High transposase abundance in the deep ocean is linked to 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

Abstract

Transposases are mobile genetic elements (MGEs) that move within and between genomes, enabling genomic plasticity within microorganisms. Transposases often carry functional genes, termed cassette sequences, when they migrate within and between genomes. In marine microbial communities, transposase abundance increases with depth, but the reasons for this trend are unclear. Through analysis of metagenomes from the *Tara* Oceans and Malaspina Expeditions, we identify the particle-associated lifestyle as a main covariate for the high occurrence of transposases in the deep ocean. The abundance of biofilm-associated genes correlates with transposases in a strong and depth-independent manner, suggesting that MGEs readily propagate among microbial communities in crowded biofilms. We show that particle association positively correlates with larger genome size, which is linked to higher transposase abundance. Genes related to defense mechanisms are enriched in cassette sequences associated with transposons and are more highly expressed in the deep sea. Thus, while transposons spread at the expense of their microbial hosts, they introduce novel genes and potentially benefit the hosts in accessing heavily contested resources. Together, our results reveal an emerging picture of deep ocean particles as gene sharing highways that result in expanded and defensively oriented microbial genomes.

Introduction

Mobile genetic elements (MGEs) are DNA segments that can mediate the movement of genetic sequences both within and between bacterial and archaeal genomes [1]. MGEs generally encode enzymes to mediate this process [1]: in transposons, the mediating enzymes are transposases, through which the mechanism of transfer results in inverted repeats [2, 3]. When MGEs migrate from one genomic region to another, some carry cassette sequences, many of which have been

reported to encode functional genes such as antibiotic resistance [4] and metal uptake [5, 6], and regulatory genes that affect gene expression in host cells [7, 8].

Transposases are the most abundant and ubiquitous genes in nature [9], despite their uneven distribution across taxa [10, 11]. In marine systems, one of the most striking – and as of yet unexplained – trends in microbial genomics is the distinct increase in transposase abundance with depth. Genomic analyses from station ALOHA in the North Pacific revealed a substantial increase in transposase abundance with depth from 500 m to their observed maximum at 4000 m [12, 13]. As one of the most overrepresented COG-categories in ALOHA deep waters, transposases accounted for 1.2% of all fosmid sequences from 4000 m [12, 14]. Similarly, a study of hydrothermal chimneys showed that transposases were particularly abundant in biofilms and accounted for 8% of all metagenomic reads, a 10-fold increase compared to metagenomes in other habitats [15]. Based on these observations, our analyses focused on three central questions: (1) why is transposase abundance associated with greater depths and biofilms in marine systems, (2) are transposases selfish genes that spread widely with neutral or deleterious effects on host genomes, and (3) do transposases perform useful functions for microbial hosts that encode them?

To identify reasons for the high transposase abundance in the deep sea and to expand our knowledge of the role of MGEs in marine microbial communities, we analyzed 138 microbial metagenomes and 152 microbial metatranscriptomes from the *Tara* Oceans Expedition [16, 17]. The *Tara* Oceans samples were from 5-1000 m in depth; to represent bathypelagic microbial communities, we also incorporated 58 metagenomes, sampled between 2400-4000 m in depth, from the 2010 Malaspina Expedition [18]. To explore the role of transposases at a genome-resolved level, we analyzed 1402 metagenome assembled genomes (MAGs) previously generated from the *Tara* Oceans [19] and Malaspina [18] metagenomes. Previous work has demonstrated that deep-sea prokaryotes have a predominantly particle-associated lifestyle [20]. Here, we show that the high

abundance of transposases in deep ocean microbial communities was strongly linked to the shift towards a particle-associated lifestyle. We identified taxonomy, particle association, and genome size as key covariates to transposase abundance in MAGs. Further, we found that among the open reading frames (ORFs) in cassette sequences, the functional categories "defense mechanism" were highly abundant, suggesting that transposons introduce beneficial genes to their microbial hosts inhabiting highly competitive habitats.

Results and Discussion

Previous studies have revealed that transposase abundance in microbial genomes is associated with increased ocean depth, lower dissolved oxygen (DO) concentrations, and a particle-associated lifestyle (as opposed to a planktonic lifestyle)[10, 13, 21]. However, the reason and nature of these correlations has not been further explored. To confirm these findings on a broader scale, we screened for significant covariates for transposase abundance in the *Tara* Oceans and Malaspina metagenomes. We plotted the transposase abundance against depth and DO. In our metagenome analysis, transposase abundance was defined as the proportion of reads mapped back to all transposase ORFs (the database used to identify putative transposase ORFs is shown in **Table S1**). We confirmed that transposase abundance increases steadily as depth increases, despite large variations across samples (**Fig. 1**). In every ocean, the transposase abundances of the deep water samples were higher than those of the shallow water samples (**Table S2**). Here, "deep water" was defined as mesopelagic (depth 250-1000 m) and bathypelagic (2500-4000 m) waters, and "shallow water" was defined as surface (<10 m) and deep chlorophyll maximum (DCM, depth 17-120 m) waters.

The depth of a sample (e.g., surface, DCM) was a significant predictor for transposase abundance (Table S3). Adding dissolved oxygen, however, did not significantly improve the

prediction of transposase abundance in a sample (ANOVA *F*-test $p = 0.21$, Table S3). This was partially due to the inconsistent relationship between DO and transposase abundance across depths. In the surface and bathypelagic samples, DO positively correlated with transposase abundance, but this correlation was reversed in DCM and mesopelagic samples (Fig. S1).

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

To further investigate the relationship between transposase abundance and particle-associated lifestyles, we examined the correlation between transposase abundance and the percentage of predicted secretory carbohydrate-active enzymes (CAZymes) and peptidases among all CAZyme and peptidase ORFs (see Methods). The percentage of secretory enzymes among all CAZymes and peptidases has previously been applied to quantify the degree to which microbial communities rely on a particle-associated lifestyle [20]. Microorganisms rely on extracellular enzymes to degrade large particulate organic carbon into compounds of smaller molecular weight to incorporate into the cell [20, 22] and CAZymes and peptidases are key enzymes for carbohydrate and protein degradation, respectively [22]. To account for differences in biological processes at different depths, we normalized the gene and transcript abundance of predicted secretory CAZymes and peptidases by those of all CAZyme and peptidase ORFs in metatranscriptomic and metagenomic samples (the bathypelagic zone was excluded for all transcript analysis, because no metatranscriptomes were available for Malaspina samples).

From the *Tara* Oceans metagenomes, we identified 52,890 secretory CAZymes out of 421,080 CAZyme ORFs, and 179,932 secretory peptidases out of 1,294,210 peptidase ORFs; from the Malaspina metagenomes, we identified 7,002 secretory CAZymes in 40,415 CAZyme ORFs, and

18,014 secretory peptidases out of 67,617 peptidases (see Methods). We observed an increasing percentage of secretory CAZymes and peptidases with depth in both metagenomic and metatranscriptomic samples (**Fig. S2**), which was consistent with previous findings [20]. The increase in the percentage of secretory CAZymes and peptidases with depth indicates increased extracellular enzyme activity, suggesting a shift toward a particle-associated lifestyle toward the deep ocean.

Marine particles are thought to be hotspots for horizontal gene transfer and transposase propagation [23], and the increased importance of a particle-associated lifestyle with depth in the ocean offers a promising explanation for the elevated transposase abundance in the deep ocean. In both metagenomic and metatranscriptomic samples, the abundance of transposases and the percentage of secretory CAZymes and peptidases were strongly correlated (**Fig. 2**). Given a model with only depth as predictor of the transposase abundance in a sample, the addition of secretory CAZymes and peptidases each significantly improved the accuracy of prediction (both ANOVA *F*-test $p = 3\text{x}10^{-7}$, Table S3). In addition, the percentage of secretory CAZymes and peptidases each had consistent positive correlations with transposase gene and transcript abundance across depths (**Fig. S3**). Correlations in metagenomes and metatranscriptomes suggested that transposases are more abundant and more frequently transcribed in microbial communities with a greater reliance on a particle-associated lifestyle. After observing such trends at the community level, we sought to determine if these trends hold true at a genome-resolved level to support our hypothesis that the particle-associated lifestyle is a main driver for the high transposase abundance in the deep ocean.

Genome size and taxonomy are key factors to transposase abundance in populations.

To identify factors affecting transposase abundance at a genome-resolved level, we analyzed 1147 MAGs recovered from the *Tara* Oceans metagenomes [19] and 255 MAGs from the Malaspina

metagenomes [18]. All MAGs in our analysis had percent completeness >70%, redundancy <10%, and contained <5% of eukaryotic sequences (see Methods). The transposase abundance in MAGs was quantified by the percentage of transposase ORFs among all ORFs in the MAG (abbreviated as *%-transposase*). Similar to metagenomes, MAGs from deep waters had higher transposase abundance than MAGs from shallow waters. Specifically, MAGs from the mesopelagic and bathypelagic zones had a mean %-transposase 6 times (95% CI: 5.38 to 6.57) that of MAGs from the surface and DCM.

Given the enrichment of transposase ORFs in deep water MAGs, we ask 3 questions: (1) whether the correlation between transposase abundance and the particle-associated lifestyle persists in MAGs; (2) since larger genomes have an elevated rate of HGT [25, 26], whether genome size correlates to transposase abundance in MAGs; and (3) whether specific taxa encode more transposases than others, and thus whether community composition plays a role in determining transposase abundance.

As with the metagenomes, we quantified the degree to which individual genomes were particle-associated by calculating the percentage of secretory CAZymes and peptidases out of all CAZyme and peptidase ORFs in a MAG. Taxonomy was analyzed at a class level, and the "complete genome size" was estimated using the genome length of a MAG divided by its estimated percent completeness. Although MAG assembly might bias against MGEs [28] and the completeness of MAGs might be underestimated for rare taxonomy groups [29, 30], we can still gain valuable insights by analyzing the distribution of transposase ORFs in MAGs of various genome sizes. To address the first question, we observed a positive correlation between the percentage of transposase ORFs and the percentage of secretory CAZymes and peptidases in MAGs (Spearman $\rho = 0.208$ and $\rho = 0.210$, respectively; both $p < 10^{-14}$), thus confirming that the correlation between transposases and particle association persists on a genome-by-genome basis.

To examine the relationship between genome size and transposase abundance, we compared the estimated genome size of each MAG with the percentage of ORFs characterized as transposases within each MAG. In every depth, we observed a positive correlation between the estimated genome size and the percentage of transposase ORFs within a MAG (Spearman ϱ was 0.56, 0.50, 0.46, 0.60 for surface, DCM, mesopelagic, and bathypelagic samples, respectively. All $p < 10^{-20}$), confirming previous speculations that larger genomes were linked to high transposase abundance [10, 13]. While deep water genomes were bigger (**Fig. 3A**), the particle-associated lifestyle correlated with greater complete genome size in every depth (**Fig. 3B**). Previous work has shown that large genomes are associated with a high fraction of transposase ORFs in isolated NCBI microbial genomes [11, 26, 27], and in a study of metagenomes from the Baltic Sea [10]. The correlation between transposase abundance and genome size was previously attributed to a higher frequency of HGT in larger genomes [25, 26]. The correlation between the particle-associated lifestyle and genome size has also been previously observed and may be linked to the expanded metabolic versatility of particle-associated microbial populations [31]. Thus, the particle-associated lifestyle might facilitate transposase propagation by reducing the distance between microorganisms, and may lead to larger genome size through increased rates of HGT. It is worth noting that the transposase abundance in MAGs peaked when the complete genome size reached 5 Mbp, but it stabilized or declined beyond the peak (**Fig. 3C**). A previous study has also reported a decreased proportion of transposase ORFs after bacterial genomes surpass 6 Mbp in size [11]. However, most (88.9%) MAGs recovered from the *Tara* Oceans and Malaspina metagenomes were under 5 Mbp, so the association between large genomes and high transposase abundance generally held true.

To address the third question, we found that taxonomic class was also a key predictor of transposase abundance in MAGs; information on the taxonomy and depth of MAGs together explained 53% of the variance in transposase abundance (**Table 1**). MAGs of the taxonomic classes

Alpha-, *Beta*-, *Gammaproteobacteria*, and *Actinobacteria* were enriched in transposases compared to other MAGs (for each class with ≥10 MAGs, The %-transposase of MAGs in that class were compared against those of all other MAGs; Wilcox-test p cutoff: 0.05). MAGs of *Flavobacteria*, *Acidimicrobidae*, *novelClass_E*, and *SAR202-2* had low transposase abundance. Taxa that were enriched/low in transposases mostly matched with previous studies [10, 11], except that *Actinobacteria* had previously been reported to be low in transposase abundance.

The link between taxonomy and transposase abundance could partially be attributed to depth. MAGs of transposase-enriched classes were more abundant in deep waters, and MAGs from low-transposase classes were more abundant in shallow waters (2 χ^2-tests, both $p < 1\text{x}10^{-6}$; **Fig. S4**). Thus, although community composition is a key covariate to transposase abundance, the distribution of taxa is nonetheless linked to the depth of a microbial community.

Relaxed selection does not explain transposase abundance in genomes.

It is possible that transposases accumulate in deep ocean microbial populations as a result of genetic drift, because those populations tend to be smaller in size [32] and experience slow growth rates [33]. To examine this possibility, we tested whether the strength of selection experienced by populations correlated with their transposase abundance. For each ORF in each MAG, we calculated its pN/pS ratio in each sample from its designated depth (pN/pS was only calculated for ORFs with ≥20x coverage). pN/pS represents the proportion of nonsynonymous mutations to the proportion of synonymous mutations, and characterizes selection at the level of the population, in contrast to dN/dS, which characterizes selection between individual species [34–36].

We observed that MAGs from the mesopelagic zone had a higher median pN/pS than MAGs from the surface and DCM, and MAGs from the bathypelagic zone had a higher median pN/pS than MAGs from the mesopelagic zone (both $p < 5 \text{ x } 10^{-6}$, **Fig. 4A**). The high median pN/pS

of populations from the deeper layers suggested relaxed selection pressure relative to populations from shallower layers, which was in agreement with previous results [13]. The general explanation for such a trend is a slower replication rate, which leads to smaller effective population sizes and a reduced selection effect in the deep ocean [12, 13]. However, further work is needed to substantiate this hypothesis.

Since transposases and deep-sea ORFs were both under relaxed selection, we tested whether high transposase abundance correlated with relaxed selection in a population. However, we did not observe a correlation between the median *pN*/*pS* of MAGs and their transposase abundance (Spearman ρ = -0.075, p = 0.08). Furthermore, the relationship between the selective pressure experienced by a population and its transposase abundance was inconsistent across different depths (**Fig. 4B**). It has been previously suggested that small populations would experience relaxed selection against mobile genetic elements [38, 39]. However, our analyses show that although relaxed selection and transposase enrichment co-occur in deep oceans, we did not observe this trend on a genome-by-genome basis, suggesting that there is not a direct connection between transposase abundance and relaxed selection at a genome-resolved scale, and thus within individual microbial populations.

Transposons/integrons carry a high proportion of ORFs related to defense mechanisms

One important question regarding high transposase abundance in the deep sea is whether these transposases are neutral or deleterious, or whether they perform important functions in the genomes encoding them. In some cases, transposases are deleterious and proliferate as a result of genetic drift [40], but transposases can also benefit the host by introducing functional and regulatory genes [7, 23, 41, 42].

228 Transposases mediate the migration and integration of cassette sequences into host genomes;
229 a transposon or integron is the combination of a transposase and its cassettes. Thus, cassette
230 sequences are functionally similar to auxiliary metabolic genes (AMGs) [43], in that they are carried
231 by a selfish MGE (or viruses in the case of AMGs) to increase the fitness of the host and therefore
232 increase the fitness of the MGE.

233 To determine whether transposons in marine habitats carry advantageous genes, we began
234 by querying the functional categories of cassette sequences. The software package Integron Finder
235 [11] was used to locate cassettes on contigs; the program searches for the two palindromic flanking
236 sites necessary for transposase-mediated recombination as well as a nearby transposase/integrase.
237 We treated transposons and integrons as equivalent, due to the sequence and function similarity
238 between integrases and transposases [11]. We found 8,519 cassette ORFs from the co-assembled
239 Malaspina [18] and *Tara* Oceans [17, 19] metagenomes (see Methods). Only 27% of the cassette
240 sequences were assigned with a clusters of orthologous genes (COG) annotation, which was a lower
241 annotation rate compared to other ORFs in the *Tara* Oceans (55-60%) and Malaspina (68%)
242 metagenomes.

243 Compared to the rest of the metagenome, cassettes were enriched in ORFs of the COG
244 categories "replication, recombination, and repair", "defense mechanisms", and "mobilome:
245 prophages, transposons" (**Fig. 5A**). Transposons/integrons are expected to be enriched in ORFs
246 related to the mobilome and replication/recombination. The prevalence of defense mechanism
247 genes in cassettes was noteworthy: defense mechanisms accounted for 14.3% of known function
248 calls in cassettes, but only accounted for 2.47% of known function calls in non-cassette ORFs
249 (*known* was defined as all functional groups except "function unknown" and "general function
250 prediction"). Out of the 276 defense mechanism cassettes, 123 of them encoded toxin/antitoxin
251 genes. Since transposases are more pervasive on particles, toxin and defense genes they carry would

be useful in competition and protection on crowded particles [45–48]. Moreover, the percentage of secretory CAZymes and peptidases in metagenomes both correlated with the abundance of defense mechanism ORFs (**Fig. 5B**). Thus, if defense mechanism genes in the deep ocean were more highly expressed, this would substantiate the hypothesis that deep ocean microorganisms benefit from novel genes introduced by integrons/transposons.

Expression of transposase, biofilm, and defense mechanisms ORFs

To determine whether defense mechanism genes are more highly expressed and how this correlates to expression of particle-associated genes, we calculated their RNA/DNA ratios (defined as the transcript abundance of a target gene divided by its gene abundance) of secretory CAZyme, secretory peptidase, defense mechanism, and transposase ORFs in each *Tara* Oceans sample with a paired metagenome and metatranscriptome (n = 94). Malaspina samples were excluded from this analysis, because no metatranscriptomes were sequenced.

The ORFs related to secretory CAZymes and defense mechanisms had greater RNA/DNA ratios in the mesopelagic zone than in the surface and DCM (**Fig. 6**), demonstrating a greater need for these genes as deep ocean microbial communities switch toward a particle-associated lifestyle. In contrast, although transposases were more abundant in the mesopelagic zone (**Fig. S5**), their RNA/DNA ratios were similar across all depths (**Fig. 6**). A possible explanation is that although transposases are not upregulated, they sometimes carry beneficial cassette genes, which are increasingly expressed in deep-sea microbial genomes.

Conclusion

Our analysis suggests a cyclic interaction between the particle-associated lifestyle, expanded genome size, and selection for defense mechanism genes to explain the increasing abundance of

transposases with depth (**Fig. 7**). From the surface to the bathypelagic zone, microbial communities shift from a predominantly planktonic lifestyle toward a predominantly particle-associated lifestyle, observed both in the literature [20, 49] and here via the increased transcript abundance of secretory CAZymes and peptidases in deeper oceans. On densely populated particles, the close proximity between microorganisms leads to resource competition and allows transposases to spread more rampantly. Particle-associated microorganisms also tend to have larger genomes, which are associated with high transposase abundance. Thus, deep-sea prokaryotic genomes sustain high abundances of MGEs, allowing transposons to actively shape the ecology of particle-associated microbial communities with the novel genes they introduce. Most notably, cassette sequences are enriched in defense mechanism genes, which are more highly expressed in the deep ocean and can benefit microorganisms engaging in the intense competition on particles [45, 47]. Altogether, transposons not only exploit the dense microbial population on particles to propagate but also contribute functional genes that help hosts compete in such an environment.

Investigating the elevated transposase abundance in the deep ocean is crucial for understanding gene flow, competition, and other ecological characteristics of the dark ocean, one of the largest habitats on Earth. Here, we show a strong association between transposase abundance and the particle-associated lifestyle, suggesting that transposons can enable microbial lineages to compete in crowded biofilm-associated habitats. We did not find correlation between the strength of selection and transposase abundance at a genome-resolved scale, so the relaxed selection pressure is unlikely to be the primary driver for the high transposase abundance in the deep ocean. Overall, our results reveal an emerging picture of the ocean as a stratified system in which the deep ocean acts as a gene sharing highway with increased networks of gene exchange, particularly on particles, resulting in large and defensively oriented genomes in the deep-sea microbial communities, compared to the streamlined and specialized genomes that dominate the surface oceans.

Materials and Methods

Analysis of the Ocean Microbial Reference Catalog v2 (OM-RGC.v2) from the *Tara* Oceans Project. OM-RGC.v2 contained 47 million non-redundant open reading frames (ORFs) [16, 17], and the coverage of each ORF in every metagenomic (n = 138) and metatranscriptomic (n = 152) sample (website)[16]. Samples from the Arctic Ocean were excluded. We identified transposases and biofilm-associated ORFs in OM-RGC.v2 through TBLASTN (e-value < 10^{-5}). The transposase database contained "transposase" and "integrase" genes from the Pfam database [50] (Table S1). We identified ORFs of the "defense mechanisms" category through COG annotations in OM-RGC.v2. Finally, the coverage data in OM-RGC.v2 was used to determine the relative abundance of a target gene in a sample.

Acquisition of global ocean metagenomes. Metagenomic reads of the *Tara* Oceans Project and Malaspina Expedition were downloaded from the European Nucleotide Archive (ENA) under study accessions PRJEB402 and PRJEB44456, respectively. Methods for sample collection and Illumina sequencing were described by Sunagawa *et al.* [54] for the *Tara* Oceans Project and Acinas *et al.* [18] for the Malaspina Expedition. For the assembled contigs used as references for mapping of the *Tara* Oceans metagenomes, we downloaded (website) 10 existing co-assembled metagenomes from the 10 oceanic provinces [19]. For the reference contigs of the Malaspina metagenomes, we downloaded (website) the co-assembled metagenome of all 58 samples [18]. Samples from the *Tara* Oceans were filtered through 1.6- or 3-μm filters and collected on 0.22-μm filters [54], whereas samples from the Malaspina expedition were collected from the 0.2 – 0.8 μm and 0.8 – 5 μm size fractions [18]. We used EukDetect [55] to confirm that samples collected with a larger size fraction did not have a higher contribution from eukaryotic reads (see Supplementary Text).

Mapping and calculating *pN/pS* ratios. To assess the strength of selection of target ORFs, we calculated the *pN/pS* ratio (the proportion of nonsynonymous mutations over the proportion of synonymous mutations) for all ORFs. To do this, we mapped the metagenomic reads against the assembled contigs for each metagenome. We mapped the raw reads of each of the *Tara* Oceans metagenomes to the co-assembly from their corresponding provinces, and all raw reads of each of the Malaspina metagenomes to one co-assembled metagenome using Bowtie2 [56] (v2.2.9; paired-end alignment with default parameters). We calculated *pN/pS* using anvi'o [57] with the script "anvi-script-calculate-pn-ps-ratio" with the "--min-coverage" flag set to 20. The *pN/pS* ratios were calculated on a sample-per-sample basis, so an ORF from a co-assembled metagenome might have multiple *pN/pS* ratios from different samples.

Prediction of secretory CAZyme and peptidase ORFs. We adapted the methods employed in Zhao et al [20] to assess whether microbial communities were largely planktonic or particle-associated by quantifying the relative abundance of secretory CAZymes and peptidases. Predicted CAZymes and peptidases were annotated using DIAMOND (2.0.15) [58] BLASTp (e-value $<10^{-10}$) to search against the dbCAN [59] and MEROPS [60] databases, respectively. SignalP (5.0) [61] was used to identify signal peptides. We used the Gram-positive mode for ORFs affiliated with *Actinobacteria* and *Firmicutes*, Gram-negative mode for other bacterial phylum, and Archaea mode for ORFs affiliated with the *Archaea* domain. Only ORFs in the *Bacteria* or *Archaea* domains were included in the analysis. For MAGs, counts of secretory CAZymes and peptidases were used instead of abundance; validation based on predicted lifestyles of known taxa confirmed that this gave an accurate prediction of lifestyle (see Supplementary Text).

MAG Selection. To conduct genome-resolved analyses, we selected 1147 out of 2631 MAGs that had previously been recovered from the Tara Ocean metagenomes [19] and 255 out of 317 MAGs recovered from the Malaspina metagenomes [18]. All selected MAGs had <5% of eukaryotic sequences (in base pairs, Tiara [62] was used to identify putative eukaryotic contigs), completeness > 70%, and redundancy < 10% (both criteria were calculated with CheckM [24]).

Determination of MAG's depth and lifestyles. MAGs that originated from co-assemblies in a specific province in the *Tara* Oceans dataset (*e.g.*, Red Sea), as in Tully *et al.* [19], were used to recruit reads from all *Tara* Oceans samples (n = 180) using Bowtie2 (default parameters). Each province was performed separately. SAM files were converted to BAM files using samtools [63] and used to determine RPKM (reads per kilobase pair MAG per Million pair metagenome) as in Graham et al. [64].

To assign the depth origin of a MAG from the *Tara* Oceans metagenomes, we summed up its RPKM in surface, DCM, and mesopelagic samples (3 groups, separately). The layer with the highest sum RPKM was the depth of that MAG. All Malaspina MAGs belonged to the bathypelagic layer.

Integron and cassette sequences detection. We used Integron Finder (v2) [11] to identify cassette sequences from the *Tara* Ocean and Malaspina metagenomes. Integron Finder uses HMMER to locate the integron-integrase *intI*, which is conserved for most integrons. Then, cassette sequences are identified with the near-palindromic flaking regions. anvi'o [47] was used to identify and annotate ORFs on metagenomic contigs using the script 'anvi-run-ncbi-cogs.' If the anvi'o ORF calls from Prodigal [65] were within 100bps (start and stop) of the Integron Finder's cassette calls, the anvi'o COG annotations were used for those cassette ORFs.

Statistical analysis

We used the Wilcoxon signed-rank test to compare the differences between means of two numeric variables, and the Spearman's rank correlation was used to determine the association between two numeric variables. Multiple hypothesis p-values were adjusted using the Benjamini–Hochberg procedure. All linear regressions had a normal distribution of residuals (Shapiro–Wilk test). Statistical significance was assumed if $p < 0.05$. All statistical analyses were performed using base R packages (v4.1.3).

Code and data availability.

All Python, R scripts, code explanation, and raw data for analysis are publicly accessible on GitHub 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.

References

1. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* 2005; **3**: 722–732.
2. Transposons. http://www.nature.com/scitable/topicpage/transposons-the-jumping-genes-518. Accessed 30 Nov 2021.
3. Muñoz-López M, García-Pérez JL. DNA transposons: nature and applications in genomics. *Curr Genomics* 2010; **11**: 115–128.
4. Collis C M, Hall R M. Expression of antibiotic resistance genes in the integrated cassettes of integrons. *Antimicrob Agents Chemother* 1995; **39**: 155–162.
5. Rankin DJ, Rocha EPC, Brown SP. What traits are carried on mobile genetic elements, and why? *Heredity* 2011; **106**: 1–10.
6. Hacker J, Carniel E. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep* 2001; **2**: 376–381.
7. Jones JM, Grinberg I, Eldar A, Grossman AD. A mobile genetic element increases bacterial host fitness by manipulating development. *Elife* 2021; **10**: e65924.
8. Casacuberta E, González J. The impact of transposable elements in environmental adaptation. *Mol Ecol* 2013; **22**: 1503–1517.
9. Aziz RK, Breitbart M, Edwards RA. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res* 2010; **38**: 4207–4217.
10. Vigil-Stenman T, Ininbergs K, Bergman B, Ekman M. High abundance and expression of transposases in bacteria from the Baltic Sea. *ISME J* 2017; **11**: 2611–2623.
11. Cury J, Jové T, Touchon M, Néron B, Rocha EP. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res* 2016; **44**: 4539–4550.
12. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U, et al. Community

genomics among stratified microbial assemblages in the ocean's interior. *Science* 2006; **311**: 496–503.

13. Konstantinidis KT, Braff J, Karl DM, DeLong EF. Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre. *Appl Environ Microbiol* 2009; **75**: 5345–5355.
14. Worden AZ, Cuvelier ML, Bartlett DH. In-depth analyses of marine microbial community genomics. *Trends Microbiol* 2006; **14**: 331–336.
15. Brazelton WJ, Baross JA. Abundant transposases encoded by the metagenome of a hydrothermal chimney biofilm. *ISME J* 2009; **3**: 1420–1424.
16. Salazar G, Paoli L, Alberti A, Huerta-Cepas J, Ruscheweyh H-J, Cuenca M, et al. Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* 2019; **179**: 1068-1083.e21.
17. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. *Science* 2015.
18. Acinas SG, Sánchez P, Salazar G, Cornejo-Castillo FM, Sebastián M, Logares R, et al. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Communications Biology* 2021; **4**: 1–15.
19. Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data* 2018; **5**: 1–8.
20. Zhao Z, Baltar F, Herndl GJ. Linking extracellular enzymes to phylogeny indicates a predominantly particle-associated lifestyle of deep-sea prokaryotes. *Sci Adv* 2020; **6**: eaaz4354.
21. Ganesh S, Parris DJ, DeLong EF, Stewart FJ. Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. *ISME J* 2014; **8**: 187–211.
22. Arnosti C. Microbial extracellular enzymes and the marine carbon cycle. *Ann Rev Mar Sci* 2011;

3: 401–425.

23. Stewart FJ. Where the genes flow. *Nat Geosci* 2013; **6**: 688–690.

24. Cordero OX, Hogeweg P. The impact of long-distance horizontal gene transfer on prokaryotic genome size. *Proc Natl Acad Sci U S A* 2009; **106**: 21748–21753.

25. Iranzo J, Gómez MJ, López de Saro FJ, Manrubia S. Large-scale genomic analysis suggests a neutral punctuated dynamics of transposable elements in bacterial genomes. *PLoS Comput Biol* 2014; **10**: e1003680.

26. Meziti A, Rodriguez-R LM, Hatt JK, Peña-Gonzalez A, Levy K, Konstantinidis KT. The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample. *Appl Environ Microbiol* 2021; **87**.

27. Garcia SL, Buck M, McMahon KD, Grossart H-P, Eiler A, Warnecke F. Auxotrophy and intrapopulation complementary in the "interactome" of a cultivated freshwater model community. *Mol Ecol* 2015; **24**: 4449–4459.

28. Rodríguez-Gijón A, Nuy JK, Mehrshad M, Buck M, Schulz F, Woyke T, et al. A Genomic Perspective Across Earth's Microbiomes Reveals That Genome Size in Archaea and Bacteria Is Linked to Ecosystem Type and Trophic Strategy. *Front Microbiol* 2021; **12**: 761869.

29. Touchon M, Rocha EPC. Causes of Insertion Sequences Abundance in Prokaryotic Genomes. *Mol Biol Evol* 2007; **24**: 969–981.

30. Leu AO, Eppley JM, Burger A, DeLong EF. Diverse Genomic Traits Differentiate Sinking-Particle-Associated versus Free-Living Microbes throughout the Oligotrophic Open Ocean Water Column. *MBio* 2022; **13**: e0156922.

31. Costello MJ, Chaudhary C. Marine Biodiversity, Biogeography, Deep-Sea Gradients, and Conservation. *Curr Biol* 2017; **27**: R511–R527.

32. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, et al. Microbial diversity in the deep sea and the underexplored "rare biosphere." *Proc Natl Acad Sci U S A* 2006; **103**: 12115–12120.

33. Simmons SL, Dibartolo G, Denef VJ, Goltsman DSA, Thelen MP, Banfield JF. Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. *PLoS Biol* 2008; **6**: e177.

34. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 1991; **351**: 652–654.

35. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, et al. Genomic variation landscape of the human gut microbiome. *Nature* 2013; **493**: 45–50.

36. Massana R. Eukaryotic Picoplankton in Surface Oceans. *Annu Rev Microbiol* 2011; **65**: 91–110.

37. Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, et al. Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. *Nat Genet* 2003; **35**: 32–40.

38. Moran NA, Plague GR. Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev* 2004; **14**: 627–633.

39. Escobar-Páramo P, Ghosh S, DiRuggiero J. Evidence for Genetic Drift in the Diversification of a Geographically Isolated Population of the Hyperthermophilic Archaeon Pyrococcus. *Molecular Biology and Evolution* . 2005. , **22**: 2297–2303

40. Flynn KJ, Swanson MS. Integrative conjugative element ICE-βox confers oxidative stress resistance to Legionella pneumophila in vitro and in macrophages. *MBio* 2014; **5**: e01091-14.

41. Sullivan JT, Ronson CW. Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. *Proc Natl Acad Sci U S A* 1998; **95**: 5145–5149.

42. Breitbart M, Bonnain C, Malki K, Sawaya NA. Phage puppet masters of the marine microbial

realm. *Nat Microbiol* 2018; **3**: 754–766.

43. Nadell CD, Drescher K, Foster KR. Spatial structure, cooperation and competition in biofilms. *Nat Rev Microbiol* 2016; **14**: 589–600.

44. Basler M, Ho BT, Mekalanos JJ. Tit-for-tat: type VI secretion system counterattack during bacterial cell-cell interactions. *Cell* 2013; **152**: 884–894.

45. Hayes CS, Aoki SK, Low DA. Bacterial contact-dependent delivery systems. *Annu Rev Genet* 2010; **44**: 71–90.

46. Ho BT, Dong TG, Mekalanos JJ. A view to a kill: the bacterial type VI secretion system. *Cell Host Microbe* 2014; **15**: 9–21.

47. Mestre M, Ruiz-González C, Logares R, Duarte CM, Gasol JM, Sala MM. Sinking particles promote vertical connectivity in the ocean microbiome. *Proc Natl Acad Sci U S A* 2018; **115**: E6799–E6807.

48. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res* 2021; **49**: D412–D419.

49. Taglialegna A, Matilla-Cuenca L, Dorado-Morales P, Navarro S, Ventura S, Garnett JA, et al. The biofilm-associated surface protein Esp of Enterococcus faecalis forms amyloid-like fibers. *npj Biofilms and Microbiomes* 2020; **6**: 1–12.

50. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* 2021; **49**: D545–D551.

51. The UniProt Consortium, Bateman A, Martin M-J, Orchard S, Magrane M, Agivetova R, et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2020; **49**: D480–D489.

52. Sunagawa S, Acinas SG, Bork P, Bowler C, Tara Oceans Coordinators, Eveillard D, et al. Tara Oceans: towards global ocean ecosystems biology. *Nat Rev Microbiol* 2020; **18**: 428–445.

53. Lind AL, Pollard KS. Accurate and sensitive detection of microbial eukaryotes from whole

metagenome shotgun sequencing. *Microbiome* 2021; **9**: 58.

54. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; **9**: 357–359.

55. Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, et al. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* 2021; **6**: 3–6.

56. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 2021; **18**: 366–368.

57. Huang L, Zhang H, Wu P, Entwistle S, Li X, Yohe T, et al. dbCAN-seq: a database of carbohydrate-active enzyme (CAZyme) sequence and annotation. *Nucleic Acids Res* 2018; **46**: D516–D521.

58. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res* 2018; **46**: D624–D632.

59. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol* 2019; **37**: 420–423.

60. Karlicki M, Antonowicz S, Karnkowska A. Tiara: Deep learning-based classification system for eukaryotic sequences. *Bioinformatics* 2021; **38**: 344–350.

61. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015; **25**: 1043–1055.

62. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078–2079.

63. Graham ED, Heidelberg JF, Tully BJ. Potential for primary productivity in a globally-

distributed bacterial phototroph. *ISME J* 2018; **12**: 1861–1866.

64. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010; **11**: 119.

Figure Legends:

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

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

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

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

sample basis, a MAG might have several median *pN/pS* ratio from multiple samples. **B** The relationship between the selection pressure (median *pN/pS*) and the transposase abundance of a MAG, separated by depth. Only MAGs with ≥ 100 *pN/pS* were computed for median *pN/pS*.

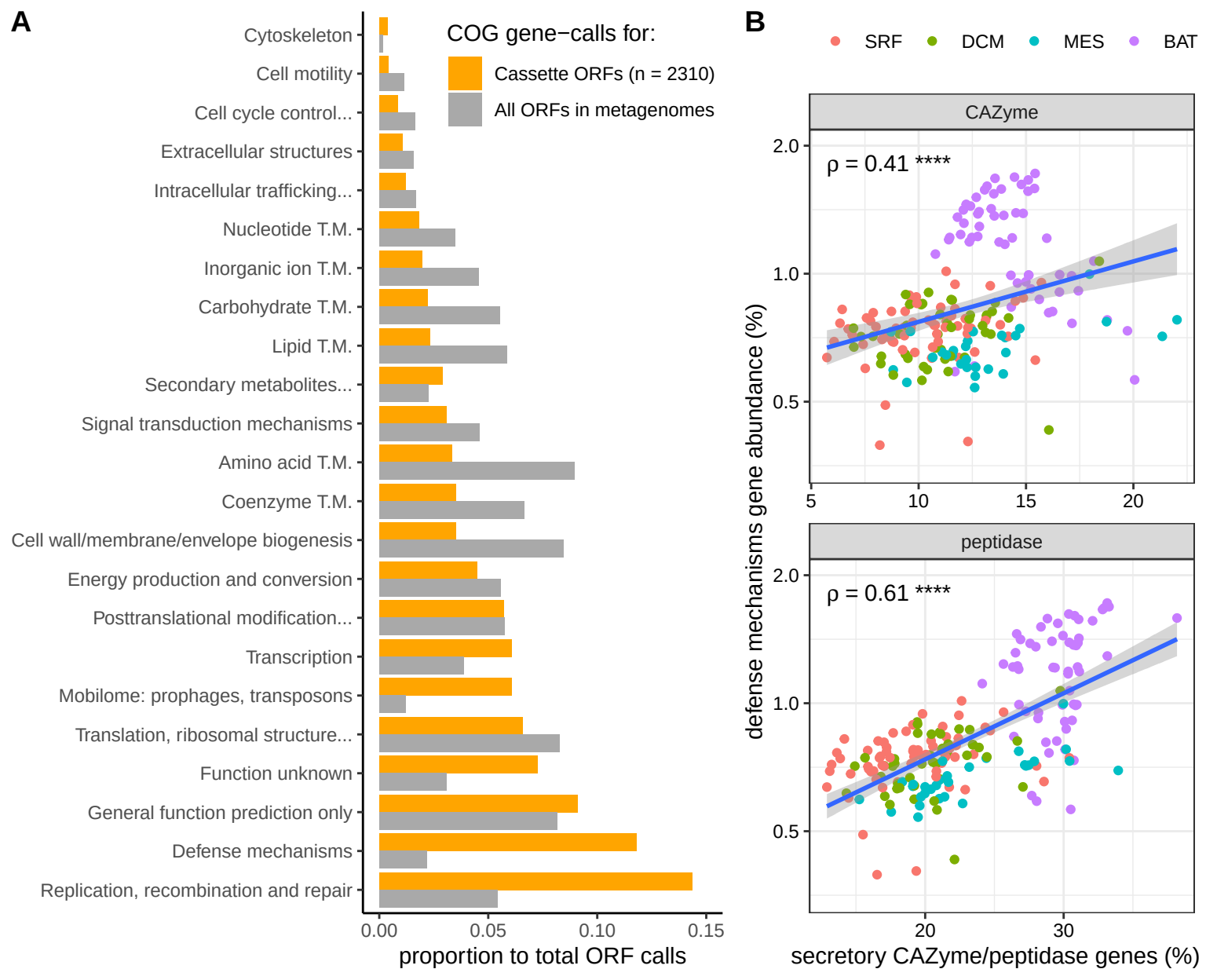
Fig. 5: Cassette sequences include high proportions of defense mechanism genes, which are more abundant in microbial communities that rely heavily on a particle-associated lifestyle. **A** The distribution of COG functional categories in cassette and non-cassette metagenome ORFs. *TM: transport and metabolism. 1951 cassette ORFs with COG function calls were identified from the 10 co-assembled *Tara* metagenomes, and 359 from the co-assembled Malaspina metagenome. 5 million ORFs with COG function calls were then sampled from the *Tara* Oceans and Malaspina metagenomes according to the ratio of identified cassettes described above. **B** The correlation between the abundance of secretory CAZyme/peptidase ORFs and defense mechanism ORFs in metagenomes. Samples from different depths are distinguished by different colors. The Spearman's correlation coefficients, ϱ, are shown on top left. **** indicates that $p < 0.0001$.

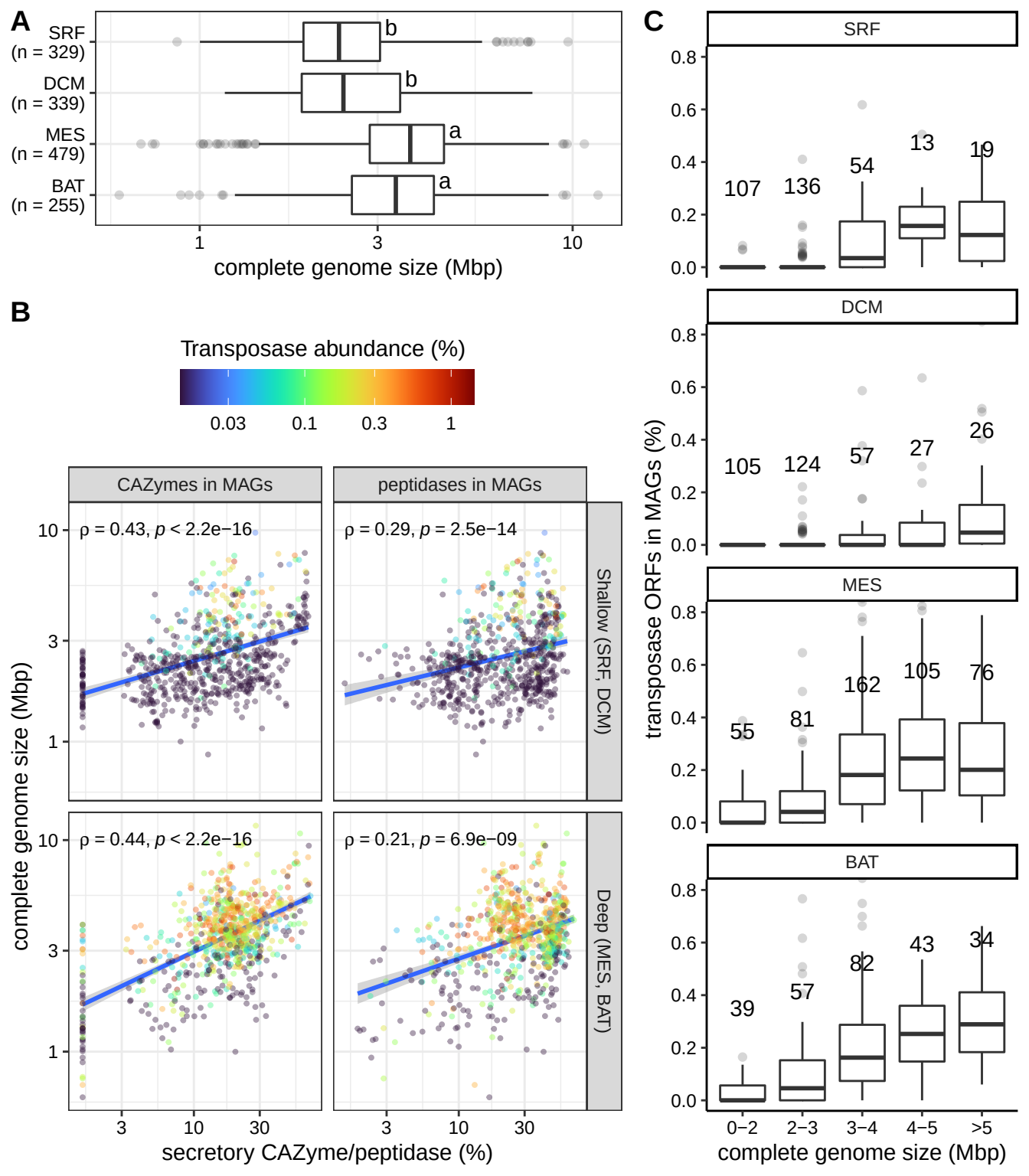
Fig. 6: Particle-associated and defense mechanism ORFs are highly expressed in deeper waters, but transposase are not. Log-transformed RNA/DNA ratios of target genes in each sample, separated by depth. 2-tailed *t*-tests were performed on log-transformed RNA/DNA ratios. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

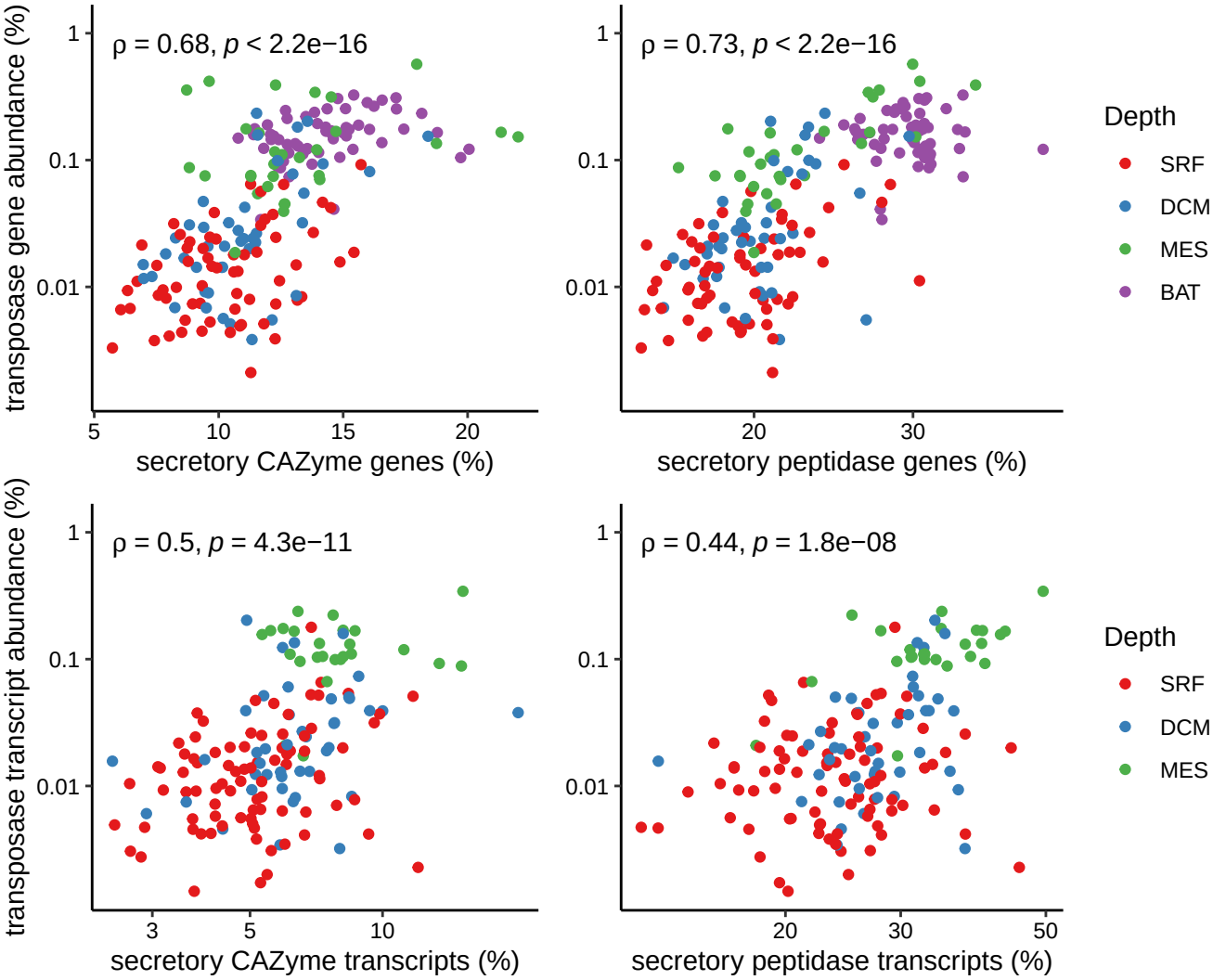
Fig. 7: Linking a particle-associated lifestyle to the high abundance of transposons in the deep ocean.

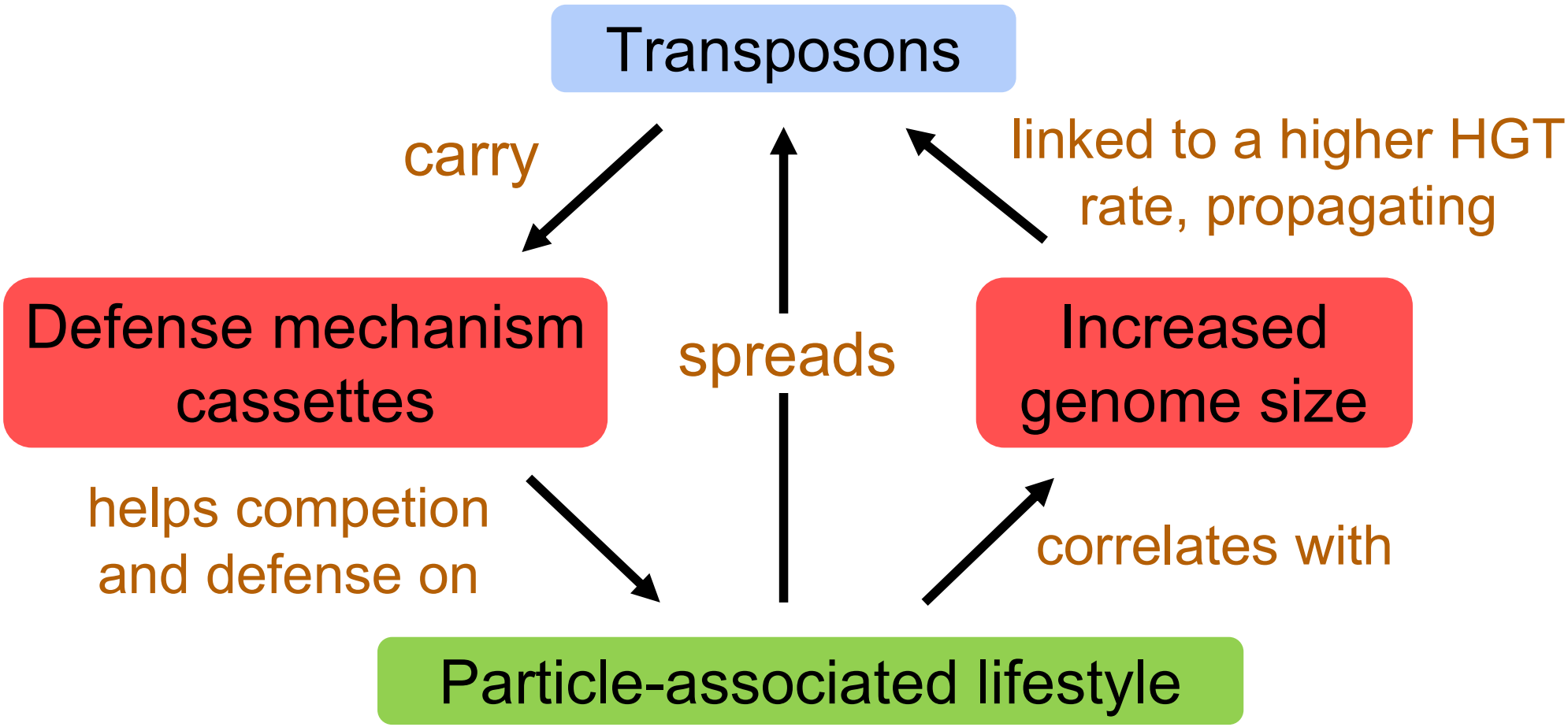
Table 1. Multiple Stepwise Regression for log-transformed transposase abundance (%) of 1402 MAGs. An ANOVA test was performed on each newly added covariate.

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

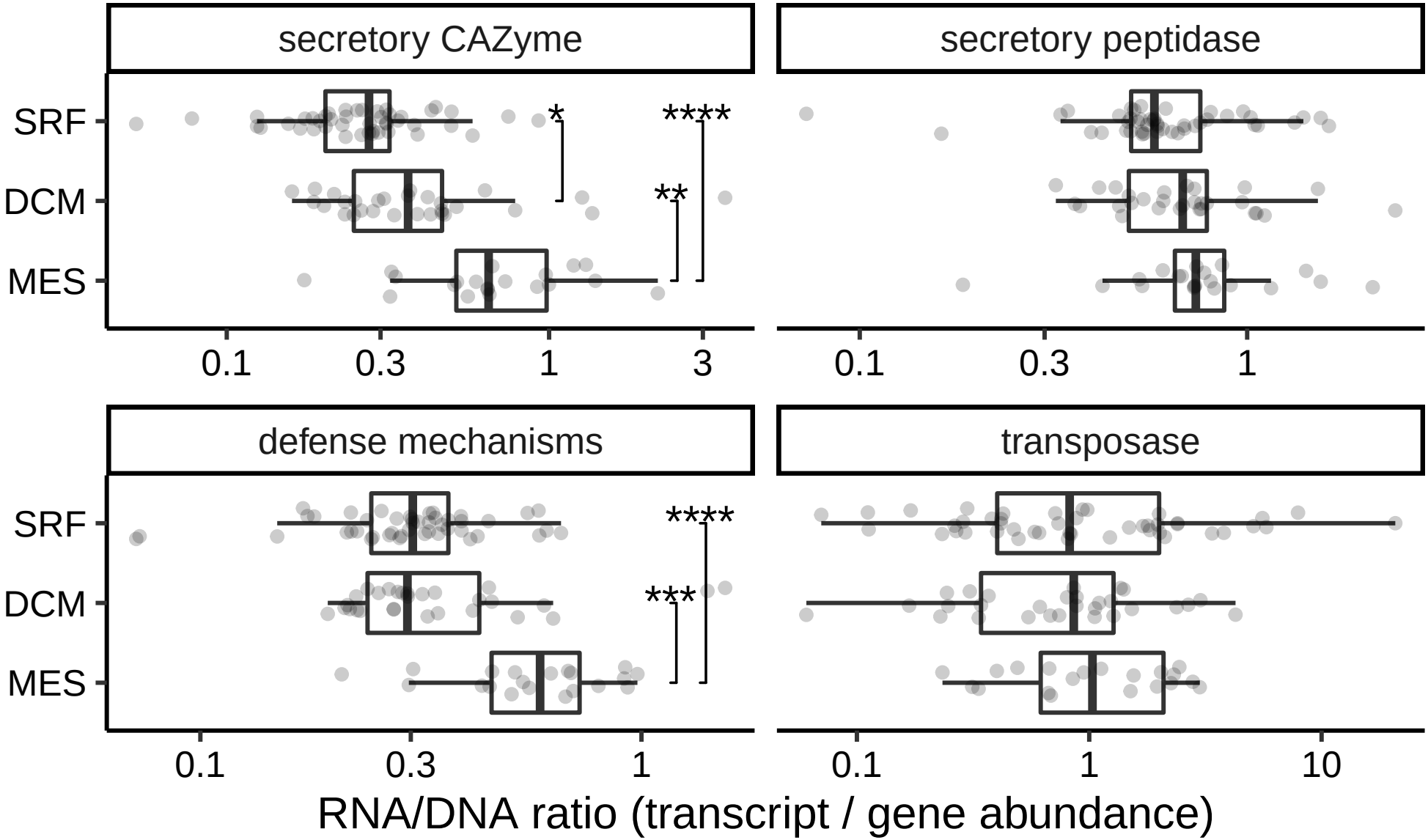

A
COG gene−calls for:
Cassette ORFs (n = 2310)
All ORFs in metagenomes
Cytoskeleton
Cell motility
Cell cycle control...
Extracellular structures
Intracellular trafficking...
Nucleotide T.M.
Inorganic ion T.M.
Carbohydrate T.M.
Lipid T.M.
Secondary metabolites...
Signal transduction mechanisms
Amino acid 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 (%)

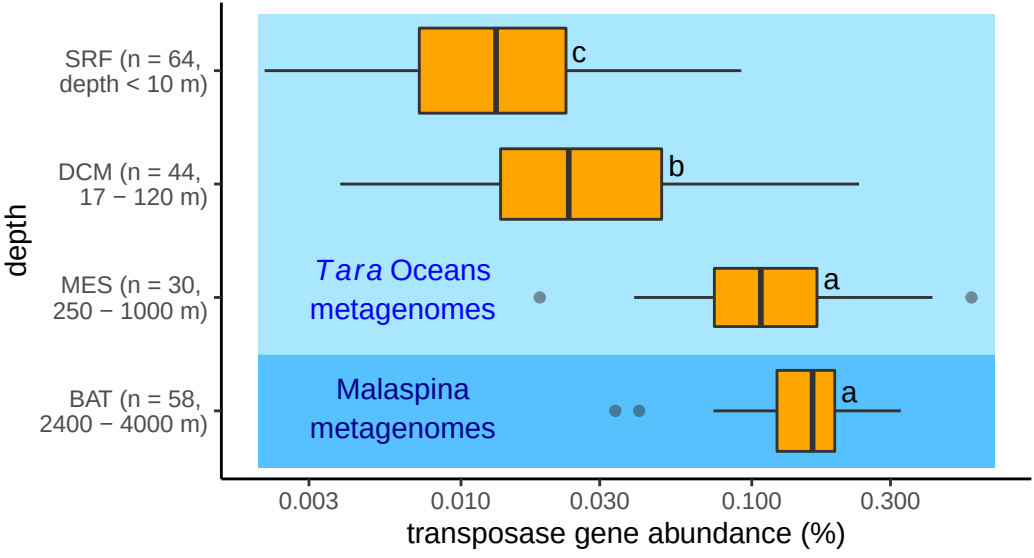

A
SRF
(n = 329)
DCM
(n = 339)
MES
(n = 479)
BAT
(n = 255)
b
b
a
a
1
3
10
complete genome size (Mbp)
B
Transposase abundance (%)
0.03
0.1
0.3
1
CAZymes in MAGs
peptidases in MAGs
ρ = 0.43, p < 2.2e−16
ρ = 0.29, p = 2.5e−14
ρ = 0.44, p < 2.2e−16
ρ = 0.21, p = 6.9e−09
Shallow (SRF, DCM)
Deep (MES, BAT)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

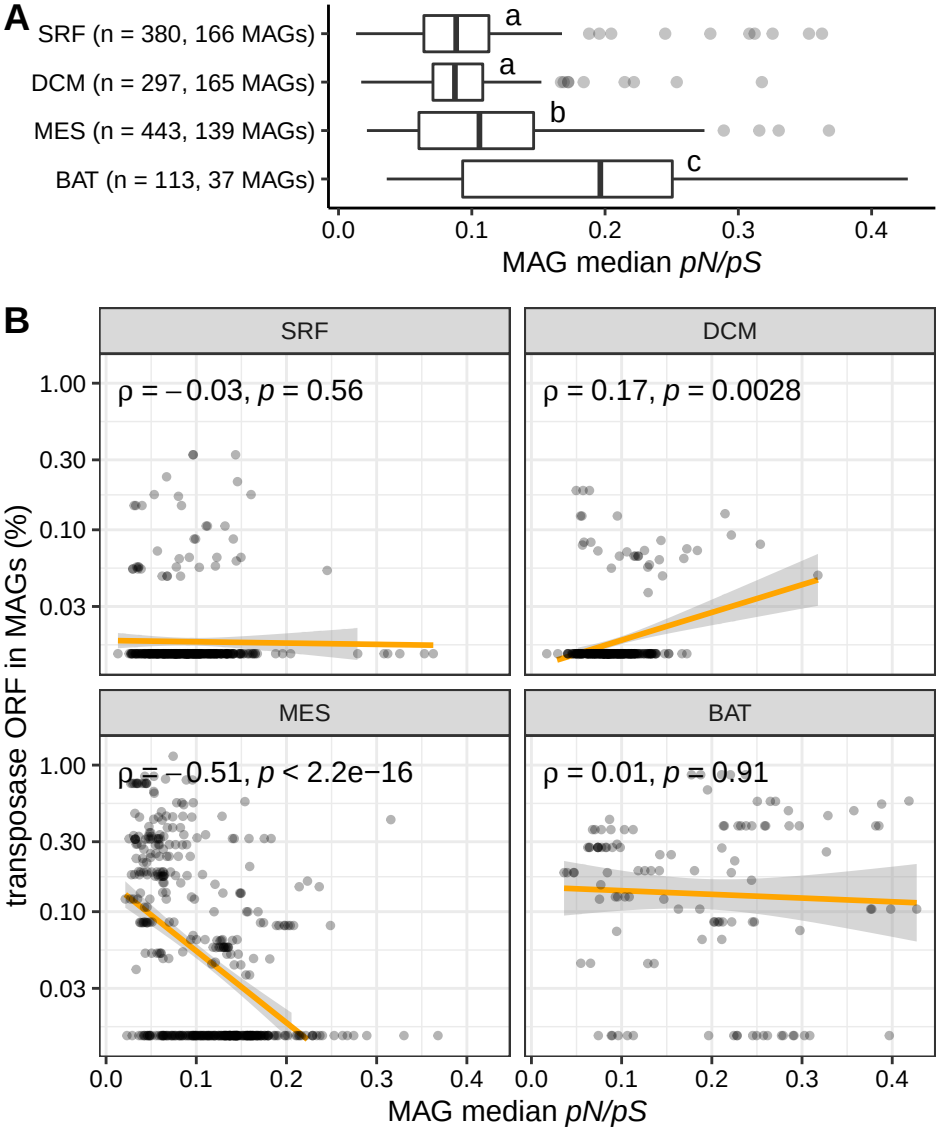

ρ = 0.68, p < 2.2e−16
ρ = 0.73, p < 2.2e−16
ρ = 0.5, p = 4.3e−11
ρ = 0.44, p = 1.8e−08
transposase gene abundance (%)
transposase transcript abundance (%)
secretory CAZyme genes (%)
secretory peptidase genes (%)
secretory CAZyme transcripts (%)
secretory peptidase transcripts (%)
Depth
SRF
DCM
MES
BAT



A deep-ocean microbial community


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


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 (%)


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