Resource conservation manifests in the genetic code

Liat Shenhav[1,2,*] and David Zeevi[1,*,^]

[1]Center for Studies in Physics and Biology, Rockefeller University, New York, NY, USA

[2]Department of Computer Science, University of California Los Angeles, Los Angeles, CA, USA

[*]denotes equal contribution

[^]dzeevi@rockefeller.edu

Abstract

Nutrient limitation is a strong selective force, driving competition for resources. However, much is unknown about how selective pressures resulting from nutrient limitation shape microbial coding sequences. Here, we study this 'resource-driven' selection using metagenomic and single-cell data of marine microbes, alongside environmental measurements. We show that a significant portion of the selection exerted on microbes is explained by the environment and is strongly associated with nitrogen availability. We further demonstrate that this resource conservation optimization is encoded in the structure of the standard genetic code, providing robustness against mutations that increase carbon and nitrogen incorporation into protein sequences. Overall, we demonstrate that nutrient conservation exerts a significant selective pressure on coding sequences and may have even contributed to the evolution of the genetic code.

Introduction

Nitrogen and carbon are major limiting factors in many ecosystems, with recent studies linking their availability to core genomic properties (*1*, *2*). In low-nitrogen environments there is a lower incorporation of nitrogen-rich side-chains into proteins, a strong A+T bias in nucleotide sequences, and smaller genome sizes (*3–5*). An opposite trend was shown for carbon (*2*) and indeed nitrogen and carbon concentrations are typically inversely correlated (*6*). These studies suggest a purifying selective pressure associated with resource conservation, which we term 'resource-driven' selection.

This resource-driven selection postulates that DNA mutations which result in excess incorporation of nutrients such as nitrogen and carbon into proteins, are disfavored. However, not all mutations have the same effect on protein sequences due to constraints imposed by the structure of the genetic code. The standard genetic code, common to virtually all of life on earth, is known to be highly efficient in mitigating deleterious effects of mistranslation errors and point mutations (*7*), specifically those leading to radical changes in amino acids. This error minimization is prominent among theories regarding the origin of the genetic code (*8–11*), which propose that the code had evolved through selection to minimize potential adverse effects of mutations on protein structure and function (*12–14*). To quantify code optimality, some theories provide structurally-informed amino acid metrics based on hydropathy and stereochemistry [e.g., the polar requirement scale (*11*) and hydropathy index (*15*)]. To our knowledge, an optimization of nutrient conservation in the genetic code has not been studied thus far.

Here, we use publicly available metagenomic and single-cell data from global marine habitats, coupled with their corresponding measurements of environmental conditions, to comprehensively characterize and quantify the selective pressure resulting from nutrient limitation. We estimate selection in 746 marine samples (**Fig. S1A**) and quantify its 'resource-driven' component, which we show to be a ubiquitous purifying selective force. We further demonstrate that the structure of the genetic code, known to minimize the effects of DNA mutations on the structure and function of proteins, also minimizes nutrient expenditure given a random mutation. We show that this optimization for resource conservation is orthogonal to known

44 mechanisms of error minimization in protein structure and function encoded in the genetic code. This
45 analysis exposes a hitherto unknown facet of the genetic code suggesting resource conservation as an
46 important contributor to its evolution.

47 **Results**

48 **Widespread purifying selection in the marine environment**

49 To comprehensively characterize how coding sequences of marine microbes are affected by resource
50 availability, we devised a computational pipeline that calculates selection metrics from marine
51 metagenomic samples (**Fig. S1B**). We downloaded 746 samples from Tara oceans (*16*) (n=136),
52 bioGEOTRACES (*17*) (n=480), and the Hawaii Ocean- (HOT; n=68); and Bermuda Atlantic- (BATS;
53 n=62) Time-series (*17*) [**Fig. S1A**; (*18*)]. We aligned these reads to the Ocean Microbiome Reference Gene
54 Catalog (OM-RGC) (*16*), a database of genes from marine environments that is accompanied by functional
55 information. We then searched for single nucleotide polymorphisms (SNPs) in genes that had sufficient
56 high-quality coverage [**Fig. S1B**; (*18*)]. Overall, we found 71,921,864 high-confidence SNPs (*18*), in a total
57 of 1,590,843 genes.

58 To quantify purifying selection on different gene functions, we annotated genes from the OM-RGC
59 database using either KEGG orthology [KO; (*19*)] or eggNOG orthologous groups [OG; (*18*, *20*)]. Using
60 called SNPs, we calculated for each orthologous group and in each sample, the ratio of nonsynonymous
61 polymorphisms which lead to a change in the coded amino acid (pN), to synonymous polymorphisms,
62 which maintain it (pS) [pN/pS; (*18*)]. Across all samples, we found pN/pS ratios to be close to zero with an
63 average of 0.074 in OGs (CI [0.072, 0.075]) and 0.079 in KOs (CI [0.077, 0.080]; **Fig. 1A**), indicating
64 widespread purifying selection in the marine environment. To corroborate the validity of calculating
65 selection metrics from metagenomic samples, we compared nonsynonymous mutations leading to

66 “conservative” amino acid substitutions to those leading to “radical” substitutions and found conservative
67 mutations to be significantly more common (**Supplementary text**).

68 **Significant resource-driven selection across marine microbial genes**

69 Based on recent studies (*2–5*), we hypothesized that nutrient availability is a central driver of this purifying
70 selection. We sought to quantify the environmental contribution to this selective pressure. We considered
71 environmental measurements taken alongside each sample, including depth, water temperature and salinity,
72 and the concentrations of nitrate, nitrite, oxygen, phosphate and silicate [**Fig. S2A-H**; (*18*)].

73 These environmental measurements presented consistent correlation patterns with the pN/pS of many
74 KEGG and eggNOG orthologs (**Fig. S3**). However, as they are also highly correlated with each other (**Fig.**
75 **S2I**), we cannot accurately estimate their individual effects. We therefore sought to estimate the fraction of
76 variance in pN/pS explained by the environment while accounting for these covariations. We used a linear
77 mixed model (LMM) with variance components (*18*), which allows to estimate the fraction of variance in
78 pN/pS ratios (dependent variable) that is explained by the environment (random effect) while controlling
79 for the complex correlation structure between the environmental parameters. We term the fraction of
80 variance in pN/pS explained by resource availability ‘environmental explained variance’ [EEV; (*18*)].
81 Across both KEGG and eggNOG orthologs we found that a significant fraction of the variance in pN/pS
82 can be attributed to the environment (**Fig. 1B,C, S4A,B**; Mann-Whitney *U* test $P<10^{-16}$), with nitrate being
83 more strongly correlated with pN/pS ratios than any other environmental parameter (**Fig. S5**; Kolmogorov-
84 Smirnov test; $P<10^{-30}$ for all comparisons). Examining typical DNA mutations and amino acid substitutions
85 in nitrate-rich versus nitrate-poor environments, we find that environmental nitrate is associated with
86 specific changes to both DNA and protein sequences, favoring lower nitrogen incorporation into protein
87 sequences when nitrate is scarce (**Supplementary text**).

This association between environmental measurements and the magnitude of purifying selection is significant even after controlling for potential confounders such as time or effective population size (**Fig. S4C,D**; Mann-Whitney *U* test $P<10^{-16}$; **Supplementary text**) as well as in specific environmental niches (**Fig. S6**; Mann-Whitney *U* test $P<10^{-20}$; **Supplementary text**). Additionally, these results were replicated using benchmarking data of assembled genomes from uncultivated single cells from three dominant lineages of the surface ocean [SAR-11, SAR-86 and Prochlorococcus; **Fig. S4E-G; Supplementary text**; (*21*)]. These validations demonstrate that the association between selective pressure and environmental conditions is robust to both data type and selection metric and is not confounded by population properties and clade-specific metabolism.

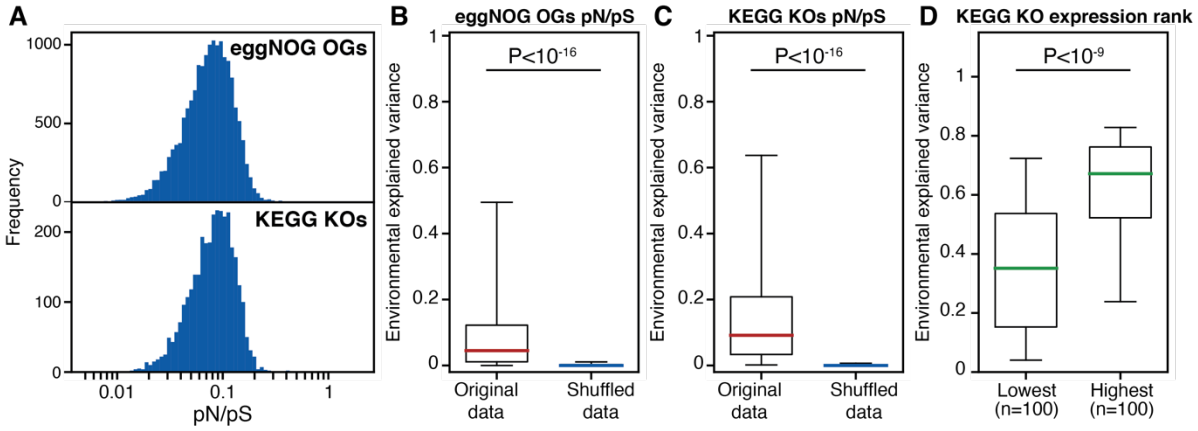


Figure 1. Analysis of pN/pS reveals resource-driven selection. *(A) Histogram of pN/pS ratios for eggNOG orthologous groups [OG (20); top] and KEGG orthologs [KO (19); bottom] across all marine samples. (B) Box plot (line, median; box, IQR; whiskers, 5th and 95th percentiles) of the variance of eggNOG OG pN/pS that was explained by the environment in a linear mixed model [LMM;(18)] as compared to the same data with shuffled labels. (C) Same as B, for KEGG KO pN/pS. P, Wilcoxon signed-rank test. (D) Box plot of variance in pN/pS explained by the environment in the 100 lowest and highest expressed KEGG KOs. P, Mann-Whitney U test.*

105 **Environmental association is stronger in resource-consuming genes**

106 With nitrate being the environmental factor most strongly associated with pN/pS, we sought to provide a
107 population-level perspective to the theory and experiments demonstrating selection against mutations that
108 increase the nitrogen requirements of cells, especially in nitrogen-limited conditions (*3*, *5*, *22*). This implies
109 stronger purifying selection in highly expressed genes (*3*, *4*), where one DNA mutation could translate to
110 thousands of proteins, each consuming more resources (illustrated in **Fig. S7**). We used an additional
111 expression dataset for marine microbial genes (*23*) to rank KEGG orthologs by their mean expression (*18*)
112 and found that the 100 most highly expressed KEGG KOs had a significantly higher EEV as compared to
113 the 100 least expressed ones (**Fig. 1D, S8**; Mann-Whitney *U* test $P<10^{-9}$). We replicated these results using
114 single-cell data pertaining to specific bacterioplankton lineages (Mann-Whitney *U* test $P<10^{-7}$).
115 Additionally, we found that extracellular protein-coding genes, where resources excreted from the cell
116 cannot be recycled, had significantly higher EEV as compared to other gene groups [**Fig. S9**; $P<0.05$; (*18*)].
117 This higher EEV in resource-consuming genes further strengthens our results regarding the breadth of
118 resource-driven selection.

119 **Resource-conservation as an optimization mechanism in the genetic code**

120 We observed selection against DNA mutations which result in excess incorporation of nutrients, such as
121 nitrogen and carbon, into proteins. However, not all mutations have the same effect on protein sequences
122 due to constraints imposed by the structure of the genetic code. Indeed, the genetic code was shown to
123 minimize the impact of point mutations on protein structure and function (*12–14*). We hypothesized that
124 the genetic code also minimizes the impact of point mutations on nutrient incorporation into proteins.
125 Specifically, that the genetic code acts as a buffer between the DNA level, where mutations occur, and the
126 protein level, where resource-driven selection is exerted, such that the expected cost of a random DNA
127 mutation, in terms of added resources to proteins, is minimized.

To rigorously test this hypothesis, we first defined, for each element *e* (e.g., carbon, nitrogen), a function quantifying the cost of a single mutation as the added number of amino acid atoms resulting from it. For example, a missense mutation from codon CCA to CGA results in an amino acid substitution from proline to arginine, with an increase of one carbon and three nitrogen atoms, setting the nitrogen cost of such mutation to 3 and the carbon cost to 1 (**Fig. 2A**).

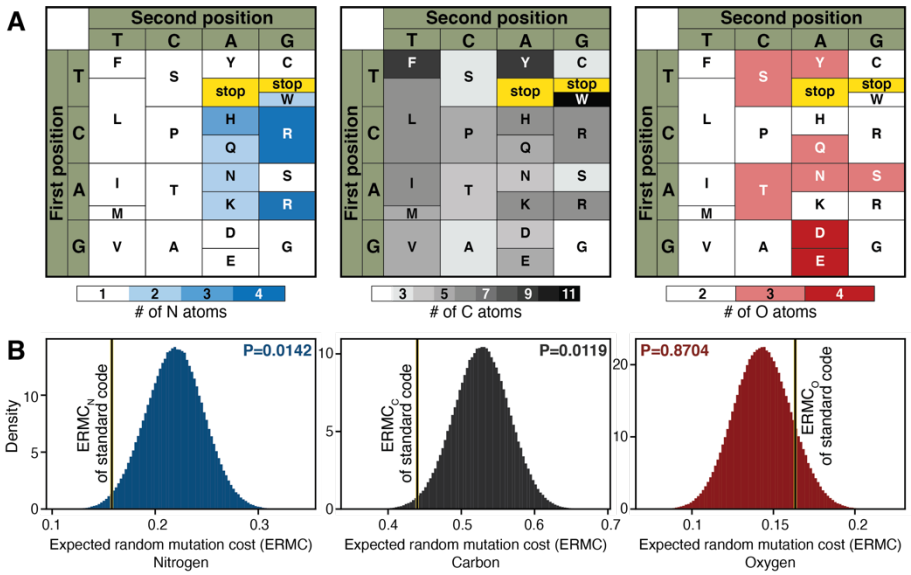


Figure 2. Resource-conservation is facilitated by the genetic code. *(A) Nitrogen (left), carbon (center) and oxygen (right) content of different amino acids depicted along their positions in the standard genetic code. (B) Histograms of the expected random mutation cost (ERMC), in 1,000,000 random permutations of the genetic code for nitrogen (left, blue), carbon (center, black) and oxygen (right, red). Black-yellow bar marks the ERMC of the standard genetic code, ERMC(Vs), for each of the elements.*

To estimate the cost of random mutations across the entire genetic code, we calculated, per element (e.g., nitrogen), the Expected Random Mutation Cost (ERMC):

$$ERMC_e(V) = \sum_{v,v' \in V_S} P(v)P(mut(v,v'))c_e^+(v,v'),$$

Where V is a genetic code; P(v) the abundance of codon v, here calculated from all marine samples; P(mut(v,v′)) the probability of mutation from codon v to codon v′, set to be the relative abundance of the single nucleotide mutation driving this codon change (e.g., for mutation from GCA to CCA, we use the

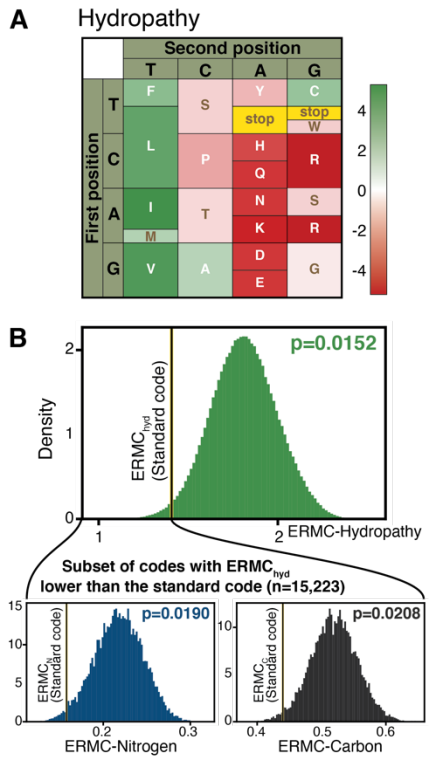
145 abundance of G-to-C transversions); and the cost function $c_e^+(v,v')$ the number of atoms of element *e* added
146 post-mutation (*18*). For the standard genetic code, and codon abundances and mutation rates calculated
147 from marine microbes, we report an ERMC of 0.440, 0.158 and 0.163 for carbon, nitrogen and oxygen,
148 respectively, corresponding to an average increase of this number of atoms per random mutation (**Fig. 2B**).

149 To check if the standard genetic code, along with codon abundances and mutation rates, is indeed robust to
150 resource-consuming mutations, we compared it to other hypothetical codes. We simulated alternative
151 genetic codes by randomizing the first and second position in all codons one million times, to create a null
152 distribution of ERMC (*18*). We found that the standard genetic code, common to most life forms, is
153 parsimonious in terms of carbon and nitrogen utilization, given a random mutation. This is marked by a
154 significantly low ERMC for nitrogen (**Fig. 2B**; $ERMC_N$ P=0.0142) and carbon ($ERMC_C$ P=0.0119), but not
155 oxygen ($ERMC_O$ P=0.8704).

156 We compared the extent of robustness against an addition of carbon or nitrogen to well-studied structure
157 and function error minimization mechanisms (*12–14*). We calculated ERMC for changes in hydropathy
158 (*15*) and polar requirement [PR; (*11*)], both structurally-informed amino acid properties used to determine
159 code error minimization (*18*). We found that these optimization mechanisms are of a similar magnitude to
160 those shown here for nitrogen and carbon [**Fig. S10C,D**; $ERMC_{hyd}$ P=0.0152; $ERMC_{PR}$ P=0.0137; (*18*)].

161 We next sought to confirm that optimization for carbon and nitrogen is not confounded by these known
162 optimization mechanisms. To this end, we devised a hierarchical model which examined the subset of
163 genetic codes (out of 1M hypothetical codes tested) that have a lower ERMC than the standard code for PR
164 or hydropathy, and tested whether this subset is also optimized for nitrogen or carbon (*18*). If nutrient
165 optimization exists separately from structural optimization, we expect the standard code to be optimized
166 for carbon and nitrogen even as compared to this subset, featuring significantly lower $ERMC_N$ and $ERMC_C$.
167 We find that out of 15,223 hypothetical codes that have $ERMC_{hyd}$ lower than the standard code, only 270
168 have lower $ERMC_N$ (**Fig. 3, S10E**; P=0.019) and only 249 have lower $ERMC_C$ (**Fig. 3, S10E**; P=0.0208).
169 Similarly, out of the 13,729 hypothetical codes that have $ERMC_{PR}$ lower than the standard code, only 83

170 have lower $ERMC_N$ (**Fig. S10F**; P=0.0058) and only 442 have lower $ERMC_C$ (**Fig. S10F**; P=0.037). This
171 is in contrast with the observed overlap between hydropathy and PR: out of the 15,223 hypothetical codes
172 that have $ERMC_{hyd}$ lower than the standard code, 6,736 have lower $ERMC_{PR}$ (**Fig. S10E**; P=0.4425). These
173 results indicate that the detected carbon and nitrogen optimization is not confounded by previously reported
174 optimization properties such as hydropathy and PR.



175

176 ***Figure 3. Optimization for carbon and nitrogen is not confounded by hydropathy.*** *(A) Hydropathy of*
177 *different amino acids depicted along their positions in the standard genetic code. (B; top) As in Fig. 2B,*
178 *for hydropathy; (B; bottom) histograms of ERMC for nitrogen (left) and carbon (right) for the subset of*
179 *hypothetical genetic codes with ERMC lower than the standard code for hydropathy. P, permutation test.*

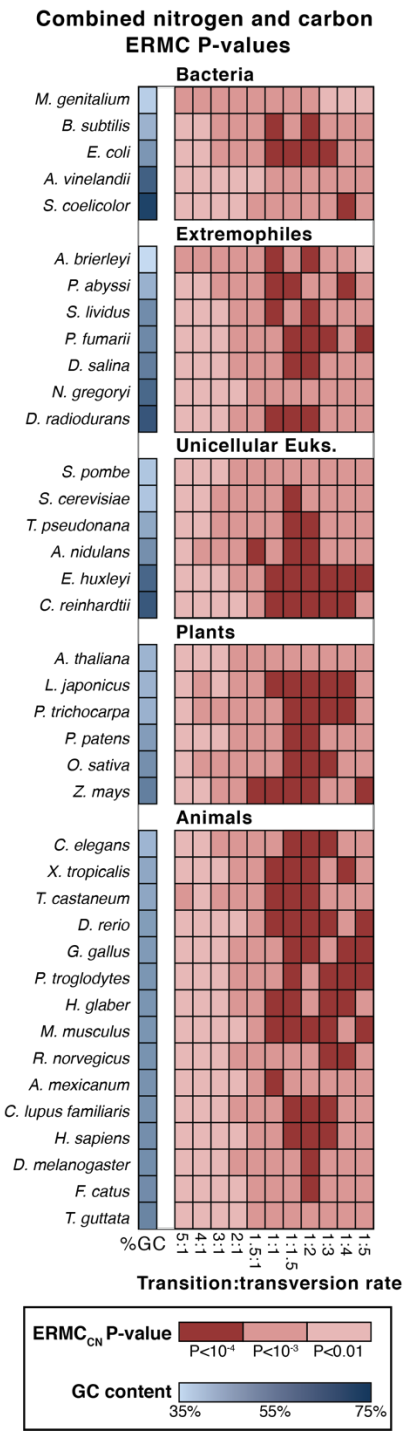
180 Remarkably, only 128 out of 1M randomized genetic codes were better than the standard code in
181 conservation of nitrogen and carbon together (**Fig. S10G,H**; $ERMC_{CN}$ P=1.28×10^{-4}). We note that this
182 number is significantly smaller than the number of hypothetical codes expected to have both a lower
183 $ERMC_C$ and $ERMC_N$ (Chi-square test of independence P=0.0013; **Table S1**). This is possibly driven by a

184 small overlap between the positions of high-nitrogen and high-carbon amino acids. This property of the
185 standard code potentially enables concurrent optimization for both carbon and nitrogen. These results
186 highlight a new optimization principle of the genetic code that is of similar magnitude- and independent of-
187 previously proposed principles.

188 **The genetic code facilitates resource conservation across kingdoms**

189 To show that the resource-robustness of the genetic code was not limited to our dataset and analytic
190 approach, we calculated the ERMC of 187 strains of genera *Prochlorococcus* and *Synechococcus*. We
191 computed codon abundances and mutation rates using published protein-coding sequences (*1*) and the
192 accepted transition:transversion rate of 2:1 (*18*, *24*). We corroborated significant conservation of carbon,
193 nitrogen, and both elements combined (**Fig. S11A**; $ERMC_C$ mean P=0.013, P=0.020; $ERMC_N$ mean
194 P=0.049, P=0.032; $ERMC_{CN}$ P=0.0004, P=0.0007 for *Prochlorococcus* and *Synechococcus*, respectively).

195 To explore whether this nutrient conservation optimization mechanism of the standard code extends across
196 organisms, we performed a similar calculation using codon abundances from 39 organisms across all
197 domains of life, including all human protein-coding sequences, and a range of transition:transversion rates
198 (*18*). Similarly to marine microbes, we find that the standard code is optimized in terms of resource
199 utilization for all tested organisms, marked by a significantly low ERMC for nitrogen and carbon combined,
200 across all transition:transversion rates (P<0.01, **Fig. 4**). Moreover, we find significant optimization even in
201 the theoretical case where all codon abundances are the same (**Fig. S11B**). The codon abundances of a great
202 majority of organisms also demonstrate significantly low ERMC for nitrogen (**Fig. S11C**) and carbon (**Fig.
203 S11D**), for a wide range of transition:transversion rates. These results indicate that resource optimization
204 in the genetic code transcends taxonomy, codon choices, and mutation rates.



205

206 ***Figure 4. The genetic code is optimized for resource conservation across organisms.*** *Heat map of*
207 *ERMC$_{CN}$ P-values across 39 organisms and 11 transition:transversion rates. Organisms in each group are*
208 *ordered by the GC content of their coding sequences.*

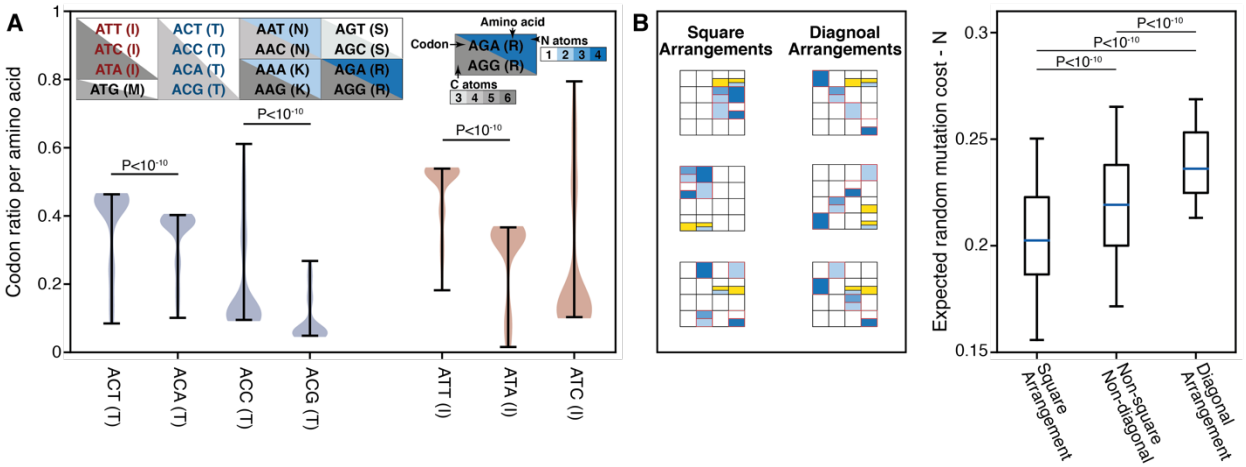
209 **Resource conservation may bias codon usage**

210 We hypothesized that variation in the third codon position for a single amino acid and may also be biased
211 due to the differential cost of a random mutation for each codon. We therefore focused on codon usage of
212 the amino acid threonine. We note that a C-to-G transversion in the second position for codons ACT and
213 ACC yields serine (AGT and AGC, respectively), but the same mutation for codons ACA and ACG yields
214 arginine (AGA and AGG; **Fig. 5A**, inset). Arginine has higher carbon and nitrogen content than serine. We
215 thus hypothesized that for a cell to conserve nutrients in case of a random mutation, codons ACT and ACC
216 should have a higher abundance than codons ACA and ACG, respectively, given a known GC-bias that is
217 a genomic quantity. To test this hypothesis, we examine codon usage in 187 *Prochlorococcus* and
218 *Synechococcus* strains, and show a significantly higher use of ACT as compared to ACA (**Fig. 5A**;
219 Wilcoxon signed-rank test $P<10^{-20}$) and ACC as compared to ACG (**Fig. 5A**; Wilcoxon signed-rank test
220 $P<10^{-20}$). Similarly, isoleucine codon ATT has higher abundance as compared to ATA (**Fig. 5A**; Wilcoxon
221 signed-rank test $P<10^{-20}$). These results demonstrate that resource conservation may guide codon usage and
222 thereby affect not only protein sequence but also cellular translation efficiency.

223 **Structural principles drive optimization in the genetic code**

224 We next examined the organizing principles underlying code resource optimization. We observed that
225 codons of the nitrogen-rich amino acids histidine, glutamine, asparagine, lysine and arginine span only two
226 nucleotides in their first position and two in their second position. We define this organization to be a
227 'square' arrangement, and hypothesize that it amplifies nitrogen conservation [**Fig. 5B**; (*18*)]. Specifically,
228 in the square arrangement, some codons require at least two mutations to increase the number of nitrogen
229 atoms (e.g., those coding for alanine and valine). This is in contrast to other hypothetical arrangements,
230 including a 'diagonal' one in which nitrogen-rich amino acid codons span all possible nucleotides in the
231 first and second positions [**Fig. 5B**; (*18*)]. We suggest that this arrangement would be nutrient-wasteful, as
232 in such arrangements a single mutation in either the first or second codon position could increase the

233 nitrogen content of a protein. To test this, we generated 10,000 hypothetical codes, with 220 arrangements
234 happening to embody a square structure, and 127 a diagonal one. We found that, when compared to all
235 other possible arrangements, square arrangements present a significantly lower $ERMC_N$ while diagonal
236 arrangements exhibit a significantly higher $ERMC_N$ (**Fig. 5B**; Mann-Whitney *U* test $P<10^{-10}$ for both). This
237 result demonstrates that resource optimization in the standard code is driven by structural principles,
238 perhaps underlying the optimization observed across kingdoms.



239

240 ***Figure 5. Structural properties and codon usage bias underlying optimality in the genetic code.*** *(A) Violin*
241 *plot of codon usage among 187 species of Prochlorococcus and Synechococcus showing significant*
242 *preference of threonine codons ACT and ACC as compared to ACA and ACG, and of isoleucine codon ATT*
243 *as compared to ACA. P, Wilcoxon signed-rank test. (B) Box plot (line, median; box, IQR; whiskers, 5th and*
244 *95th percentiles) of $ERMC_N$ of square arrangements (left) and diagonal arrangements [right, (18)], as*
245 *compared to all other arrangements (center) out of 10,000 hypothetical arrangements. P, Mann-Whitney*
246 *U test.*

247 **Discussion**

248 In this work, we characterize and quantify the selective forces exerted by nutrient availability on protein
249 coding genes in marine environments. We provide a data-driven, population-level perspective and show
250 that resource-driven selection is a ubiquitous force. We further show that a significant portion of DNA
251 mutations may not result in increased nutrient incorporation into protein sequences, due to the structure of

252 the standard genetic code. We provide evidence that the genetic code, shared among most lifeforms, along
253 with organismal codon frequencies, is robust against the incorporation of additional atoms of nitrogen and
254 carbon given a random mutation. We quantitatively demonstrate that this nutrient conservation is not
255 confounded by known error minimization mechanisms of the genetic code.

256 In light of these results, we hypothesize that resource-driven selection is equally exerted on all parts of the
257 protein-coding gene. This sets it apart from selection to maintain the structural integrity of a protein or the
258 function of its active site, which occur predominantly in structurally important regions (*25*). Thus,
259 accounting for resource-driven selection may improve the identification of alternative translation start sites,
260 alternatively spliced introns, or readthrough stop codons, as intermittently translated regions of the protein
261 may be under weaker resource-driven selection as compared to constitutively translated ones (*26*).

262 We note that in addition to resource-driven selection there are other forces associated with the environment,
263 driving both selective and neutral mechanisms. Genomic GC-content is associated with exposure to UV
264 radiation (*27*) and nitrogen availability (*2*, *28*) yet nonsynonymous mutations, especially in codons
265 encoding structurally and functionally important protein positions, are likely to be less affected by GC-bias.
266 Furthermore, codon usage bias was also shown to be associated with the environment (*29*).

267 We show that the genetic code optimizes nutrient conservation, calling upon an examination of code
268 evolution theories. Hypotheses regarding the origin of the genetic code include stereochemical affinity
269 between a codon or anticodon and their amino acid (*11*, *30*); a frozen accident theory (*31*); a co-evolution
270 of the genetic code with the emergence of amino acid-producing biosynthetic pathways (*9*); and an early
271 fixation of an optimal genetic code, suggesting selection for error minimization (*10*, *32*). Our results are in
272 line with the latter theory and suggest that the genetic code may have also been optimized for nutrient
273 conservation. This implies that the genetic code can be viewed as a buffer between the evolutionary forces
274 of mutation and selection, the former occurring in DNA sequences and the latter predominantly in proteins.
275 In the case of nutrient conservation, many DNA mutations do not result in the incorporation of additional
276 nutrients into proteins, and are thus not selected against, which may allow more 'freedom' to explore the

mutation-space. Notwithstanding, while optimized for resource conservation, the genetic code is also near-immutable. This is evident in heterotrophic eukaryotes, which while being net-nitrogen producers, still harbor a nitrogen-conservative code.

Since the genetic code is common to virtually all organisms, we can assume that it was already fixed in its current form, or one very similar, in the last universal common ancestor. Thus, if the genetic code evolved to optimize resource conservation, it may reflect selective pressures in the beginning of life on earth. An organism harboring a nitrogen- and carbon-efficient genetic code would have had a selective advantage over its peers, especially in the absence of fully evolved DNA mutation repair mechanisms.

References

1. P. M. Berube, A. Rasmussen, R. Braakman, R. Stepanauskas, S. W. Chisholm, Emergence of trait variability through the lens of nitrogen assimilation in Prochlorococcus. *Elife*. **8** (2019), doi:10.7554/eLife.41043.

2. D. R. Mende *et al.*, Environmental drivers of a microbial genomic transition zone in the ocean's interior. *Nature Microbiology*. **2** (2017), pp. 1367–1373.

3. J. J. Grzymski, A. M. Dussaq, The significance of nitrogen cost minimization in proteomes of marine microorganisms. *ISME J.* **6**, 71–80 (2012).

4. H. Akashi, T. Gojobori, Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 3695–3700 (2002).

5. F. L. Hellweger, Y. Huang, H. Luo, Carbon limitation drives GC content evolution of a marine bacterium in an individual-based genome-scale model. *The ISME Journal*. **12** (2018), pp. 1180–1187.

6. E. R. M. Druffel, S. Griffin, A. I. Coppola, B. D. Walker, Radiocarbon in dissolved organic carbon of the Atlantic Ocean. *Geophys. Res. Lett.* **43**, 5279–5286 (2016).

7. R. T. Hinegardner, J. Engelberg, RATIONALE FOR A UNIVERSAL GENETIC CODE. *Science*. **142**, 1083–1085 (1963).

8. G. Gamow, Possible Relation between Deoxyribonucleic Acid and Protein Structures. *Nature*. **173**, 318–318 (1954).

9. J. T. Wong, A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **72**, 1909–1912 (1975).

10. S. J. Freeland, L. D. Hurst, The genetic code is one in a million. *J. Mol. Evol.* **47**, 238–248 (1998).

11. C. R. Woese, D. H. Dugre, W. C. Saxinger, S. A. Dugre, The molecular basis for the genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **55**, 966–974 (1966).

12. S. E. Massey, A neutral origin for error minimization in the genetic code. *J. Mol. Evol.* **67**, 510–516 (2008).

13. A. S. Novozhilov, E. V. Koonin, Exceptional error minimization in putative primordial genetic codes. *Biol. Direct*. **4**, 44 (2009).

14. N. Torabi, H. Goodarzi, H. Shateri Najafabadi, The case for an error minimizing set of coding amino acids. *J. Theor. Biol.* **244**, 737–744 (2007).

15. D. Haig, L. D. Hurst, A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* **33**, 412–417 (1991).

16. S. Sunagawa *et al.*, Structure and function of the global ocean microbiome. *Science*. **348**, 1261359 (2015).

17. S. J. Biller *et al.*, Marine microbial metagenomes sampled across space and time. *Sci Data*. **5**,

180176 (2018).

18. Materials and methods are available as supplementary materials at the Science website.

19. M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

20. J. Huerta-Cepas *et al.*, Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).

21. M. G. Pachiadaki *et al.*, Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell*. **179** (2019), pp. 1623–1635.e11.

22. A. Dufresne, L. Garczarek, F. Partensky, Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol.* **6**, R14 (2005).

23. B. C. Kolody *et al.*, Diel transcriptional response of a California Current plankton microbiome to light, low iron, and enduring viral infection. *ISME J.* (2019), doi:10.1038/s41396-019-0472-2.

24. E. Urbach, D. J. Scanlan, D. L. Distel, J. B. Waterbury, S. W. Chisholm, Rapid Diversification of Marine Picophytoplankton with Dissimilar Light-Harvesting Structures Inferred from Sequences of Prochlorococcus and Synechococcus (Cyanobacteria). *Journal of Molecular Evolution*. **46** (1998), pp. 188–201.

25. J. Liu, Y. Zhang, X. Lei, Z. Zhang, Natural selection of protein structural and functional properties: a SNP perspective. *Genome Biology*. **9** (2008), p. R69.

26. R. W. Read *et al*., Nitrogen cost minimization is promoted by structural changes in the transcriptome of N-deprived Prochlorococcus cells. *ISME J.* **11**, 2267–2278 (2017).

27. C. E. Singer, B. N. Ames, Sunlight ultraviolet and bacterial DNA base ratios. *Science*. **170**, 822–825 (1970).

28. C. E. A. Mcewan, D. Gatherer, N. R. Mcewan, Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. *Hereditas*. **128**, 173–178 (1998).

29. M. Botzman, H. Margalit, Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biol.* **12**, R109 (2011).

30. M. Yarus, J. J. Widmann, R. Knight, RNA–Amino Acid Binding: A Stereochemical Era for the Genetic Code. *J. Mol. Evol.* **69**, 406 (2009).

31. F. H. Crick, The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379 (1968).

32. S. J. Freeland, R. D. Knight, L. F. Landweber, L. D. Hurst, Early fixation of an optimal genetic code. *Mol. Biol. Evol.* **17**, 511–518 (2000).

33. D. Zeevi *et al*., Structural variation in the gut microbiome associates with host health. *Nature*. **568**, 43–48 (2019).

34. 1000 Genomes Project Consortium, G. R. Abecasis *et al*., A map of human genome variation from population-scale sequencing. *Nature*. **467**, 1061–1073 (2010).

35. J. Huerta-Cepas *et al.*, eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

36. T. Seemann, Prokka: rapid prokaryotic genome annotation. Bio-Informatics 30: 2068--2069 (2014).

37. A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).

38. J. H. Ward, Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **58**, 236–244 (1963).

39. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

40. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. **30**, 1312–1313 (2014).

41. N. Goldman, Z. Yang, A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736 (1994).

42. R. Schlitzer *et al*.,The GEOTRACES Intermediate Data Product 2017. *Chem. Geol.* **493**, 210–223 (2018).

43. J. Honaker, G. King, M. Blackwell, Others, Amelia II: A program for missing data. *J. Stat. Softw.* **45**, 1–47 (2011).

44. A. L. Price, N. A. Zaitlen, D. Reich, N. Patterson, New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).

45. S. Schloissnig *et al.*, Genomic variation landscape of the human gut microbiome. *Nature*. **493**, 45–50 (2013).

46. M. R. E. Symonds, S. P. Blomberg, A Primer on Phylogenetic Generalised Least Squares. *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology* (2014), pp. 105–130.

47. M. Westoby, M. R. Leishman, J. M. Lord, On misinterpreting thephylogenetic correction'. *J. Ecol.* **83**, 531–534 (1995).

48. F. E. Rheindt, T. U. Grafe, Rapidly evolving traits and the comparative method: how important is testing for phylogenetic signal? *Evol. Ecol.* (2004) (available at http://www.evolutionary-ecology.com/abstracts/v06/1507.html).

49. J. Wang, L. Raskin, D. C. Samuels, Y. Shyr, Y. Guo, Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics*. **31**, 318–323 (2015).

50. J. Athey *et al.*, A new and updated resource for codon usage tables. *BMC Bioinformatics*. **18**, 391 (2017).

51. J. Zhang, Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *J. Mol. Evol.* **50**, 56–68 (2000).

392 52. S. J. Giovannoni *et al*., Genome streamlining in a cosmopolitan oceanic bacterium. *Science*. **309**,
393 1242–1245 (2005).

394 53. F. L. Hellweger, E. van Sebille, N. D. Fredrick, Biogeographic patterns in ocean microbes emerge in
395 a neutral agent-based model. *Science*. **345**, 1346–1349 (2014).

396 54. B. K. Swan *et al*., Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in
397 the surface ocean. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 11463–11468 (2013).

398 55. R. Milo, What is the total number of protein molecules per cell volume? A call to rethink some
399 published values. *Bioessays*. **35**, 1050–1055 (2013).

400 56. Y. Taniguchi, Quantifying E. coli Proteome and Transcriptome with Single-molecule Sensitivity in
401 Single Cells. *Seibutsu Butsuri*. **51** (2011), pp. 136–137.

402 57. T. R. Anderson, P. J. le B. Williams, A one-dimensional model of dissolved organic carbon cycling
403 in the water column incorporating combined biological-photochemical decomposition. *Global*
404 *Biogeochemical Cycles*. **13** (1999), pp. 337–349.

405 58. B. Charlesworth, Effective population size and patterns of molecular evolution and variation. *Nat.*
406 *Rev. Genet.* **10**, 195–205 (2009).

407 59. M. Lynch, J. S. Conery, The evolutionary fate and consequences of duplicate genes. *Science*. **290**
408 (2000), pp. 1151–1155.

409 60. A. L. Hughes, R. Friedman, P. Rivailler, J. O. French, Synonymous and nonsynonymous
410 polymorphisms versus divergences in bacterial genomes. *Mol. Biol. Evol.* **25**, 2199–2209 (2008).

411 61. J. H. McDonald, M. Kreitman, Adaptive protein evolution at the Adh locus in Drosophila. *Nature*.
412 **351**, 652–654 (1991).

Acknowledgements: We thank Tal Korem and Nimrod D. Rubinstein for discussions, suggestions and help in analyses. We thank Jean-Pierre Eckmann, Otto X. Cordero, Itzhak Mizrahi and Niv Antonovsky for their suggestions. **Funding:** DZ was funded by the James S. McDonnell Foundation. **Authors contributions:** LS and DZ conceived the project, performed analyses and wrote this manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data is available in the manuscript, supplementary materials and referenced datasets. Code is available at: https://github.com/zeevilab/resource-conservation

Supplementary Materials

Materials and Methods

Supplementary text

Fig S1 – S16

Table S1 – S2

References (33-61)

426 **Materials and Methods**

427 Marine microbiome samples

428 Marine samples collected with Tara oceans (*16*), bioGEOTRACES (*17*), the Hawaii Ocean Timeseries
429 (HOT) and the Bermuda Atlantic Timeseries (BATS) (*17*) were downloaded from ENA with accessions
430 ENA:PRJEB1787 (TARA oceans prokaryotic fraction), ENA:PRJNA385854 (bioGEOTRACES) and
431 ENA:PRJNA385855 (HOT/BATS), each sample with a minimum of 5 million reads. For all samples for
432 which a size parameter was available, we selected the 0.22µm+ size category (typically between 0.22 and
433 5µm). This includes TARA ocean samples (prokaryote size) and data pertaining to expression (*22*). In
434 samples from the bioGEOTRACES, HOT and BATS, we used the available size fraction, larger than
435 0.2µm.

436 Mapping of Illumina reads to reference gene sequences

437 Samples were mapped to nucleotide sequences from the Ocean Microbiome Reference Gene Catalog
438 (OM-RGC) (*16*) using bowtie2 with parameters --sensitive -a 20 --quiet -p 8 and saved as a bam file using
439 the 'samtools view' command. As gene sequences are relatively short, reads from both ends of the
440 metagenomic sequencing samples were mapped separately, and reunited prior to variant calling.

441 Determining metagenomic read assignment probability

442 We determined the probability of assignment of metagenomic reads to marine microbial genes using the
443 Iterative Coverage-based Read-assignment Algorithm (ICRA) (*32*) with parameters max_mismatch=12,
444 consider_lengths=True, epsilon=1e-6, max_iterations=30, min_bins=4, max_bins=100, min_reads=10,
445 dense_region_coverage=60, length_minimum=300, length_maximum=2e5, use_theta=False. To prevent
446 spurious mapping, alignments were considered for downstream analysis only if the probability of
447 alignment was higher than 0.9.

448 Variant calling

449 Alignments from both ends of all sequencing runs pertaining to the same sample were united using the
450 samtools cat command and sorted using the samtools sort command, with default parameters.
451 To facilitate variant calling in tractable timescales, each filtered, untied and sorted bam file was split into
452 chunks, each encompassing 10,000 reference sequences (out of about 40 million reference sequences).
453 For each such batch of reference sequences, we called variants across all samples using the following
454 command: bcftools mpileup --threads 4 -a FORMAT/AD -Q 15 -L 1000 -d 100000 -m 2 -f <OM-RGC

455 fasta> <bam filenames> | bcftools call --threads 4 -Ov -mv -o <output vcf>, where <OM-RGC fasta> is
456 the fasta file of OM-RGC nucleotide sequences, <bam filenames> are the filenames of all bam files
457 pertaining to the reference sequence chunk in question, and <output vcf> is the output vcf file pertaining
458 to that same chunk.
459 To ensure that sequencing errors do not skew our results, we considered single nucleotide variants as
460 SNPs for downstream analysis only if they had a GATK quality score of at least 30, and an allele
461 frequency of at least 1% (*33*). While the GATK quality requirements filter out sequencing errors
462 randomly distributed across the genes, the 1% criterion eliminates random sequencing errors that
463 accumulate in the same position when depth of coverage reaches very high numbers.
464 For a sample mapped to a reference gene to be considered for downstream analysis, we demanded that at
465 least 60% of SNPs called along the length of the reference gene for that sample would be supported by at
466 least 4 reads, thereby enabling accurate calculation of pN/pS ratios. For a gene to be considered for
467 downstream analysis, we demanded for that gene to have at least one SNP common to 20 or more
468 samples.

469 <u>Calculation of pN/pS in single genes</u>
470 While comparing SNP patterns across samples, it is instrumental to avoid biases due to differences in
471 coverage. We therefore downsampled the read coverage (depth) to the minimum across all samples, for
472 each position in a sample that was supported by more than 4 reads. For positions that had minimum
473 support of fewer than 4 reads, no subsampling was performed. Subsampling was performed by drawing
474 from a multinomial distribution, with n trials and variant probabilities p, where n was set to the calculated
475 minimum depth and p set to the relative abundance of each variant in a given sample.
476 The ratio of non-synonymous and synonymous polymorphisms was calculated by considering all called
477 SNPs in every gene. First, we calculate a consensus sequence for each gene by taking, for each SNP
478 position, the variant that was overall more common across all samples (after the subsampling performed
479 above). We counted, for each gene, the number of non-synonymous and synonymous sites across the
480 consensus sequence. For each SNP position in each sample, we counted the number of synonymous and
481 nonsynonymous substitutions. As more than one variant can exist in a single sample, we considered the
482 relative abundance of synonymous to nonsynonymous substitutions dictated by the different variants. For
483 example, if the reference codon was CAC, coding for histidine, one variant, a C-to-G transversion in the
484 third position abundant at 50%, led to a nonsynonymous mutation that resulted in glutamine (CAG) while
485 another variant in the same sample and the same position was a synonymous C-to-T transition, we
486 counted 0.5 synonymous substitutions and 0.5 nonsynonymous substitutions. We followed by calculating
487 the pN/pS ratio:

488 $$pN/pS = \frac{nonsynonymous\ substitutions}{nonsynonymous\ sites} / \frac{synonymous\ substitutions}{synonymous\ sites}$$

489 Aggregation of calculated metrics using KEGG and eggNOG orthologies

490 Functional assignments to KEGG KOs and eggNOG OGs for all OM-RGC genes were computed using
491 eggNOG-mapper v2 based on the eggNOG v5.0 database (*19*, *34*). For each functional assignment in each
492 sample, all OM-RGC genes annotated with the same functional assignment were concatenated and treated
493 as one long genomic sequence per the calculation of pN/pS ratios. To reduce noise in pN/pS calculation,
494 we considered only KOs and OGs that had at least 5 genes per sample, in at least 50 samples.
495

496 Calculating nonsynonymous mutations leading to radical and conservative amino acid replacements

497 We calculated the ratio of nonsynonymous mutations leading to radical amino acid replacements as
498 compared to those leading to conservative replacements using two known amino acid properties: (1) the
499 polar requirement scale (PR) and (2) amino-acid hydropathy index, and multiple thresholds of "radical"
500 and "conservative" (**Table S2**). For each KEGG ortholog and each radical/conservative threshold, we
501 compared the proportion of radical nonsynonymous substitutions (out of all possible radical sites) to the
502 proportion of conservative ones (out of all conservative sites) across all samples using a Mann-Whitney *U*
503 test. P-values were adjusted to control for multiple hypotheses.
504 Formally, our statistical analysis postulates:

505 H_0: proportion of radical changes $\geq$ proportion of conservative changes
506 H_1: proportion of radical changes $<$ proportion of conservative changes

507 For each radical/conservative threshold, we calculated the proportion of KEGG orthologs in which we
508 reject this null hypothesis and conducted a permutation test. We permuted the sample labels of each
509 radical/conservative threshold across 10,000 iterations, where in each iteration we calculated the
510 proportion of KEGG orthologs in which we reject the null hypothesis. The P-value, per
511 radical/conservative threshold, is defined to be the proportion of iterations in which the ratio of KEGG
512 orthologs in which we reject the null hypothesis is greater than in the original data.
513 To examine whether highly expressed genes have fewer radical amino acid substitutions as compared to
514 lowly expressed genes, we used data from Pachiadaki et al. (*22*) to rank KEGG orthologs by their mean
515 expression.

516 Benchmarking data of assembled genomes

517 Benchmarking data was employed in the form of uncultivated single cells of dominant and representative
518 lineages of the surface ocean (SAR-11, SAR-86 and Prochlorococcus) from different parts of the world.
519 We used genomes varying at the strain level (*20*), with potentially minimal batch effects, as those have
520 been analyzed in a single laboratory.

521 Calculation of dN/dS in single genes

522 For every gene ortholog in each one of the lineages dN/dS was calculated as follows:

523 1. We extracted all gene sequences pertaining to each ortholog from all strains within each
524 clade. We determined orthologous genes using the annotation that was published
525 alongside the single-cell data. For each clade, we chose the annotation type that had the
526 highest coverage of genes: prokka (*35*) for Prochlorococcus, and Swissprot (*36*) for SAR-
527 11 and SAR-86.
528 2. To account for the possibility that gene sequences may be misannotated resulting in the
529 grouping of more than one orthologous group under the same name, we performed
530 additional curation by clustering genes using Ward hierarchical clustering (*37*) on protein
531 sequences using normalized edit distance (Levenshtein) as a dissimilarity metric. We cut
532 off clusters that had dissimilarity lower than 30%, to ensure structural similarity.
533 3. For each of the resulting clusters that had more than 20 members (i.e., genes
534 corresponding to strains within the clade) we generated a multiple sequence alignment
535 (MSA) of protein sequences using MAFFT (*38*) , with the following command:
536 mafft --quiet --maxiterate 1000 --localpair <input_fasta> > <output_msa>
537 Following that, we computed a codon alignment using the generated MSA and nucleotide
538 sequences, with the biopython AlignIO package.
539 4. Using the MSA, we computed a phylogenetic tree of the sequences using raxml (*39*)
540 (raxmlHPC command line version) with model GTRGAMMA; rooted it using the -f I
541 modifier; and computed the ancestral sequence using the -f A modifier.
542 5. For each leaf node (strain ortholog) in the phylogenetic tree of gene sequences, we
543 calculated dN/dS with respect to the tree root, using the biopython cal_dn_ds function
544 with the "ML" model (Goldman and Yang 1994)(*40*) and default parameters.

6. To quantify purifying selection on different gene functions, we annotated genes from the OM-RGC database using KEGG orthology [KO; (*18*)] and aggregated the dN/dS across genes from the same KO using a geometric mean.

Environmental variables

For each sample, we compiled measurements pertaining to the following environmental parameters: Depth [m], Nitrate [µmol/kg], Nitrate+Nitrite [µmol/kg], Oxygen [µmol/kg], Phosphate [µmol/kg], Silicate [µmol/kg], Temperature [°C] and Salinity [g/kg].

Tara oceans metadata was downloaded from PANGAEA (https://doi.pangaea.de/10.1594/) with accession numbers PANGAEA.875575 (Nutrients) and PANGAEA.875576 (Watercolumn sensor), and recorded median values for all the above nutrients were extracted. Tara nutrient concentrations were given as [µmol/l]. Conversion to [µmol/kg] was done by dividing the measured concentration by the measured specific gravity for the same sample.

bioGEOTRACES metadata was compiled from CTD sensor data and discrete sample data from the GEOTRACES intermediate data product v.2 (*41*). HOT metadata was downloaded from the ftp server of the University of Hawai'i at Manoa (ftp://ftp.soest.hawaii.edu/hot/) and BATS metadata was downloaded from the Bermuda Institute of Ocean Sciences (http://bats.bios.edu/bats-data/). As GEOTRACES/HOT/BATS ocean water samples are not linked to specific biological samples as is the case with Tara oceans samples, we considered only water samples from the exact same geographic location, within a day from biological sample collection time, and within 5% difference in depth of collection, and chose the closest sample in terms of time and depth of collection. As all these measurements of environmental conditions are highly correlated with each other (**Fig. S2I**), we utilized this correlation structure to impute missing values using the EM algorithm (*42*).

Linear mixed models

We used a linear mixed model (LMM) with variance components, commonly used in population genetics (*43*), which allows to estimate the fraction of variance in pN/pS ratios (dependent variable) that is explained by the environment (random effect) while controlling for the complex correlation structure between the environmental parameters, in a data-driven manner.

Generative model. Consider a collection of $M_i, where\ i \in \{1,2\}$ features (i.e., $M_{1,2}$number of KEGG and eggNOG orthologs respectively), each measured across N samples. We get as input an $(M_i \times N)$ matrix O^i, where ${O_{mn}}^i$ is the pN/pS of ortholog m in sample n. Let ${y^i}_m = ({O_{m1}}^i, \ldots, {O_{mN}}^i)$ be a $N \times 1$ vector

representing the pN/pS in ortholog m, according to grouping i, across N samples (e.g., pN/pS in KO K02274 across N samples). Let W be a $(N \times q)$ normalized matrix of environmental measurements. This included the depth of the sample, water temperature and salinity, as well as concentration of the key molecules nitrate, nitrite, oxygen, phosphate and silicate.

With these notations, we assume the following generative linear model:

$$y^{i}{}_{m} = Wu_m + \epsilon_m, \qquad (1)$$

where u_m and ϵ_m are independent random variables distributed as $u_m \sim N\left(0, \sigma^2{}_{u_m} I\right)$ and $\epsilon_m \sim N(0, \sigma^2_{\epsilon_m} I)$. The parameters of the model are $\sigma^2{}_{u_m}$ and $\sigma^2_{\epsilon_m}$.

An equivalent mathematical representation of model (1) is given by

$$y^{i}{}_{m} \sim N(0, \sigma_K^2 K + \sigma^2_{\epsilon_m} I), \qquad (2)$$

where $\sigma_K^2 = M_i \sigma^2{}_{u_m}, K = \frac{1}{M_i} WW^T$. We will refer to K as the environmental kinship matrix, which represents the similarity, in terms of environmental covariates, between every pair of samples across grouping i (i.e., represent the correlation structure to the data).

Environmental explained variance (EEV). The explained variance of a specific feature $y^{i}{}_{m}$ by the environmental measurements

$$\chi^{i}{}_{m} = \frac{\sigma_K^2}{\sigma_K^2 + \sigma^2_{\epsilon_m}}. \qquad (3)$$

Estimating effective population size

To approximate effective population size (N_e) we use fourfold degenerate synonymous diversity. Calculation was adapted from Schloissnig et al.(*44*):

$$\pi_{ffsyn}(S, G) = \frac{1}{|G_{ffsyn}|} \sum_{i=1}^{|G_{ffsyn}|} \sum_{B_1 \in \{A,C,G,T\}} \sum_{B_2 \in \{A,C,G,T\} \setminus B_1} \frac{x_{i,B_1}}{c_i} \times \frac{x_{i,B_2}}{c_i - 1}, \qquad (4)$$

598 where S is the sample, G is the gene of interest, $|G_{ffsyn}|$ is the number of fourfold degenerate sites, x_{i,B_j} is
599 the number of times nucleotide B_j was seen in the fourfold degenerate position i, and c_i is the coverage of
600 position i.

601 Controlling for the effects of time and effective population size
602 To control for the effects of time and effective population size, we added these two potential confounders
603 to our LMM as fixed effects. We formulate our LMM as follows:
604

605 $$y^i{}_m = \beta_{o_i} + \beta_{1_i} pS_i + \beta_{2_i} SD_i + Wu_m + \epsilon_m \ , \qquad (5)$$

606 where β_{0_i} is the intercept, pS (as a proxy for time) is the first fixed effect with β_{1_i} as its coefficient, SD_i
607 (as a proxy for effective population size) is the second fixed effect with β_{2_i} as its coefficient. $W$is a
608 normalized matrix of environmental factors, u_m and ϵ_m are independent random variables distributed as
609 $u_m \sim N(0\,, \sigma^2{}_{u^n{}_m} I\,)$and $\epsilon_m \sim N(0\,, \sigma^2_{\epsilon_m} I\,)$. The parameters of the model are β_{o_i} , β_{1_i} , β_{2_i} $\sigma^2{}_{u_m}$and
610 $\sigma^2_{\epsilon_m}$.

611 LMM on samples from the epipelagic zone
612 To demonstrate that the association of pN/pS with the environment is not confounded by depth, we reran
613 the LMM (eq. 5) only on samples from the epipelagic zone (up to 76m, median depth across 746
614 samples).

615 Linear model for dN/dS rates
616 We modeled the benchmarking data by constructing a linear model for dN/dS (the dependent variable)
617 with the environmental factors as independent variables:

618 $$dN/dS_m = \beta_{0_m} + Wu_m + \epsilon_m, \qquad (6)$$

619 where β_{0_m} is the intercept, W is a normalized matrix of environmental factors and u_m is a vector of
620 environmental coefficients. $\epsilon_m \sim N(0, \sigma^2{}_{\epsilon_m} I)$. $m \in \{1, \ldots, M\}$, where M is the number of genes tested.
621 To account for the effect of time as potential confounder we constructed the following linear model for
622 dN/dS (the dependent variable) with the environmental factors and dS (as a proxy for time) as
623 independent variables.

$$dN/dS_m = \beta_{0_m} + \beta_{1_m} dS_m + Wu_m + \epsilon_{2_m}, (7)$$

where β_{0_m} is the intercept, dS (as a proxy for time) is a fixed effect with β_{1_m} as its coefficient, W is a normalized matrix of environmental factors and u_m is a vector of environmental coefficients that were available with these samples (phosphate, depth, nitrate + nitrite). $\epsilon_{2_m} \sim N(0, \sigma^2{}_{\epsilon_{2_m}} I)$. $m \in \{1, \ldots, M\}$, where M is the number of orthologs tested. In this model we only used depth, NO_2+NO_3 and phosphate, as these were the environmental factors available for the benchmarking data.

To account for potential underlying structure in terms of phylogeny, we applied phylogenetic generalized least squares (*45*) (PGLS) regression in addition to ordinary least squares (OLS) regression to model dN/dS. Phylogenetic regression models are used when there is an underlying phylogenetic clustering structure of the observations. In such case, samples may be lowly or highly phylogenetically diverged leading to residuals that are not *i.i.d.*, which can affect the fit of the model. On the other hand, the caveat of PGLS is an underestimation of the R^2 as compared to linear models (*46*, *47*). In our data PGLS and OLS essentially gave the same results and we therefore only report the latter.

KEGG KO expression data as a ranking metric

Using expression data from 4,092 KEGG KOs collected by Kolody et al. (*22*), we ranked the KO genes in our marine samples in the following way. We first represent the expression data in relative abundance space (normalize each sample by its read counts). Next, for each focal KO i, where $i \in \{1, \ldots, I\}$, we sum across its different instances within each sample. The input is an $m \times n$ matrix, where m is the number of different instances of KO i, and n is the number of samples. The output is a vector of length n: $(x_1, \ldots, x_n)$.Finally, we average the expression levels $(x_1, \ldots, x_n)$ across samples:$\frac{1}{n}\sum_{j=1}^{n} x_j$ and rank the KOs based on the calculated average expression. We limited the scope of our analysis to samples collected only in small fraction filters (0.22 μm).

Determination of extracellular genes

To determine extracellular gene groups, we searched the eggNOG v.5 OG database for the words 'secreted' or both words 'extracellular' and 'protein' in their description. We demanded that the words 'autoinducer', 'expression', 'role' and 'hypothetical' are not in the description to prevent instances where (a) the OG in question describes a hypothetical protein; and (b) where the OG produces a secreted particle but is not secreted by itself, as is the case with autoinducer producing genes. To ensure robust pN/pS calculations, descriptions encompassing 10 OGs or more were assigned a group name, while descriptions encompassing less than 10 OGs were grouped together.

654 <u>Calculation of mutation flux in divergent nitrate concentrations</u>

655 To calculate the differential rate of nutrient incorporation between environments, we compare, for each
656 pair of codons *a* and *b*, the ratio of mutations that result in *a* becoming *b* and the ratio of reverse
657 mutations, that result in *b* becoming *a*. We term this quantity mutation flux. To prevent simplex-related
658 confounding effects, mutation flux is estimated using the log-odds ratio between the two (e.g.,
659 $log\ (p(AAA \rightarrow AAC)/p(AAC \rightarrow AAA))$. We created a matrix $H^{(U)}$ as follows: consider a set U of all
660 genes for which SNP measurements exist and a subset $T \subseteq U$ of this set across Ksamples. Let ${G^{(T)}}_j =$
661 (V, E) be a codon graph for subset T and sample j, where $v \in V$ is a codon (e.g., CUU coding for
662 Alanine) and $(v, v') \in E$ if and only if v and v' are one mutation apart (e.g., CUU for Alanine and CAU
663 for Histidine). Let ${w^{(T)}}_j : (v, v') \rightarrow [0,1]$ be a weight function where

664 $$ {w^{(T)}}_j(v, v') = \frac{Number\ of\ mutations\ turning\ v\ to\ v\prime\ in\ subset\ T\ and\ sample\ j}{Number\ of\ occurrences\ of\ codon\ v\ in\ subset\ T\ and\ sample\ j}. $$

665 Let $H^{(T)}$ be a matrix of dimension $(|E| \times K)$, where ${H^{(T)}}_{(v,v\prime),j} = {w^{(T)}}_j(v, v')$. We next sum-normalized
666 H per each sample and compared the codon mutation frequencies between the 40 lowest- and 40 highest-
667 nitrate samples. Despite significant differences in codon mutation frequencies between low-nitrate and
668 high-nitrate samples, some of the difference could be driven by the simplex properties of the sum-
669 normalized codon mutation frequencies, and some could be attributed to the different rates of
670 synonymous mutations between the high- and low-nitrate groups, which, combined with simplex
671 properties, may affect observed nonsynonymous mutation rates. To address these simplex properties, we
672 employed a centered log-ratio (CLR) normalization on *H*. The CLR transformation is a mapping, per
673 codon composition, from the simplex to a Euclidean vector subspace. This log transforms each value and
674 then centers them around zero as given below:

675 $$ CLR(V) = [log\frac{v_1}{g(v)}, \ldots, log\frac{v_D}{g(v)}] = log(v) - log(g(v)) $$

676 where g(v) is the geometric mean of all of the codons.

677 To address differences in rates of the different types of mutations, for each mutation (v, v') in the CLR
678 normalized matrix $H'^{(U)}$ we calculated the log odds ratio between the mutation and its reverse mutation.
679 Namely, we computed mutation flux matrix $F^{(U)}$ where ${F^{(U)}}_{(v,v\prime),j} = {H^{(U)}}_{(v,v\prime),j} - {H^{(U)}}_{(v\prime,v),j}$. We

compared differences in codon mutation flux between low- and high nitrate samples using the Mann-Whitney *U* test.

<u>Calculation of expected random mutation cost per genetic code</u>

Let V be a genetic code with a set $V_s \subset V$ of stop codons. Let $P(v)$ be the abundance of codon $v \in V$ in a sample and $P(mut(v, v'))$ the probability of a single mutation from codon v to v'. Let $c_e: V \times V \rightarrow [0,1]$ be a cost function for element e, where:

$$c_e(v, v') = \#\ of\ atoms\ of\ e\ in\ v' - \#\ of\ atoms\ of\ e\ in\ v.$$

With these notations, we define the expected cost of genetic code V for element e as follows:

$$E[C_e(V)] = \sum_{v,v' \in V} P(v)P(mut(v, v'))c_e(v, v'),$$

and the ERMC cost as:

$$ERMC_e(V) = \sum_{v,v' \in V} P(v)P\big(mut(v, v')\big)\mathrm{c}_{\mathrm{e}}^{+}(v, v'),$$

where $c^{+}{}_{e} = max(0, c_e)$. When testing the random mutation cost on hydrophobicity or PR we consider the absolute difference in the amino-acid hydropathy index or PR value pre- and post-mutation.

We estimate $ERMC_e(V)$ as follows:

1. We define $P(v)$ as the median abundance of all codons v' coding for the same amino acid as v.
2. We calculate mutation rates in sites that are under minimal selection. To this end, we estimate $P(mut(v, v'))$ by calculating, from fourfold-degenerate synonymous mutation sites the average abundance of each single nucleotide mutation (e.g. A to C) across all genes in which there are called SNPs in all ocean samples, excluding stop codons. We then estimate $P(mut(v, v'))$ using the relative abundances of all pairs of single nucleotide mutations. We estimate $P(mut(v, v'))$ for *Prochlorococcus* and *Synechococcus* using published transition:transversion rates (*1*, *48*).
3. We calculate *c* using information on the amino acids which each codon codes for.

To compute a P-value, we generate a null distribution by calculating $ERMC_e(V)$ for hypothetical genetic codes. We randomize the first and second position of all codons, while maintaining that the two sets of first and second positions in which the stop codons reside are separated by a single transition mutation.

705 Confounding effects between cost functions for the structure of the genetic code

706 To confirm that our elemental cost function is not confounded by traditional properties of amino acids
707 such as the polar requirement (PR) and hydropathy index (*11*, *15*), we calculated the expected random
708 mutation cost (ERMC), per hypothetical genetic code, using these common cost functions across 1
709 million codes. To this end, we randomized the first and second position of all codons, while maintaining
710 that the two sets of first and second positions in which the stop codons reside are separated by a single
711 transition mutation. We next calculated a contingency table for each pair of cost functions for both
712 nitrogen and carbon (i.e., $ERMC_N(V)$:$ERMC_{PR}(V)$, $ERMC_N(V)$:$ERMC_{hyd}(V)$, (V):$ERMC_{PR}(V)$,
713 $ERMC_C(V)$:$ERMC_{hyd}(V)$). We assign each code to one of four bins in the following way: (1) surpassing
714 the standard code in both cost functions (e.g., nitrogen and PR), (2) surpassing the standard code only in
715 element e cost (e.g. only nitrogen), (3) surpassing the standard genetic only in the traditional cost function
716 (e.g., PR), (4) not surpassing the standard code in either. We applied the Chi-square test of independence
717 with two degrees of freedom to each contingency table.

718 Finally, we examined the subset of genetic codes (out of the 1M hypothetical codes) that have a lower
719 ERMC than the standard code in terms of PR ($ERMC_{PR}$) or hydropathy ($ERMC_{hyd}$), and tested whether
720 this subset is also optimized for nitrogen and carbon. To this end, we created a null model in which we
721 compute EMRC only accounting for hydropathy or PR robustness and an expanded, nested model that
722 adds minimization of nitrogen and carbon utilization. Under the null, we expect that in the subset of codes
723 optimizing for hydropathy and PR, nitrogen and carbon utilization will also be minimized. Under the
724 alternative, if such nutrient optimization indeed exists separately from structural optimization, we expect
725 that nitrogen carbon and utilization will not be optimized. Exact P-values are defined to be the fraction of
726 shuffled codes with ERMC smaller or equal to the standard code (one for each element - carbon and
727 nitrogen) in this subset of codes that have a lower ERMC than the standard code in terms of PR
728 ($ERMC_{PR}$) or hydropathy ($ERMC_{hyd}$).

729 Determination of 'square' and 'diagonal' arrangements of nitrogen in the genetic code

730 We define a 'square' arrangement of the codons coding for nitrogen-rich amino acids histidine,
731 glutamine, asparagine, lysine and arginine as one where their codons span only two nucleotides in the first
732 position and two nucleotides in the second position. In the standard code, these amino acids are coded by
733 CAN, CGN, AAN, AGR, following a square configuration. In contrast, a 'diagonal' arrangement of the
734 codons coding for these amino acids is one where they span all possible nucleotides in the first position
735 and all possible nucleotides in the second position. For example, a genetic code where TTY codes for
736 histidine, TTR for glutamine, CCN and AAR for arginine, GGY for asparagine and GGR for lysine
737 constitutes a 'diagonal' arrangement of nitrogen amino acids. In each of the alternative genetic codes we

738 generated, we tested whether either of these conditions hold and, if so, designated the code as 'square' or
739 'diagonal' accordingly.

740 *Prochlorococcus* and *Synechococcus* genomic data and mutation rates

741 We downloaded *Prochlorococcus* and *Synechococcus* protein-coding gene sequences (where available)
742 from the Joint Genome Institute (https://genome.jgi.doe.gov/portal/) following accession numbers
743 published by Berube et al. (*1*). To estimate codon relative abundance $P(v)$, we counted and sum-
744 normalized codons in all protein-coding genes for each strain. To estimate codon mutation rate
745 $P(mut(v, v'))$, we used the published transition:transversion rate of 2:1 for *Prochlorococcus* and
746 *Synechococcus* (*23*).

747 Multiple taxa ERMC calculation

748 To calculate ERMC for 39 taxa across multiple transition:transversion rates, we downloaded codon usage
749 and GC-content data collected by Athey et al. (*49*). We used codon usage counts to estimate $P(v)$ and 11
750 transition:transversion rates (1:5, 1:4, 1:3, 1:2, 2:3, 1:1, 3:2, 2:1, 3:1, 4:1, 5:1) to estimate $P(mut(v, v'))$.

751 **Supplementary Text**

752 Supplementary Note 1. Conservative nonsynonymous substitutions are more common than radical ones in
753 marine microbial metagenomes

754 To corroborate the validity of calculating selection metrics from metagenomic samples, we sought to
755 replicate a known property of mutations in amino acids, namely that mutations resulting in "conservative"
756 amino acid replacements are more common than those resulting in "radical" ones (*50*). To this end, we
757 calculated the ratio of nonsynonymous mutations leading to radical amino acid replacements and
758 compared them to those leading to conservative replacements. To best define "radical" and
759 "conservative" nonsynonymous substitutions, we used two known amino acid properties: (1) the Polar
760 Requirement (PR) scale and (2) amino-acid hydropathy index (**Table S2**). Since the definition of radical
761 and conservative may be subjective, we considered multiple thresholds as radical and conservative for
762 each property. For example, a radical substitution could be one that results in a difference of more than 3
763 units on the hydropathy index whereas a "conservative" one results in just one or less (**Table S2**;
764 hydropathy case 4).
765 We then calculated, for each KEGG ortholog and for each radical/conservative threshold, the proportion
766 of radical nonsynonymous substitutions (out of all possible radical sites) as compared to the proportion of
767 conservative ones across samples using a Mann-Whitney U test. P-values were adjusted to control for
768 multiple hypotheses.
769 Formally, our statistical analysis postulates:

770 $$H_0 : proportion\ of\ radical\ changes\ \geq proportion\ of\ conservative\ changes$$
771 $$H_1 : proportion\ of\ radical\ changes\ < proportion\ of\ conservative\ changes$$

772 Next we calculated, for each radical/conservative threshold, the proportion of KEGG orthologs in which
773 we reject the null hypothesis (i.e., the proportion of KOs in which nonsynonymous mutations leading to
774 conservative amino acid replacements are significantly more common than nonsynonymous mutations
775 leading to radical amino acid replacements) and conducted a permutation test. We permuted the sample
776 labels of each radical/conservative threshold across 10,000 iterations, where in each iteration we
777 calculated the proportion of KEGG orthologs in which we reject the null hypothesis. The P-value, per
778 radical/conservative threshold, is defined to be the proportion of iterations in which the ratio of KEGG
779 orthologs in which we reject the null hypothesis is higher as compared to the original data. We find that

780 these comparisons are significant for all cases, regardless of the amino acid property in question or the
781 implemented threshold and that this is the case for most KOs ($P<0.0001$ for all comparisons; **Table S2**).
782 Finally, we examined whether highly expressed genes have fewer radical amino acid replacements as
783 compared to lowly expressed genes. To this end, we used an additional expression dataset for marine
784 microbial gene expression to rank KEGG orthologs by their mean expression (similar to the approach
785 implemented in **Fig. 1D** (*22*)). We find that KEGG orthologs which exhibit significantly more
786 conservative mutations than radical ones have significantly higher expression as compared to the
787 alternative case (Wilcoxon signed-rank test $P<10^{-5}$). These results demonstrate that metagenomic
788 sequences are adequate for analyzing selective pressures.

789 Supplementary Note 2. Resource-driven selection drives specific amino acid changes

790 Our results indicate that mutations leading to an addition of resources would be more common in nutrient-
791 rich environments. This is supported by previous observations regarding genomic and proteomic changes
792 associated with environmental concentrations of nitrate (*2*, *5*, *21*, *51–53*). We thus sought to quantify how
793 resource conservation manifests in molecular changes to DNA and protein sequences. Mutations in
794 nitrogen-rich and, typically, carbon-poor environments were shown to drive an increase in genomic GC
795 content (which increases nitrogen requirements), as well as by higher nitrogen and lower carbon in
796 protein sequences.
797 Using SNP derived from metagenomic data, we corroborate previous findings (*5*, *21*), and show that
798 mutations leading to higher GC content and those leading to higher nitrogen proteins are more prevalent
799 in high-nitrate waters as compared to low-nitrate (**Fig. S12A,B**). We note that mutations leading to higher
800 GC content (higher nitrogen DNA) often result in higher nitrogen proteins (*2*, *3*). We further deconvolve
801 these two types of mutations and find that the ratio of mutations that lead to higher nitrogen DNA but do
802 not change protein nitrogen-content increases with environmental nitrate ($R=0.333$; **Fig. S12C**).
803 Additionally, the ratio of mutations that increase protein nitrogen but do not change DNA nitrogen-
804 content is even more strongly correlated with nitrate concentrations ($R=0.473$; **Fig. S12D**). We
805 hypothesize that this stronger association of the environment with protein sequences can be attributed to
806 the higher copy number of protein as compared to DNA and RNA, often in the order of 10^2-10^5 protein
807 copies per RNA molecule (*54*, *55*) leading to higher cost per gene, and thus to stronger selective pressure.
808 Our results suggest a differential effect of mutations on nutrient incorporation in DNA as compared to
809 protein sequences.
810 To study amino acid substitutions typical to low- and high-nitrate environments, we examined mutations
811 leading to changes in protein nutrient content by comparing codon mutation frequencies in low- and high-

812 nitrate samples, after accounting for simplex-related confounders (Methods). We found significant
813 differences in codon mutation frequencies between environments (**Fig. S13A-C**). We next sought to
814 examine the typical change in nutrient consumption in varying nitrate concentrations. We therefore
815 defined mutation flux as the ratio between a codon mutation and its reverse, thereby measuring the typical
816 flux of nutrients; and estimated it using the log-odds ratio between the two (e.g., $log\ (p(AAA \rightarrow$
817 $AAC)/p(AAC \rightarrow AAA))$. Notably, across all the mutations significantly more prevalent in samples from
818 high nitrate environments (Methods), averaged across amino acids, we find a total increase in nitrogen
819 (**Fig. S13D**; 18 N atoms summed across all significant amino acid changes, P=0.0082), decrease in carbon
820 (**Fig S13D**; -37 atoms, P=0.0165), decrease in sulfur (-6 atoms, P=0.009), a decrease in molecular weight
821 (-508.91 g/mol, P=0.0193) and a non-significant decrease in oxygen (-6 atoms, P=0.1505). These results
822 provide additional evidence that the higher the nitrate concentrations are, the weaker the selection against
823 mutations leading to higher nitrogen incorporation into protein sequences. While nitrate concentrations
824 increase with depth (**Fig. S2I**), labile dissolved organic carbon concentrations typically decrease (*56*). We
825 therefore also show the specific codon mutations reducing carbon incorporation into proteins in nitrogen
826 rich and typically carbon poor, environments (**Fig. S13D,E**).

827 <u>Supplementary Note 3. The association between the environment and pN/pS is robust to time, effective</u>
828 <u>population size and environmental niche</u>

829 The magnitude of purifying selection, approximated by pN/pS ratios, may increase as a function of both
830 time and effective population size (*57*, *58*). These two mechanisms might also be associated with
831 environmental conditions and we thus sought to control for them as potential confounders. Time may
832 affect the magnitude of pN/pS ratios via random genetic drift. Effective population size may affect
833 selection via competition, as in large populations, any slightly deleterious mutation is rapidly selected
834 against. We used the ratio of synonymous substitutions (i.e., pS) calculated from our data as an
835 approximation of dS and representing neutral polymorphism rate, as a proxy for time (*59*). To
836 approximate effective population size, we relied on population genetic theory suggesting that effective
837 population size is scaled to diversity at the silent sites when assuming constant mutation rate between
838 populations. As our data is composed of closely related sequences and hence similar mutation rate, we
839 used fourfold-degenerate synonymous diversity (SD) as a proxy for the effective population size (*44*)
840 (Methods).
841 We first tested whether SD is associated with environmental conditions, and compared the SD of
842 populations from different oceanic regions with different environmental parameters. We found that pN/pS
843 was more strongly associated with most environmental parameters than SD (Kolmogorov-Smirnov test

844 $P<10^{-20}$ for depth, nitrate, oxygen, phosphate, silicate and temperature; $P<10^{-5}$ for NO_2+NO_3
845 concentrations; non-significant for salinity; **Fig. S3**) suggesting that selection on protein sequences may
846 be more strongly affected by the environment than effective population sizes are. We further found that in
847 KOs with the highest EEV these differences are the most pronounced (Wilcoxon signed-rank test $P<10^{-15}$;
848 **Fig. S14**).
849 As LMMs also allow controlling for potential confounders as fixed effects, we next added our
850 approximations of time and the effective population size to our LMM in order to account for their effect
851 on the magnitude of pN/pS ratios. Accounting for these potential confounders as fixed effects would
852 enable a more accurate quantification of the environmental effect. Similar to the results of our original
853 LMM, across both KEGG and eggNOG orthologs, we found that a significant fraction of the variance in
854 pN/pS can be attributed to the environment, even after controlling for these confounders (**Fig. S4C,D**;
855 Mann-Whitney *U* test $P<10^{-16}$). We found a high correlation between EEV estimated by the two models
856 (R^2=0.947 and 0.929 for KEGG and eggNOG, respectively; **Fig. S15**).
857 To demonstrate that our results are not confounded by depth, the environmental variable with the second
858 strongest correlation to pN/pS (after nitrate), we recalculated our LMM only on samples from the
859 epipelagic zone (Methods). Similar to the original results of the LMM, we show significant variance
860 explained in pN/pS in this environmental niche (**Fig. S6**; Mann-Whitney *U* test $P<10^{-20}$). Taken together,
861 these results demonstrate that the association between the environment and pN/pS is robust to time,
862 effective population size and environmental niche.

863 <u>Supplementary Note 4. The association between environmental measurements and the magnitude of</u>
864 <u>purifying selection is replicated using single cell data and dN/dS</u>

865 pN/pS, the ratio of nonsynonymous polymorphisms (pN) to synonymous polymorphisms (pS),
866 approximates dN/dS which quantifies the magnitude and type of selection (purifying, neutral or positive)
867 exerted on protein-coding sequences across a phylogenetic group (*44*, *60*). While dN/dS calculation
868 requires genomic sequences at the strain-level that are available for only a small number of samples and
869 from limited strains, it allows for a more accurate quantification of purifying selection in the taxa
870 represented in the phylogeny. We therefore sought to confirm our findings using dN/dS. To this end, we
871 used benchmarking data in the form of assembled genomes from uncultivated single cells from dominant
872 and representative lineages of the surface ocean (SAR-11, SAR-86 and Prochlorococcus) (*20*), and their
873 corresponding environmental measurements (Methods).

874 We calculated dN/dS for KEGG orthologs in each lineage. Given the small number of samples, we
875 estimated the fraction of variance explained by the environment using a linear model with environmental
876 covariates (Methods). Across all bacterioplankton lineages comprising the benchmarking data, we found
877 that a significant portion of the variance in dN/dS ratios calculated from single-cell data is explained by
878 the environment (SAR-11, SAR-86, Prochlorococcus; Wilcoxon signed-rank test $P<10^{-16}$; **Fig. S4E-G**).
879 Similar to our original LMM, we note that this association holds, across all three bacterioplankton
880 lineages, even after controlling for both time as a confounder (Wilcoxon signed-rank test $P<10^{-16}$; **Fig.**
881 **S16**) and for phylogeny, using phylogenetic generalized least-square regression (Methods). These results
882 validate our findings using pN/pS calculated from metagenomic sequences and demonstrate the
883 association between selective pressure and environmental conditions, even in a single clade. This
884 indicates that associations of pN/pS or dN/dS with environmental variables are likely not confounded by
885 organismal properties which may differ in their metabolism across environments, but rather exhibit a
886 consistent trend even within a single clade.

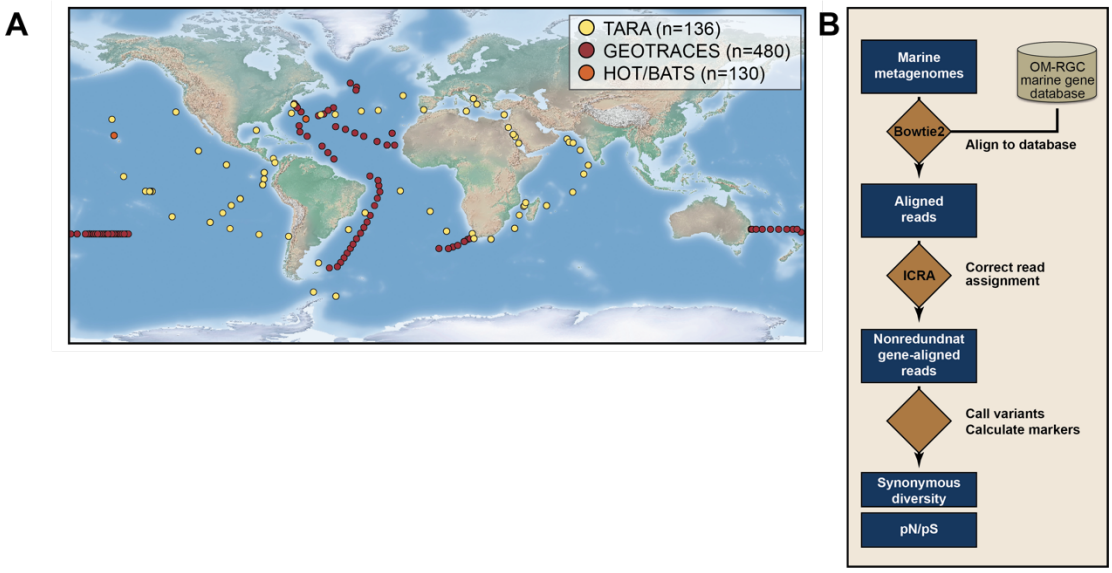


Fig. S1.

(A) Geographical overview of the samples used in this study. (B) Illustration of our computational pipeline.

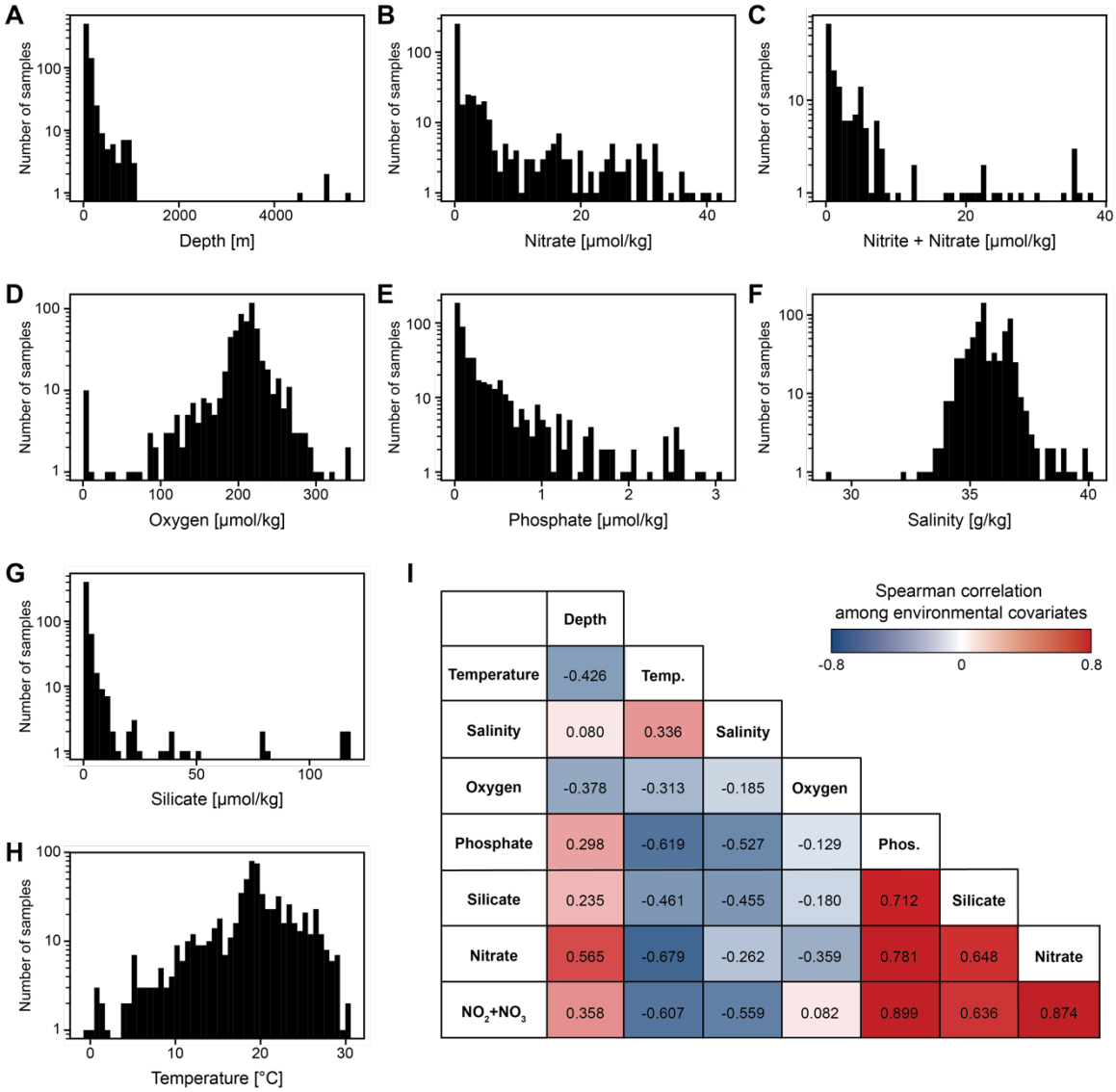


Fig. S2

(A-H) Distribution of measurements taken alongside marine microbial samples for depth (A), nitrate (B), nitrate and nitrite (C), oxygen (D), phosphate (E), salinity (F), silicate (G) and temperature (H). (I) Spearman correlation coefficients between all pairs of environmental measurements across all available samples.

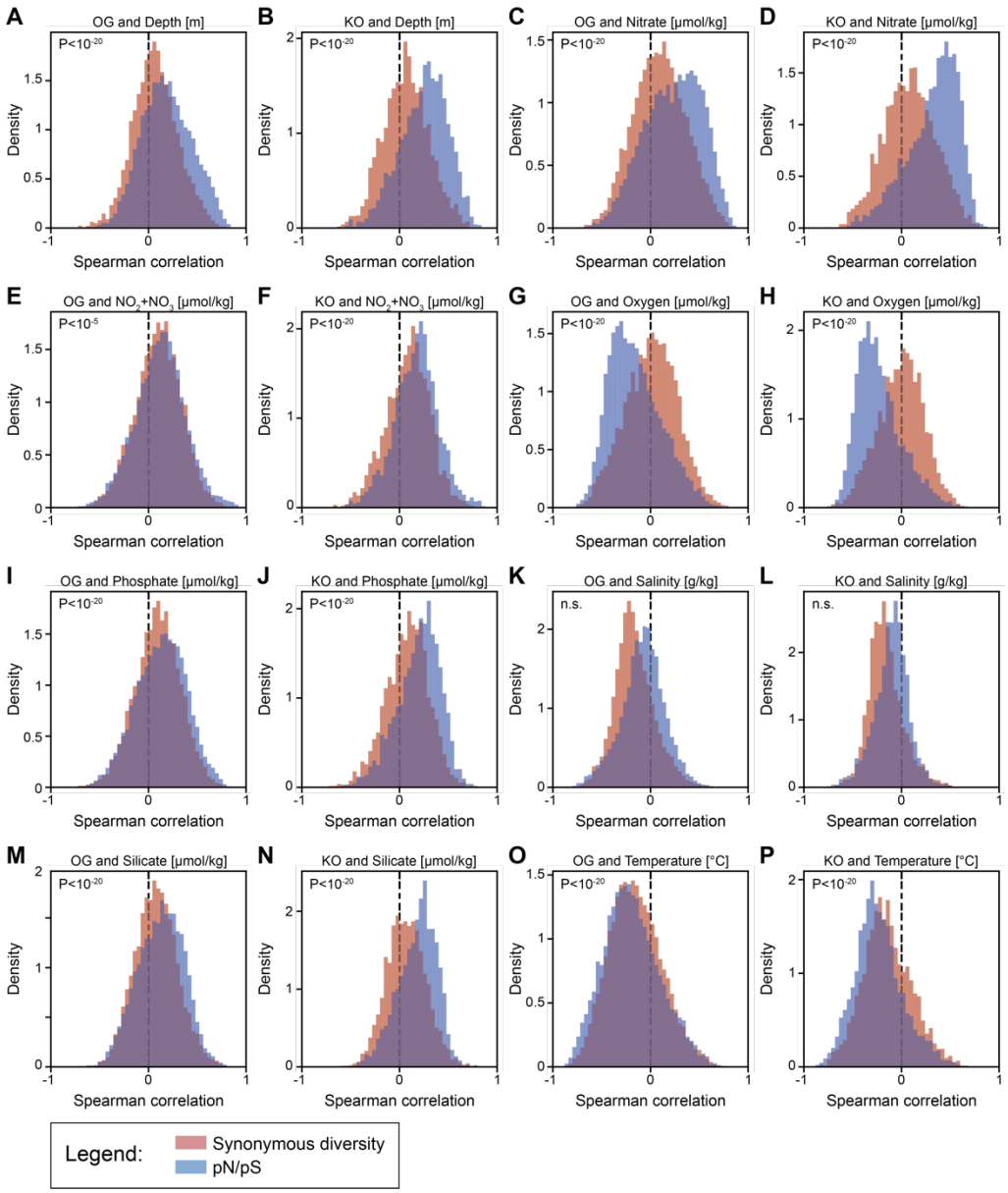


Fig. S3

(A-P) Histograms of Spearman correlations between synonymous diversity (red) or pN/pS (blue) and environmental variables, for both KEGG KOs and eggNOG OGs. Panels A, C, E, G, I, K, M and O depict correlations between OG calculated parameters and depth, nitrate, nitrite and nitrite, oxygen, phosphate, salinity, silicate and temperature, respectively. Panels B, D, F, H, J, L, N and P depict correlations between KO calculated parameters and depth, nitrate, nitrite and nitrite, oxygen, phosphate, salinity, silicate and temperature, respectively. P, one-sided Kolmogorov-Smirnov test.

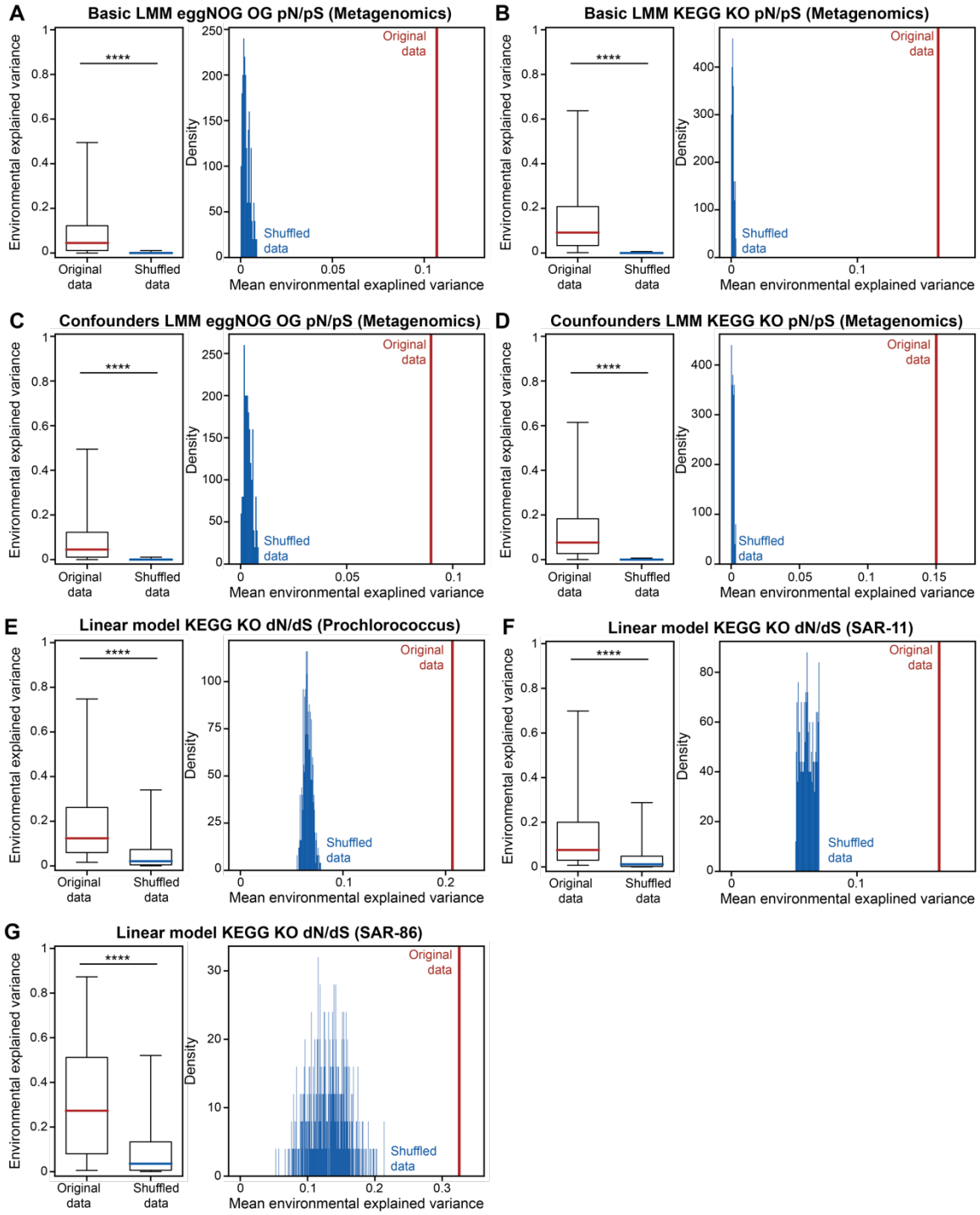


Fig. S4

(A) Left, box plot (line, median; box, IQR; whiskers, 5th and 95th percentiles) of the variance of eggNOG OG pN/pS rates that was explained by the environment in a basic linear mixed model (LMM) setting (Methods) as compared to the same data with shuffled labels; Right, mean variance explained in

909 unshuffled data (red) as compared to a histogram (blue) of mean variance explained in 500 executions
910 with shuffled data. (B) Same as A, for KEGG KO pN/pS. (C) Same as A, for variance in eggNOG OG
911 pN/pS in a LMM corrected for confounding factors of time and effective population size. (D) Same as C,
912 for KEGG KO pN/pS. (E) Box plot (line, median; box, IQR; whiskers, 5th and 95th percentiles) of the
913 variance of KEGG KO dN/dS in genus *Prochlorococcus* that was explained by the environment in a
914 linear model. (F) Same as E, for the SAR-11 clade. (G) Same as E, for the SAR-86 clade. ****, Wilcoxon
915 signed-rank test $P<10^{-16}$.

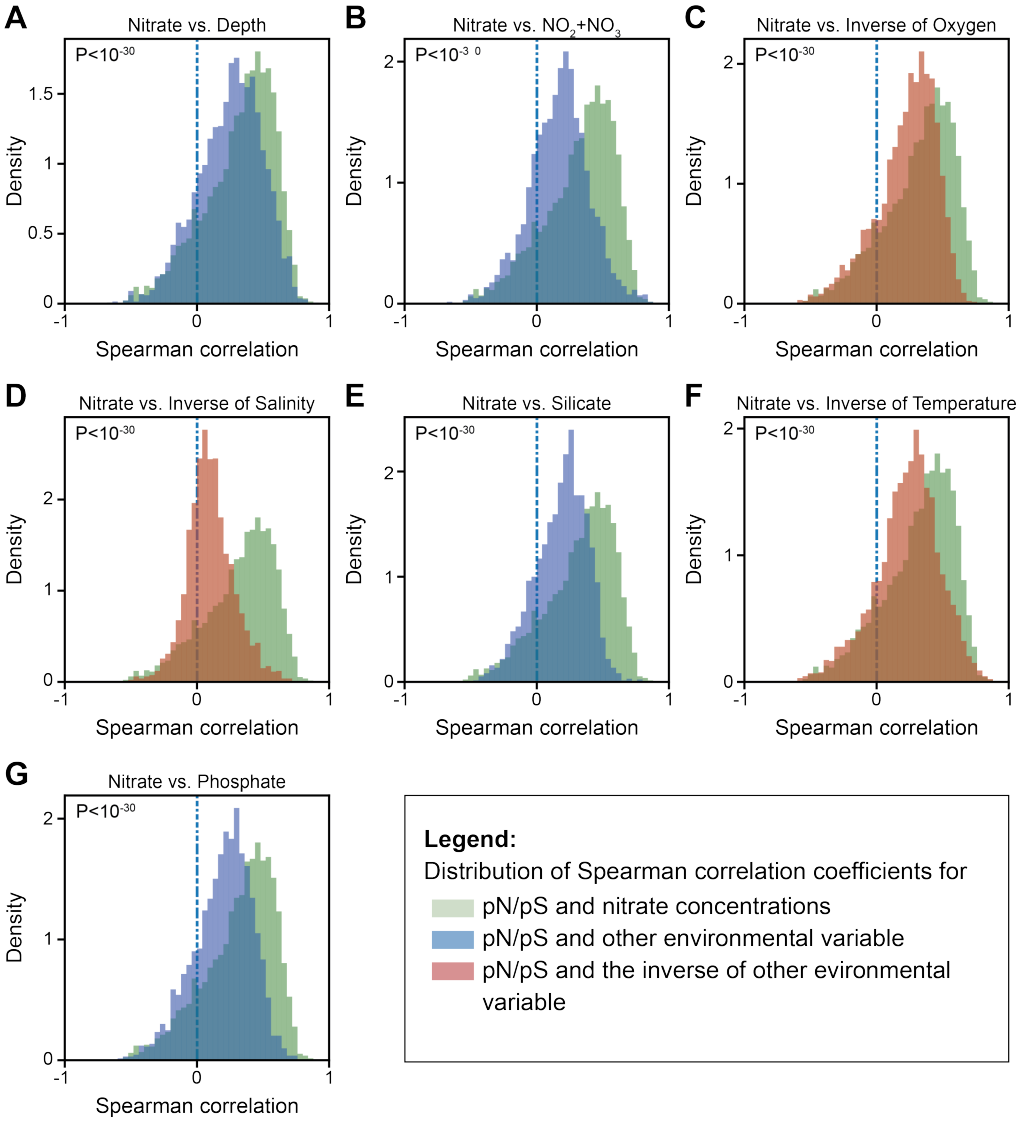


Fig. S5

Histograms of Spearman correlations between environmental variables (blue) or their inverses (red) and pN/pS ratios of KEGG KOs as compared to a Spearman correlation between the concentration of nitrate and pN/pS ratios (green). P, one-sided Kolmogorov-Smirnov test. We note that unmeasured variables correlated with nitrate, such as labile dissolved organic carbon, may also be correlated with pN/pS.

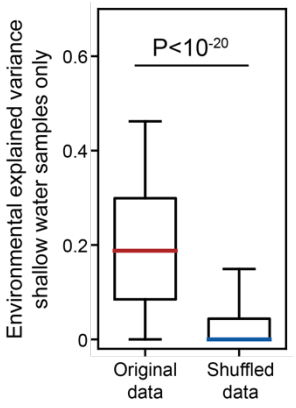


Fig. S6

Box plot (line, median; box, IQR; whiskers, 5th and 95th percentiles) of the variance of KEGG KO pN/pS ratios that was explained by the environment in epipelagic samples (up to 76m) as compared to the same data with shuffled labels.

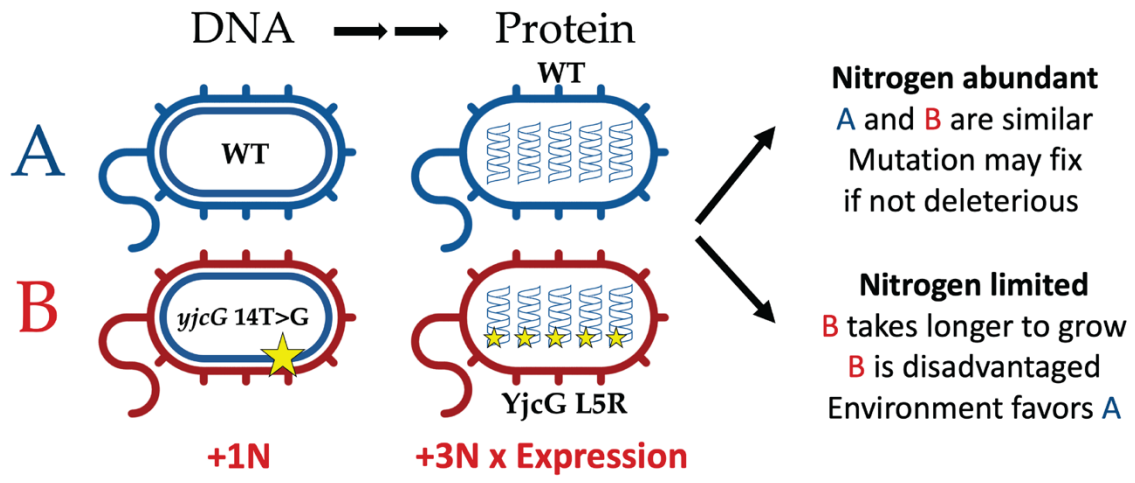


Fig. S7

Illustration of the potential environmental effect of nitrate on protein-coding sequences. Depicted are two microbes, wild-type (WT; A) and a mutant with a T-to-G transversion in the 14th position of the yjcG gene, leading to a leucine to arginine substitution, thereby consuming three extra nitrogen atoms per protein molecule (B). If environmental nitrate is scarce and growth-limiting, A is likely to have a selective advantage over B, which over multiple generations would translate into an extinction of the B variant. In highly expressed genes such a mutation is likely to have an even stronger effect, leading to a quicker extinction. On the molecular level, this phenomenon would translate to lower pN/pS in nitrate-limited waters as compared to nitrate-rich ones.

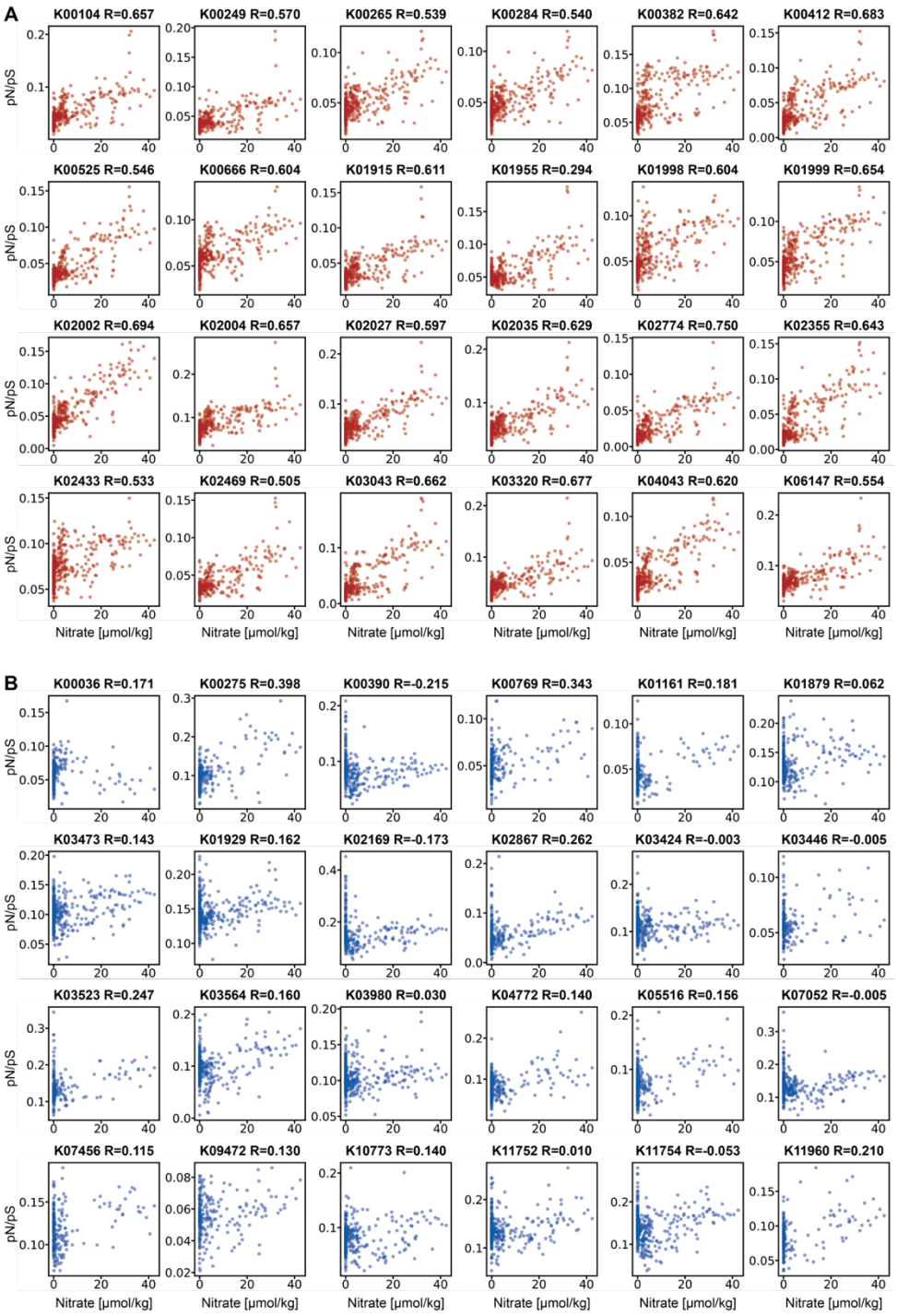


Fig. S8

(top, red) 24 highly expressed KOs, exhibiting high correlation between pN/pS and nitrate concentrations. (bottom, blue) 24 lowly expressed KOs, exhibiting low correlation between pN/pS and nitrate concentrations. R, Spearman rho.

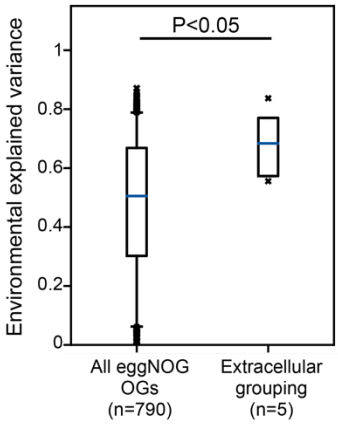


Fig. S9

Box plot (line, median; box, IQR; whiskers, 5th and 95th percentiles) of variance in pN/pS explained by the environment in extracellular gene groups versus all eggNOG OGs.

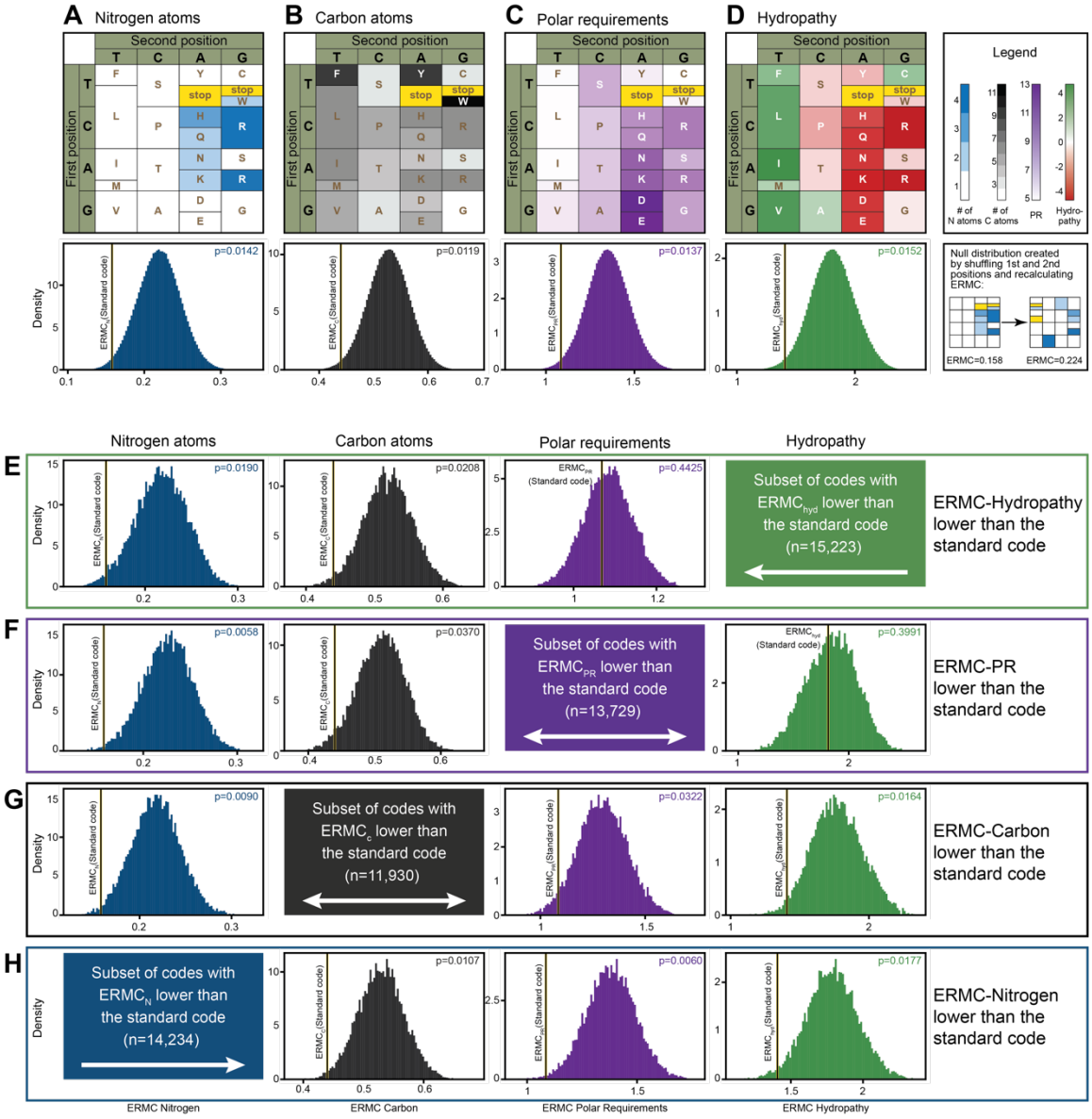


Fig. S10

Upper panel: (Top) Nitrogen content (A; blue color scale), carbon content (B; black), the polar requirement (C; purple) and hydropathy (D; green-red) of different amino acids depicted along their positions in the standard genetic code. (Bottom) Histograms of the expected random mutation cost (ERMC), in 1,000,000 random permutations of the genetic code for nitrogen (blue), carbon (black), hydropathy (green) and polar requirements (purple). Lower panel: Histograms of ERMC for nitrogen (blue), carbon (black), hydropathy (green) and polar requirements (purple) for the subset of shuffled genetic codes with ERMC lower than the standard genetic code for (E) hydropathy, (F) the polar requirement, (G) carbon and (H) nitrogen. Vertical black-yellow line, ERMC of the standard genetic code for each property. P, permutation test.

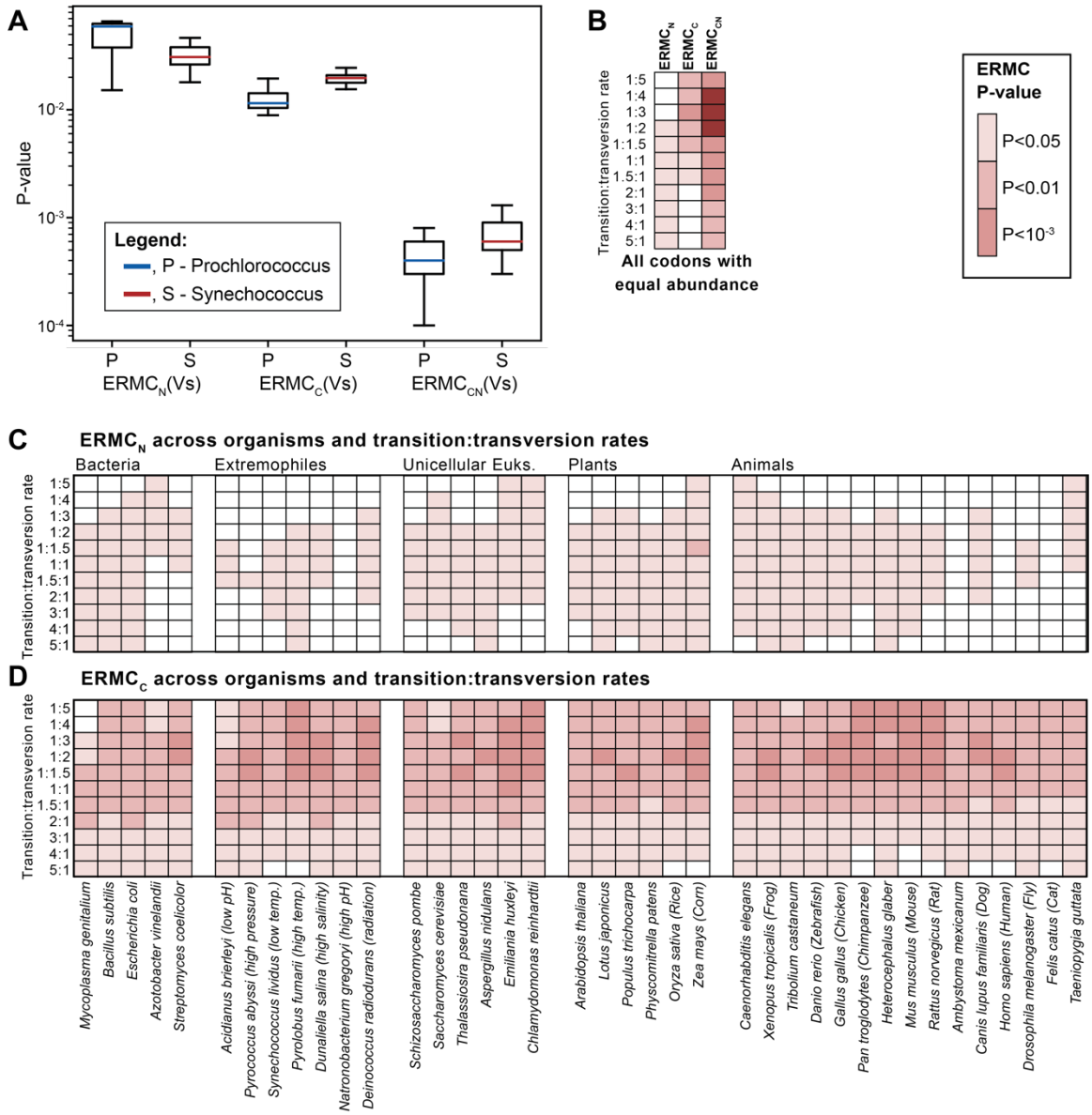


Fig. S11

(A) Box plots (line, median; box, IQR; whiskers, 5th and 95th percentiles) of P-values for the ERMC of the standard genetic code for nitrogen (left), carbon (center) and both (right) across 187 Prochlorococcus (P, blue) and Synechococcus (S, red) strains. (B) Heat map of ERMC P-values for nitrogen, carbon and both, for a theoretical case in which all codons are of the same abundance. (C,D) Same as Fig. 2C for $ERMC_N$ (D) and $ERMC_C$ (E) P-values**.**

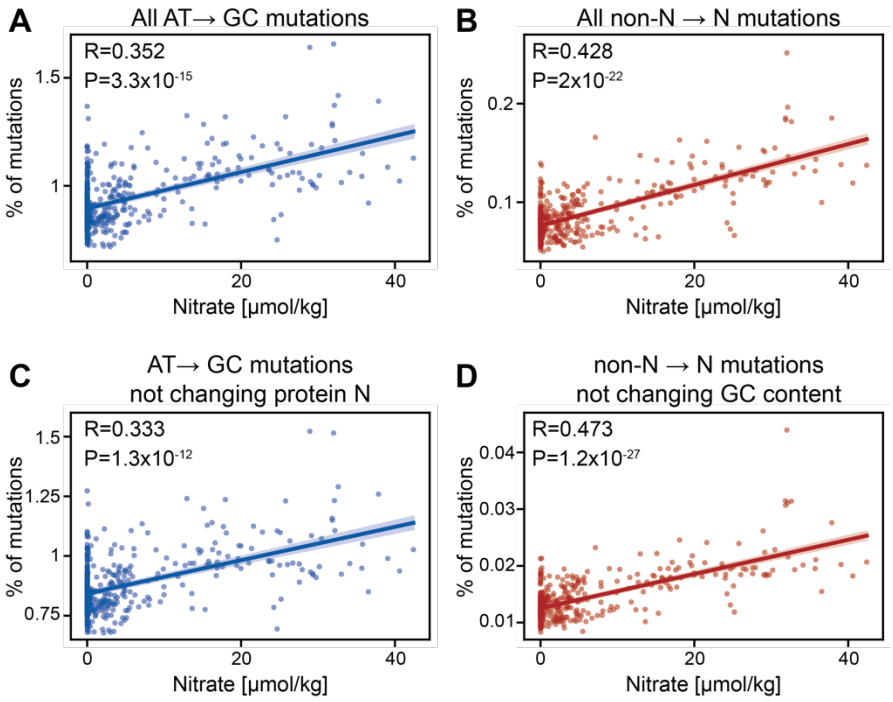


Fig. S12

Scatter plots of nitrate concentrations (x-axis) versus percent of mutations (y-axis) for (A) all AT to GC mutations; (B) All mutations in coding sequences changing an amino acid with no nitrogen in its side-chain to one with nitrogen in the side-chain (non-N to N mutations); (C) AT to GC mutations not changing protein nitrogen; (D) non-N to N mutations not changing GC content. R and P, Spearman correlation coefficient and P-value; line, linear fit; shaded area, 95% confidence interval.

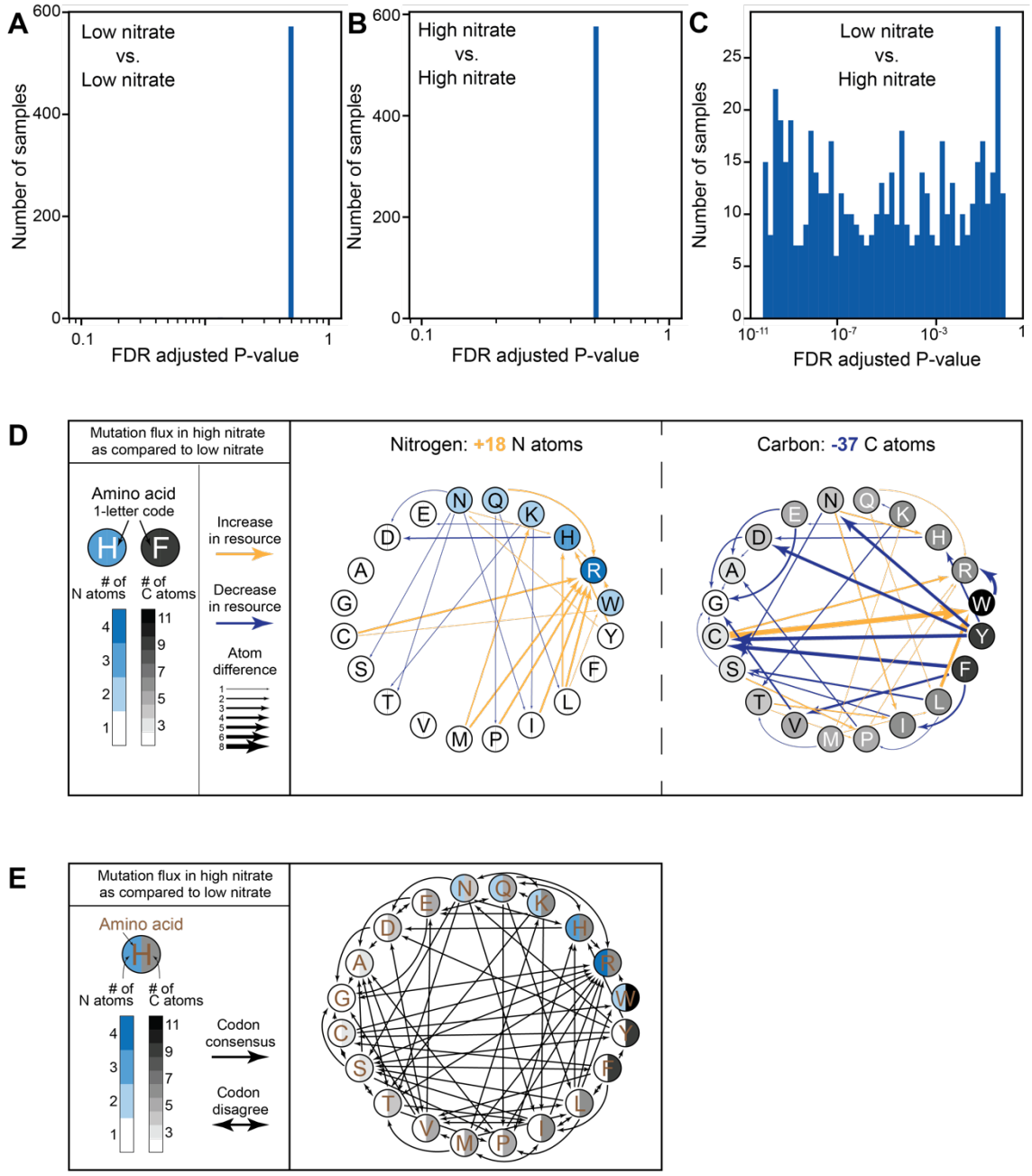
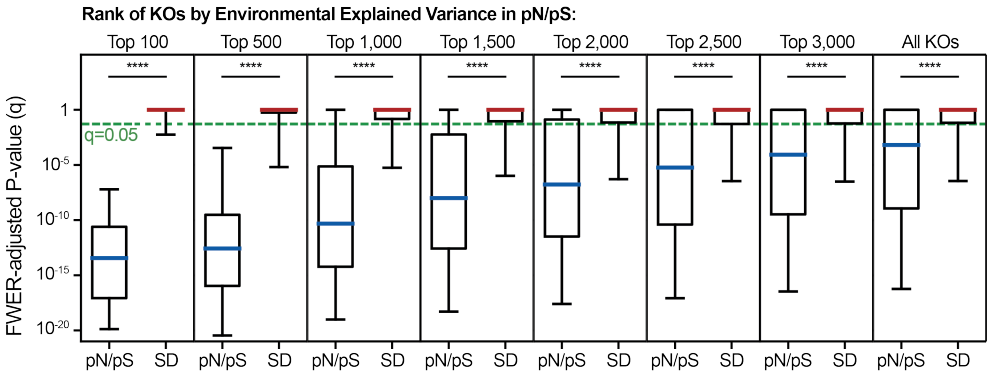


Fig. S13

(A-C) Histograms of the distribution of Mann-Whitney *U* P-values of codon-to-codon mutations compared between (A) 40 low-nitrate samples and 40 other low-nitrate samples selected randomly out of the 80 lowest-nitrate samples; (B) 40 high-nitrate samples and 40 other high-nitrate samples selected randomly out of the 80 highest-nitrate samples; (C) 40 low-nitrate samples and 40 high-nitrate samples selected randomly out of the 80 lowest and highest nitrate samples. (D) Depiction of mutation flux (Methods) common in high versus low environmental nitrate concentrations, affecting amino acid

979 nitrogen (left) and carbon (right) content. Yellow arrows, increase in resource; blue arrows, decrease in
980 resource; arrow thickness corresponds to number of atoms changed by mutation. (E) Same as D, showing
981 all fluxes, regardless of change in resources. Two-headed arrows mark cases where different codons of
982 the same amino acid have opposite mutation fluxes.

983 

984 **Fig. S14**

985 Box plots (line, median; box, IQR; whiskers, 10th and 90th percentiles) of P-values (t-test, multiple
986 hypothesis adjustment) between samples with high-N (more than 19.6 μmol/kg) and low-N (0 μmol/kg),
987 for KEGG KOs ordered by their EEV in the LMM, for both pN/pS (blue) and synonymous diversity (red).
988 ****, Wilcoxon signed-rank test $P<10^{-15}$.

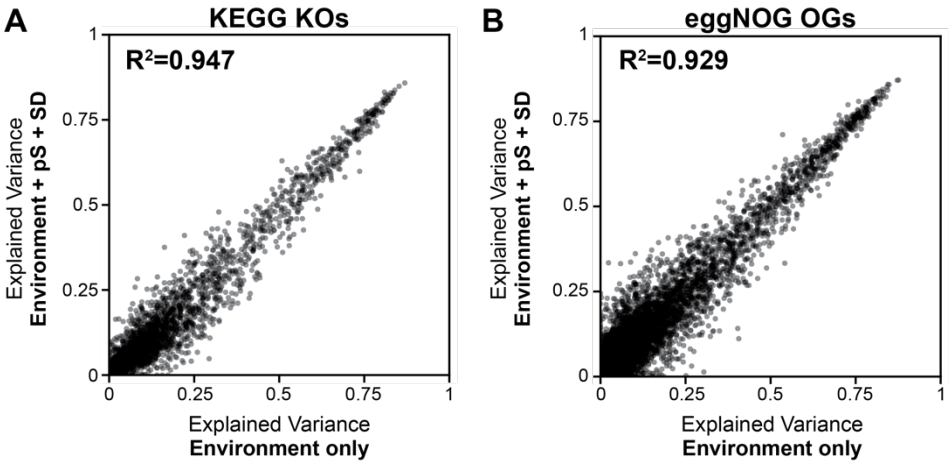


Fig. S15

(A,B) scatter plot of explained variance in a linear mixed model considering only environmental factors as random effects (x-axis) versus a model considering environmental factors as random effects plus dS (as a proxy for time) and synonymous diversity (as a proxy for effective population size) as fixed effects (y-axis) in all (A) KEGG KOs and (B) eggNOG OGs.

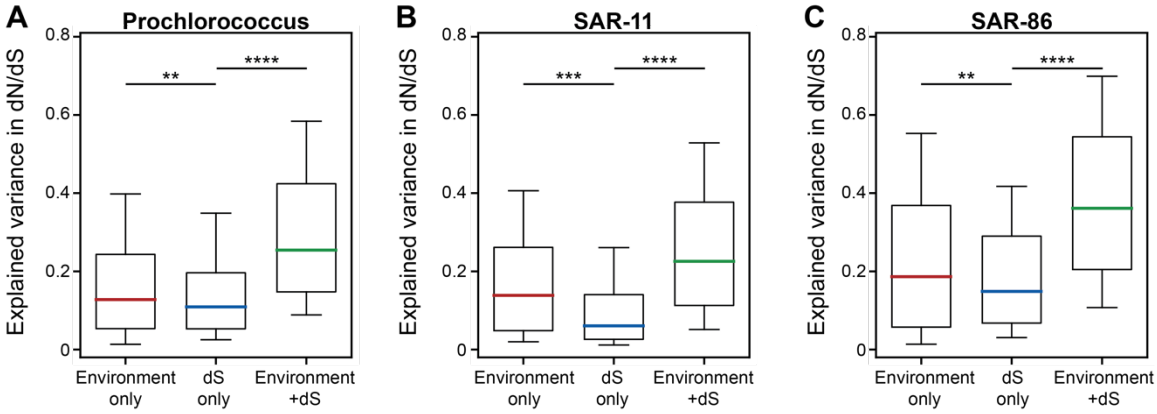


Fig. S16

Box plots (line, median; box, IQR; whiskers, 10th and 90th percentiles) of explained variance in dN/dS using linear models considering the environment only (left, red), dS only (middle, blue), or both (right, green). (A) Prochlorococcus, (B) SAR-11 clade and (C) SAR-86 clade. **, Wilcoxon signed-rank test $P<0.01$; ***, $P<10^{-5}$; ****, $P<10^{-16}$.

Table S1.

Contingency tables for each pair of cost functions for both nitrogen (n+) and carbon (c+), compared to PR (pr) and Hydropathy index (hyd), across 1 million simulated genetic codes. Each code is assigned to one of four bins: (1) surpassing the standard genetic code in both cost functions (<; <), (2) surpassing the standard genetic code only in element e cost (<; >=), (3) surpassing the standard genetic code only in the traditional cost function (>=; <), (4) not surpassing the standard genetic code in neither (>=; >=). Chi-square test of independence was applied to each contingency table.

Cost function a	Cost function b	Sign a	Sign b	Alternative code counts	Chi-squared P-value
n+	c+	<	<	128	1.32E-03
			>=	14106	
		>=	<	11802	
			>=	973964	
n+	hyd	<	<	270	2.71E-04
			>=	13964	
		>=	<	14953	
			>=	970813	
n+	pr	<	<	83	4.68E-16
			>=	14151	
		>=	<	13646	
			>=	972120	
c+	hyd	<	<	249	4.86E-07
			>=	11681	
		>=	<	14974	
			>=	973096	
c+	pr	<	<	442	4.29E-107
			>=	11488	
		>=	<	13287	
			>=	974783	

1009 **Table S2.**

1010 Comparisons between "radical" and "conservative" nonsynonymous substitutions using the polar
1011 requirement scale and amino-acid hydropathy index and multiple thresholds for each property. P,
1012 permutation test.

	Radical Threshold	Conservative Threshold	Proportion of significant Kos	Permutation test P-value
PR: Case 1	Absolute difference > 1	Absolute difference < 1	0.704	< 0.0001
PR: Case 2	Absolute difference > 2	Absolute difference < 1	0.737	< 0.0001
PR: Case 3	Absolute difference > 2	Absolute difference < 2	0.737	< 0.0001
PR: Case 4	Absolute difference > 3	Absolute difference < 1	0.928	< 0.0001
PR: Case 5	Absolute difference > 3	Absolute difference < 2	0.925	< 0.0001
PR: Case 6	Absolute difference > 3	Absolute difference < 3	0.913	< 0.0001
Hyd: Case 1	Absolute difference > 1	Absolute difference < 1	0.885	< 0.0001
Hyd: Case 2	Absolute difference > 2	Absolute difference < 1	0.880	< 0.0001
Hyd: Case 3	Absolute difference > 2	Absolute difference < 2	0.832	< 0.0001
Hyd: Case 4	Absolute difference > 3	Absolute difference < 1	0.950	< 0.0001
Hyd: Case 5	Absolute difference > 3	Absolute difference < 2	0.934	< 0.0001
Hyd: Case 6	Absolute difference > 3	Absolute difference < 3	0.932	< 0.0001
Hyd: Case 7	pre-mutation < -1 & post-mutation > 1 || pre-mutation > 1 & post-mutation < -1	Otherwise	0.943	< 0.0001
Hyd: Case 8	pre-mutation < -2 & post-mutation > 2 || pre-mutation > 2 & post-mutation < -2	Otherwise	0.968	< 0.0001

1013