 165 MAGs)
MES (n = 443, 139 MAGs)
BAT (n = 113, 37 MAGs)
a
a
b
c
0.0
0.1
0.2
0.3
0.4
MAG median pN/pS
B
SRF
DCM
MES
BAT
ρ = −0.03, p = 0.56
ρ = 0.17, p = 0.0028
ρ = −0.51, p < 2.2e−16
ρ = 0.01, p = 0.91
1.00
0.30
0.10
0.03
transposase ORF in MAGs (%)
MAG median pN/pS


A
COG gene−calls for:
Cassette ORFs (n = 2310)
All ORFs in metagenomes
Cytoskeleton
Cell motility
Cell cycle control...
Extracellular structures
Intracellular trafficking...
Nucleotide T.M.
Inorganic ion T.M.
Carbohydrate T.M.
Lipid T.M.
Secondary metabolites...
Signal transduction mechanisms
Amino acid T.M.
Coenzyme T.M.
Cell wall/membrane/envelope biogenesis
Energy production and conversion
Posttranslational modification...
Transcription
Mobilome: prophages, transposons
Translation, ribosomal structure...
Function unknown
General function prediction only
Defense mechanisms
Replication, recombination and repair
0.00
0.05
0.10
0.15
proportion to total ORF calls
B
SRF
DCM
MES
BAT
CAZyme
ρ = 0.41 ****
peptidase
ρ = 0.61 ****
defense mechanisms gene abundance (%)
secretory CAZyme/peptidase genes (%)


secretory CAZyme
secretory peptidase
defense mechanisms
transposase
SRF
DCM
MES
*
**
****
***
0.1
0.3
1
3
10
RNA/DNA ratio (transcript / gene abundance)



A deep-ocean microbial community