Resource conservation manifests in the genetic code

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

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

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

[*]denotes equal contribution

[^]dzeevi@rockefeller.edu

Abstract

Nutritional limitation is a strong selective force in many environments, driving competition for resources across organisms. However, there is still much unknown about how selective pressures resulting from nutrient limitation have shaped microbial coding sequences. Here, we study this 'resource-driven' selection using metagenomic and single-cell data of marine microbes, with concurrent environmental measurements. We present evidence indicating that a significant portion of the selection exerted on marine microbes is explained by the environment and is strongly associated with nitrogen availability. We further identify specific mutations in coding sequences associated with environmental nutrient concentrations. Finally, we demonstrate that this optimization for resource conservation is encoded in the very structure of the standard genetic code, providing robustness against random mutations that increase carbon and nitrogen incorporation into protein sequences. Notably, this robustness generalizes to codon choices from multiple taxa across all domains of life, including the Human genome. Overall, we show that nutrient conservation exerts a significant selective pressure on coding sequences and may have even contributed to the evolution of the genetic code.

Introduction

Marine microbes, which make up more than half of the biomass in the ocean[1], are involved in key cycling of nutrients that make up all living systems. They account for nearly half of the carbon compound synthesis on Earth, thereby producing about 50% of breathable oxygen[2]. These microbes also cycle nutrients to perform numerous other important roles, such as biodegradation of complex organic material and fixation of atmospheric nitrogen, while flourishing in a wide range of environments with varying ambient conditions such as oxygen and nitrogen levels, light and temperature[3,4]. Nonetheless, evolutionary forces acting on ocean microbes are not fully understood[4].

With rapidly changing climate and environment, understanding the types of stress exerted on microbes in marine habitats is of paramount importance. Recent studies[4,5] provide evidence of high variability in the core genomic properties of marine microbes, including GC content and genome size, suggesting that this variability is linked to the concentrations of nutrients in the environment. Nitrogen and carbon are major limiting factors in the marine environment and their concentrations are typically inversely correlated[6]. It was shown that 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[7,8]. An opposite trend was shown for carbon[4]. These studies suggested 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 the effects of mistranslation errors and point mutations[9–12], specifically those leading to radical changes in amino acids. This error minimization is prominent among

49 theories regarding the origin of the genetic code[13–16], postulating selection to minimize the adverse
50 effect of point mutations and translation errors as the principal factor governing code evolution[17–
51 24]. Theories of error minimization typically assert that the genetic code had evolved to minimize
52 changes to protein function and structure and suggest structurally-informed amino acid properties
53 based on hydropathy and stereochemistry (e.g., the polar requirement scale[9,25] and hydropathy
54 index[26]), for its quantitative exploration. To our knowledge, optimizations in the genetic code have
55 not been studied thus far in terms of nutrient conservation.

56 Here, we use publicly available metagenomic and single-cell data from marine habitats across
57 the globe, coupled with their corresponding measurements of environmental conditions, to
58 characterize and quantify the selective pressure resulting from nutrient limitation. For this
59 analysis, we estimate selection in 746 marine samples (**Fig. S1A**) and quantify its 'resource-
60 driven' component, while controlling for potential confounders such as effective population size
61 and time. We find that 'resource-driven' selection is a ubiquitous purifying selective force
62 associated with environmental nutrient concentrations, and particularly nitrogen. We show that
63 genes are affected by this selective force as a function of their expression and propensity for
64 secretion, which determines their impact on the organismal nutritional budget. We further identify
65 specific mutations in coding sequences associated with environmental concentrations of nitrogen,
66 in nitrate form. Based on the patterns of these mutations, we hypothesize that the structure of the
67 genetic code, known to minimize the effects of DNA mutations on the structure and function of
68 proteins, also minimizes resource expenditure given a random mutation. We test this hypothesis
69 and show that the choice of codons used in the genome of the organism, along with the structure
70 of the genetic code, provide robustness against random mutations that increase carbon and
71 nitrogen incorporation into protein sequences. We further demonstrate that this optimization for
72 resource conservation, encoded in the standard genetic code, is not confounded by previously
73 reported optimizations of protein structure and function. We extend this observation to multiple

74 mutation rates and codon distributions across the tree of life, and suggest resource conservation
75 as an important contributor to the evolution of the genetic code.

76

77 **Results**

78 **Single nucleotide polymorphisms in marine microbial genes reveal significant purifying**
79 **selection**

80 To better understand the mechanisms governing resource-driven selection acting on marine
81 microbes, we sought to characterize, at the single nucleotide level, how coding sequences of
82 marine microbes are affected by resource availability in their environment. To this end, we devised
83 a computational pipeline that calculates metrics of selection from marine metagenomic samples
84 (**Fig. S1B**). We downloaded metagenomic sequences of 746 samples from the Tara oceans[3]
85 (n=136); and bioGEOTRACES[27] (n=480) expeditions; and the Hawaii Ocean- (HOT; n=68); and
86 Bermuda Atlantic- (BATS; n=62) Time-series[27] (**Fig. S1A**; Methods). We aligned these reads to
87 the Ocean Microbiome Reference Gene Catalog (OM-RGC)[3], a database of genes from marine
88 environments that is accompanied by functional and taxonomic information. We then searched
89 for single nucleotide polymorphisms (SNPs) in genes that had sufficient high-quality coverage
90 (**Fig. S1B**; Methods). Overall, we found 71,921,864 high-confidence SNPs which passed our
91 stringent filtering (Methods), in a total of 1,590,843 genes.

92 To quantify purifying selection on different gene functions, we annotated genes from the OM-RGC
93 database to inform their functional group membership with either KEGG orthology (KO[70]) or
94 eggNOG orthologous group (OG[53]; Methods). We then calculated, using called SNPs, the ratio of
95 nonsynonymous polymorphisms (pN), which lead to a change in the resulting amino acid, to
96 synonymous polymorphisms (pS), which maintain the coded amino acid (pN/pS; Methods), for
97 each orthologous group in each sample. This measure approximates dN/dS, which requires

98 genomic sequences at the strain-level and quantifies the magnitude and type of selection
99 (purifying, neutral or positive) exerted on protein-coding sequences[28,29]. Interestingly, across all
100 samples, we found pN/pS ratios to be close to zero with an average of 0.074 across all OGs (CI
101 [0.072, 0.075]; **Fig. S1C**) and 0.079 across all KOs (CI [0.077, 0.080]; **Fig. S1D**), indicating
102 widespread purifying selection in the marine environment.

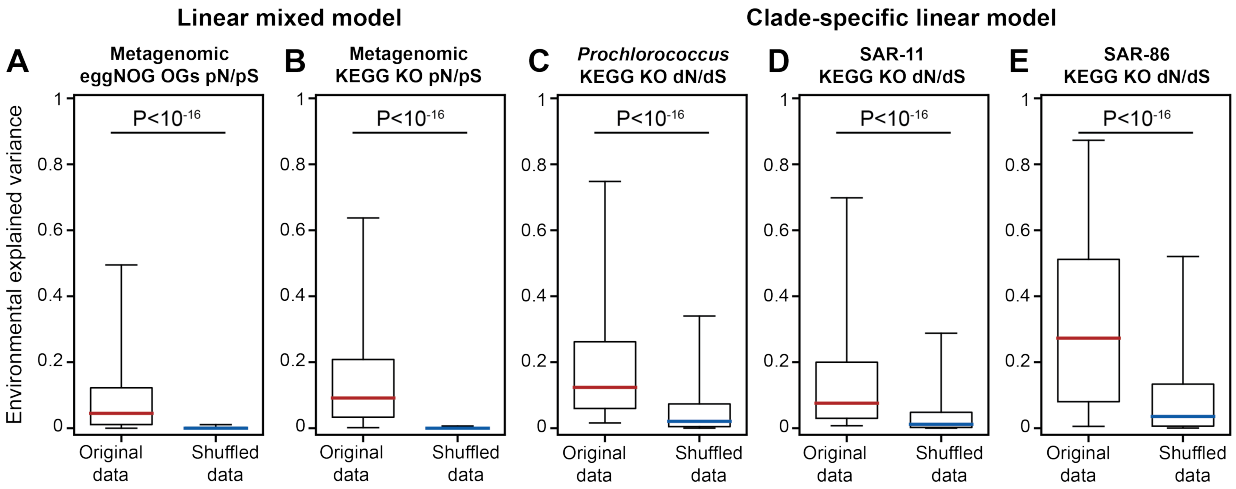
103

104 **Strong resource-driven selection apparent across marine microbial genes**

105 Based on recent studies[4,7,8], we hypothesized that nutrient availability is a central driver of this
106 purifying selection, affecting the evolution of coding sequences in marine microbes. Therefore,
107 we sought to quantify the environmental contribution to this selective pressure, estimated by
108 pN/pS. We considered environmental measurements taken alongside each sample, including the
109 depth of the sample, water temperature and salinity, and the concentrations of nitrate, nitrite,
110 oxygen, phosphate and silicate (**Fig. S2A-H**; Methods). All these measurements presented
111 consistent correlation patterns with the pN/pS of many KEGG and eggNOG orthologs (**Fig. S3**).
112 Specifically, we found lower pN/pS ratios at shallow depths and low nitrate concentrations,
113 indicating that these conditions may invoke stronger purifying selection.

114 However, as these environmental measurements are also highly correlated with each other (e.g.,
115 nitrate and phosphate concentrations increase and temperature decreases with depth; **Fig. S2I**),
116 we cannot accurately estimate their individual effects. We therefore sought to estimate the fraction
117 of variance in pN/pS explained by the environment while accounting for these covariations. We
118 used a linear mixed model (LMM) with variance components, commonly used in population
119 genetics[30] (Methods), which allows to estimate the fraction of variance in pN/pS ratios (dependent
120 variable) that is explained by the environment (random effect) while controlling for the complex
121 correlation structure between the environmental parameters, in a data-driven manner. We term
122 the fraction of the variance in pN/pS that is explained by resource availability 'environmental

123 explained variance' (EEV; Methods). Across both KEGG and eggNOG orthologs we found that a
124 significant fraction of the variance in pN/pS can be attributed to the environment (**Fig. 1A,B,**
125 **S4A,B**; Mann-Whitney *U* test $P<10^{-16}$). We note that this association between environmental
126 measurements and the magnitude of purifying selection is significant even after controlling for
127 potential confounders such as time or effective population size (**Fig. S4C,D**; Mann-Whitney *U* test
128 $P<10^{-16}$; **Supplementary Note 1**) as well as in specific environmental niches (**Fig. S5**; Mann-
129 Whitney *U* test $P<10^{-20}$; **Supplementary Note 2**).



130

131 ***Figure 1. Analysis of pN/pS reveals resource-driven selection.*** *(A) Box plot (line, median; box,*
132 *IQR; whiskers, 5th and 95th percentiles) of the variance of eggNOG OG pN/pS that was explained by*
133 *the environment in a linear mixed model (LMM; Methods) as compared to the same data with shuffled*
134 *labels. (B) Same as A, for KEGG KO pN/pS. (C) Box plot (line, median; box, IQR; whiskers, 5th and*
135 *95th percentiles) of the variance of KEGG KO dN/dS in genus Prochlorococcus that was explained by*
136 *the environment in a linear model. (D) Same as C, for the SAR-11 clade. (E) Same as C, for the SAR-*
137 *86 clade. P, Wilcoxon signed-rank test.*

138

139 As opposed to pN/pS, in which the population-wide ratio of nonsynonymous to synonymous
140 polymorphisms is calculated from metagenomic sequences, dN/dS estimates the ratio of
141 substitution rates across a phylogeny. While dN/dS calculation requires genomic sequences at
142 the strain-level that are available for only a small number of samples and from limited strains, it

143 allows for a more accurate quantification of purifying selection in the taxa represented in the
144 phylogeny. We therefore sought to confirm our findings thus far using dN/dS. To this end, we
145 used benchmarking data in the form of assembled genomes from uncultivated single cells from
146 dominant and representative lineages of the surface ocean (SAR-11, SAR-86 and
147 Prochlorococcus)[31], and their corresponding environmental measurements (Methods). We
148 calculated dN/dS for KEGG orthologs in each lineage. Given the small number of samples, we
149 estimated the fraction of variance explained by the environment using a linear model with
150 environmental covariates (Methods). Across all bacterioplankton lineages comprising the
151 benchmarking data, we found that a significant portion of the variance in dN/dS ratios calculated
152 from single-cell data is explained by the environment (SAR-11, SAR-86, Prochlorococcus;
153 Wilcoxon signed-rank test $P<10^{-16}$; **Fig. 1C-E, S4E-G**). Similar to our original LMM, we note that
154 this association holds, across all three bacterioplankton lineages, even after controlling for time
155 as a confounder (Wilcoxon signed-rank test $P<10^{-16}$; **Fig. S6**; **Supplementary Note 3**) and for
156 phylogeny, using phylogenetic generalized least-square regression (Methods). These results
157 validate our findings using pN/pS calculated from metagenomic sequences and demonstrate the
158 association between selective pressure and environmental conditions, even in a single clade. This
159 indicates that associations of pN/pS or dN/dS with environmental variables are likely not
160 confounded by organismal properties which may differ in tropism across environments, but rather
161 exhibit a consistent trend even within a single clade.

162 We next sought to find the dominant environmental factor, with respect to association with pN/pS
163 ratios, as a post hoc analysis. We found that nitrate is more strongly correlated with pN/pS ratios
164 than any other environmental parameter (**Fig. S7**; Kolmogorov-Smirnov test; $P<10^{-30}$ for all
165 comparisons). While we cannot rule out the possibility of an unmeasured confounding factor, such
166 as labile dissolved organic carbon, these results are supported by previous studies[4,5,7] which

167 examined nutrient conservation in DNA and protein sequences of marine microbes and showed
168 their association with environmental nitrogen concentrations.

169

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

171 In light of these findings which are supported by previous studies[4,7], we hypothesize that
172 mutations that increase the nitrogen requirements of cells are selected against, especially in
173 nitrogen-limited oceans. Our hypothesis implies that highly expressed genes would be under
174 stronger purifying selection[7,8,32] (illustrated in **Fig. S8**). In highly expressed genes, a mutation
175 leading to the incorporation of additional resources would have a stronger effect on the nutrient
176 economy of the cell, as one highly transcribed DNA sequence could translate to thousands of
177 proteins, each consuming more resources. To quantitatively corroborate this and examine
178 whether highly expressed genes have higher EEV as compared to lowly expressed genes, we
179 used an additional expression dataset for marine microbial genes[33] to rank KEGG orthologs by
180 their mean expression (Methods). We found that the 100 most highly expressed KEGG KOs had
181 a significantly higher fraction of the variance in their pN/pS explained by the environment (i.e.,
182 higher EEV), as compared to the 100 least expressed ones (**Fig. 2A, S9**; Mann-Whitney *U* test
183 $P<10^{-9}$). We replicated these results using dN/dS ratios calculated from single-cell data within
184 each bacterioplankton lineage (Mann-Whitney *U* test $P<10^{-7}$). This finding provides a data-driven
185 evolutionary perspective to the theory and experiments showing lower resource incorporation in
186 highly expressed genes[7,8].
187 Given the 'resource sensitivity' observed in highly expressed genes, we further hypothesized that
188 genes coding for extracellular proteins will also exhibit a higher EEV as compared to intracellular-
189 protein coding genes, as in this case the resources excreted from the cell cannot be recycled. We
190 indeed found that extracellular protein-coding genes had significantly higher EEV as compared to
191 other gene groups (Methods; **Fig. S10**; $P<0.05$). In summary, the high EEV in resource-

consuming genes, (i.e., highly expressed and extracellular protein-coding), further strengthens our results regarding the breadth of resource-driven selection.

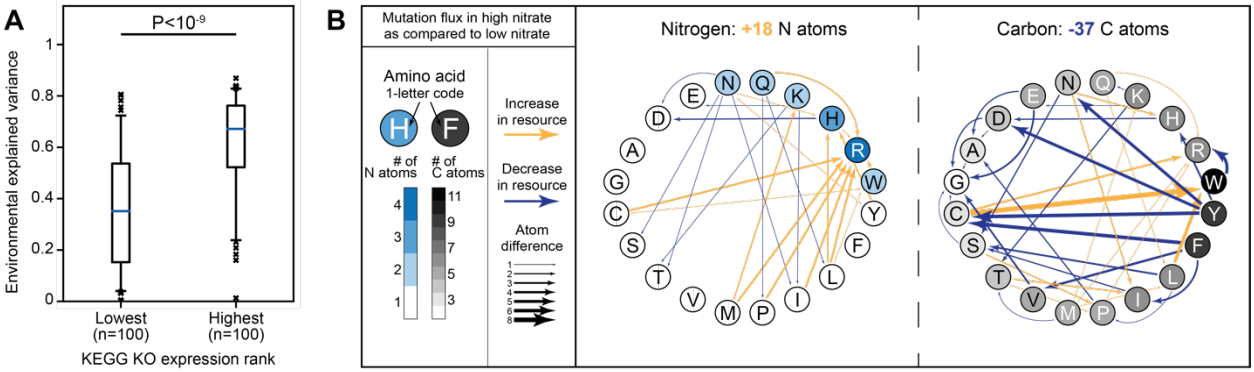


Figure 2. Effects of resource-driven selection on coding sequences. *(A) Box plot (line, median; box, IQR; whiskers, 5th and 95th percentiles) of variance in pN/pS explained by the environment in the 100 lowest and highest expressed KEGG KOs. P, Mann-Whitney U test. (B) Depiction of mutation flux (Methods) common in high versus low environmental nitrate concentrations, affecting amino acid nitrogen (left) and carbon (right) content. Yellow arrows, increase in resource; blue arrows, decrease in resource; arrow thickness corresponds to number of atoms changed by mutation.*

Resource-driven selection drives specific amino acid changes

As our results indicate that mutations leading to an addition of resources would be more common in nutrient-rich environments, we sought to quantify how they manifest in specific amino acid changes. To this end, we compared the codon mutation frequencies in low- and high-nitrate samples, after accounting for simplex-related confounders (Methods), and found significant differences in codon mutation frequencies (**Fig. S11A-C**). We sought to examine the typical change in nutrient consumption in varying nitrate concentrations. We thus defined mutation flux, as the ratio between a codon mutation and its reverse, thereby measuring the typical flux of nutrients; and estimated it using the log-odds ratio between the two (e.g., $log\,(p(AAA \rightarrow AAC)/p(AAC \rightarrow AAA))$). Notably, across all the mutations significantly more prevalent in samples from high nitrate environments (Methods), averaged across amino acids, we find a total increase in nitrogen (**Fig. 2B**; 18 N atoms summed across all significant amino acid changes, P=0.0082),

214 decrease in carbon (**Fig 2B**; -37 atoms, P=0.0165), decrease in sulfur (-6 atoms, P=0.009), a
215 decrease in molecular weight (-508.91 g/mol, P=0.0193) and a non-significant decrease in oxygen
216 (-6 atoms, P=0.1505). These results provide additional evidence that the higher the nitrate
217 concentrations are, the weaker the selection against mutations leading to higher nitrogen
218 incorporation into protein sequences. While nitrate concentrations increase with depth (**Fig. S2I**),
219 labile dissolved organic carbon concentrations typically decrease[34]. We therefore also show the
220 specific codon mutations reducing carbon incorporation into proteins in nitrogen rich and typically
221 carbon poor, environments (**Fig. 2B, S11D**).

222 Our results are supported by previous observations regarding genomic and proteomic changes
223 associated with environmental concentrations of nitrate[4]. Mutations in nitrogen-rich and typically,
224 carbon-poor environments were shown to drive an increase in genomic GC content (which
225 increases nitrogen requirements), as well as by higher nitrogen and lower carbon in protein
226 sequences. Using SNP data, we corroborate these findings, and show that mutations leading to
227 higher GC content and those leading to higher nitrogen proteins are more prevalent in high-nitrate
228 waters as compared to low-nitrate (**Fig. S12A,B**). We note that mutations leading to higher GC
229 content (higher nitrogen DNA) often result in higher nitrogen proteins due to the structure of the
230 genetic code[4,7]. We further deconvolve these two types of mutations and find that while the ratio
231 of mutations that lead to higher nitrogen DNA but do not change protein nitrogen-content
232 increases with environmental nitrate (R=0.333; **Fig. S12C**), the ratio of mutations that increase
233 protein nitrogen but do not change DNA nitrogen-content is even more strongly correlated with
234 nitrate concentrations (R=0.473; **Fig. S12D**). We hypothesize that this is due to the higher copy
235 number of protein as compared to DNA and RNA, often in the order of 10^2-10^5 protein copies per
236 RNA molecule[35,36] leading to higher cost per gene, and thus to stronger selective pressure. Our
237 results suggest a differential effect of mutations on nutrient incorporation in DNA as compared to
238 protein sequences.

239

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

241 We observed selection against DNA mutations which result in excess incorporation of nutrients,
242 such as nitrogen and carbon, into proteins. However, not all mutations have the same effect on
243 protein sequences due to constraints imposed by the structure of the genetic code, which was
244 shown to minimize the impact of point mutations on protein structure and function[25,26,37]. We
245 therefore hypothesized that the genetic code also minimizes the impact of point mutations on
246 nutrient incorporation into proteins. Specifically, we hypothesized that the genetic code acts as a
247 buffer between the DNA level, where mutations occur, and the protein level, where resource-
248 driven selection is exerted, such that the expected cost of a random DNA mutation, in terms of
249 added resources to proteins, is minimized.

250 To rigorously test this hypothesis, we first defined a cost function for each element e (e.g., carbon,
251 nitrogen), such that the 'tariff' of a single mutation is the added number of amino acid atoms
252 resulting from the mutation. As an example, a missense mutation from codon CCA to codon CGA
253 results in an amino acid change from proline to arginine, and an increase of 3 nitrogen atoms and
254 one carbon atom, setting the nitrogen cost of such mutation to 3 and the carbon cost to 1 (**Fig.**
255 **3A**).

256 To estimate the cost of a random mutation on each element, across the entire genetic code, we
257 calculated, for nitrogen, carbon and oxygen, the Expected Random Mutation Cost (ERMC) for the
258 standard genetic code V_S:

259 $$ERMC_e(V_S) = \sum_{v,v' \in V_S} P(v)P(mut(v,v'))c_e^+(v,v'),$$

260

261 where $P(v)$ is the abundance of codon v, calculated from all marine samples; $P(mut(v,v'))$ is the
262 probability of mutation from codon v to codon v', set to be the relative abundance of the single

263 nucleotide mutation driving this codon change (e.g. for mutation from GCA to CCA, we use the
264 abundance of G-to-C transversions), calculated from all mutations observed in fourfold
265 degenerate codons to avoid sites under strong selection; and the cost function $c_e^+(v, v')$ is the
266 number of atoms of element e added to the post-mutation amino acid compared to the pre-
267 mutation one (Methods). For the standard genetic code, and typical codon abundances and
268 mutation rates calculated from marine microbes, we report an ERMC of 0.440, 0.158 and 0.163
269 for carbon, nitrogen and oxygen respectively, corresponding to an average increase of this
270 number of atoms per random mutation (**Fig. 3B**).

271 To check if the genetic code, along with codon abundances and mutation rates in marine
272 microbes, is indeed robust to resource-consuming mutations, we compared it to other hypothetical
273 codes. We simulated alternative genetic codes by randomizing the first and second position in all
274 codons one million times, while constraining the permutation in stop codons, to create a null
275 distribution of ERMC (Methods). We found that the standard genetic code, common to most life
276 forms, is parsimonious in terms of carbon and nitrogen utilization, given a random mutation,
277 manifested by minimization of the ERMC for nitrogen (**Fig. 3B**; $ERMC_N$ P=0.0142) and carbon
278 (**Fig. 3B**; $ERMC_C$ P=0.0119), while in the ERMC for oxygen we did not find a significant trend
279 (**Fig. 3B**; $ERMC_O$ P=0.8704).

280 We compared the extent of the robustness observed in the standard genetic code against the
281 addition of carbon and nitrogen to well-studied structure and function error minimization
282 mechanisms[38–41]. To this end, we calculated ERMC for changes in amino acid hydropathy[26] and
283 the polar requirement (PR)[9,25], structurally-informed amino acid properties commonly used to
284 determine code error minimization (Methods). We found that these optimization mechanisms are
285 of a similar magnitude to those shown here for nitrogen and carbon (Methods; **Fig. S13C,D**;
286 $ERMC_{hyd}$ P=0.0152; $ERMC_{PR}$ P=0.0137).

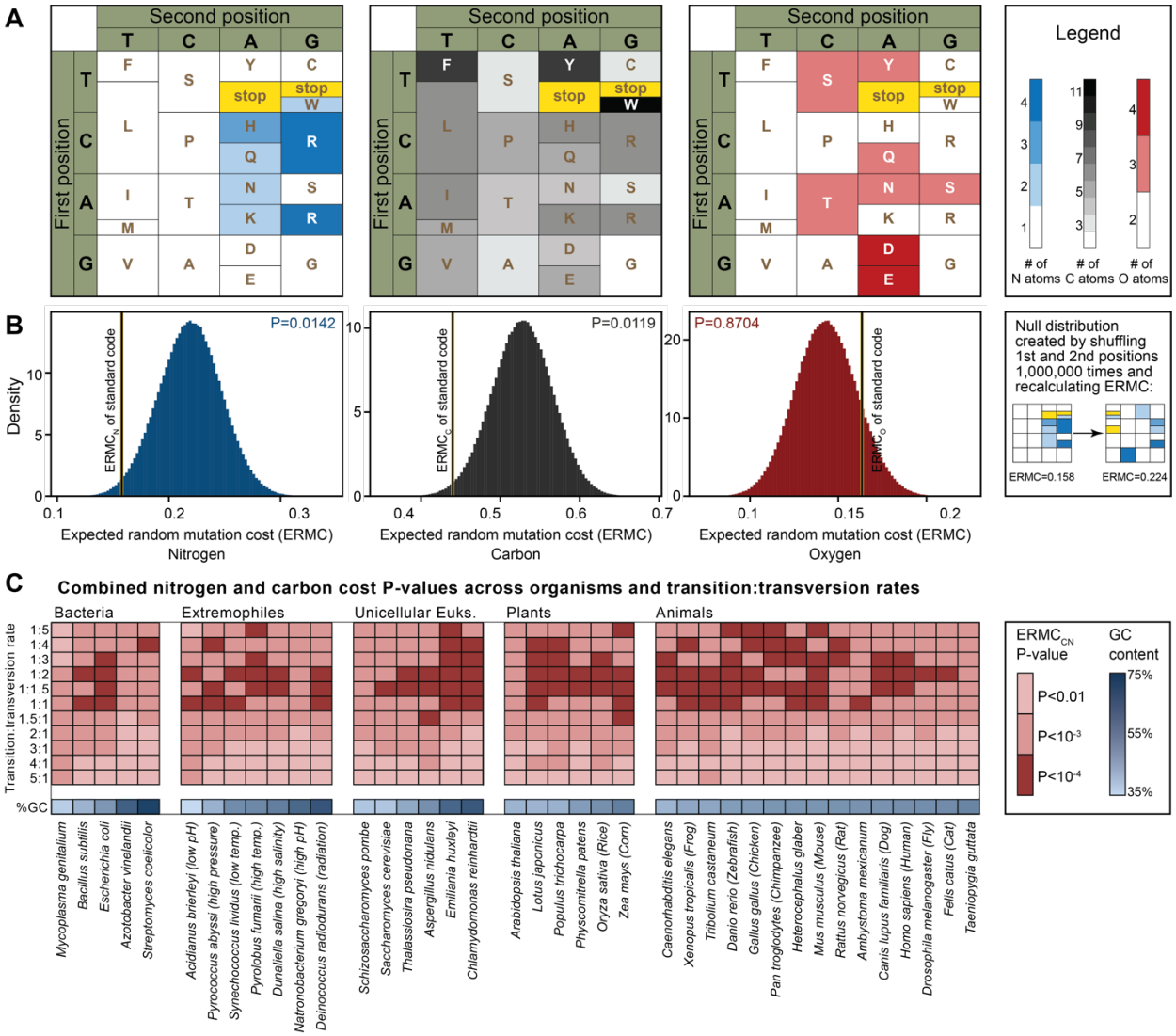


Figure 3. Resource-conservation is facilitated by the genetic code. *(A) Nitrogen (left), carbon (center) and oxygen (right) content of different amino acids depicted along their positions in the standard genetic code. (B) Histograms of the expected random mutation cost (ERMC), in 1,000,000 random permutations 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. (C) Heat map of* $ERMC_{CN}$ *P-values across 39 organisms and 11 transition:transversion rates. Organisms in each of the groups are ordered according to the GC content of their coding sequences.*

297 We next sought to confirm that our elemental cost function is not confounded by these known
298 optimization mechanisms. To this end, we devised a hierarchical model which examined the
299 subset of genetic codes (out of the 1M hypothetical codes) that have a lower ERMC than the
300 standard genetic code for PR or hydropathy, and tested whether this subset is also optimized for
301 nitrogen or carbon (Methods). If nutrient optimization exists separately from structural
302 optimization, we expect the standard genetic code to be optimized for carbon and nitrogen as
303 compared to this subset, featuring significantly lower $ERMC_N$ and $ERMC_C$. We find that out of the
304 13,729 hypothetical codes that have $ERMC_{PR}$ lower than the standard genetic code, only 83 have
305 lower $ERMC_N$ (**Fig. S13E**; P=0.0058) and only 442 have lower $ERMC_C$ (**Fig. S13E**; P=0.037).
306 Similarly, out of 15,223 hypothetical genetic codes that have $ERMC_{hyd}$ lower than the standard
307 genetic code, only 270 have lower $ERMC_N$ (**Fig. S13F**; P=0.019) and only 249 have lower $ERMC_C$
308 (**Fig. S13F**; P=0.0208). This is in contrast with the observed overlap between hydropathy and PR:
309 out of the 15,223 hypothetical codes that have $ERMC_{hyd}$ lower than the standard code, 6,736 have
310 lower $ERMC_{PR}$ (**Fig. S13F**; P=0.4425). This result indicates that the detected carbon and nitrogen
311 optimization in the genetic code is not confounded by previously reported optimization properties
312 such as hydropathy and PR.

313 Remarkably, only 128 out of 1M randomized genetic codes were more resource-robust than the
314 standard genetic code in conservation of nitrogen and carbon together (**Fig. S13G,H**; $ERMC_{CN}$
315 $P=1.28x10^{-4}$). We note that this number is significantly smaller than the number of shuffled codes
316 expected to have both a lower $ERMC_C$ and $ERMC_N$ (Chi-square test of independence P=0.0013;
317 **Table S1**). This finding suggests that there is a smaller overlap than expected between the
318 nitrogen and carbon optimizations that are encoded in the genetic code, possibly driven by a small
319 overlap between the positions of high nitrogen amino acids and high carbon amino acids. This
320 property of the genetic code potentially enables efficient optimization for both carbon and nitrogen
321 at the same time. Altogether, these results indicate that resource optimization in marine microbes

322 is driven by the structure of the genetic code, alongside specific amino acid choices. We also
323 show that this nutrient optimization mechanism in the genetic code is of the same order of
324 magnitude as known optimization mechanisms, and is not confounded by them.

325

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

327 To show that the resource-robustness of the genetic code was not limited to our dataset and
328 analytic approach, we calculated the ERMC of 187 species of genera *Prochlorococcus* and
329 *Synechococcus*, major marine microbial clades. We calculated codon abundances and mutation
330 rates using published protein-coding sequences[5] and the accepted transition:transversion rate of
331 2:1[42] (Methods). By testing the ERMC of the standard genetic code in these species against a
332 null distribution generated using these known parameters rather than ones inferred from marine
333 samples, we corroborated significant conservation of carbon, nitrogen, and both elements
334 combined (**Fig. S14A**; $ERMC_C$ mean P=0.013, P=0.020; $ERMC_N$ mean P=0.049, P=0.032;
335 $ERMC_{CN}$ P=0.0004, P=0.0007 for *Prochlorococcus* and *Synechococcus*, respectively).

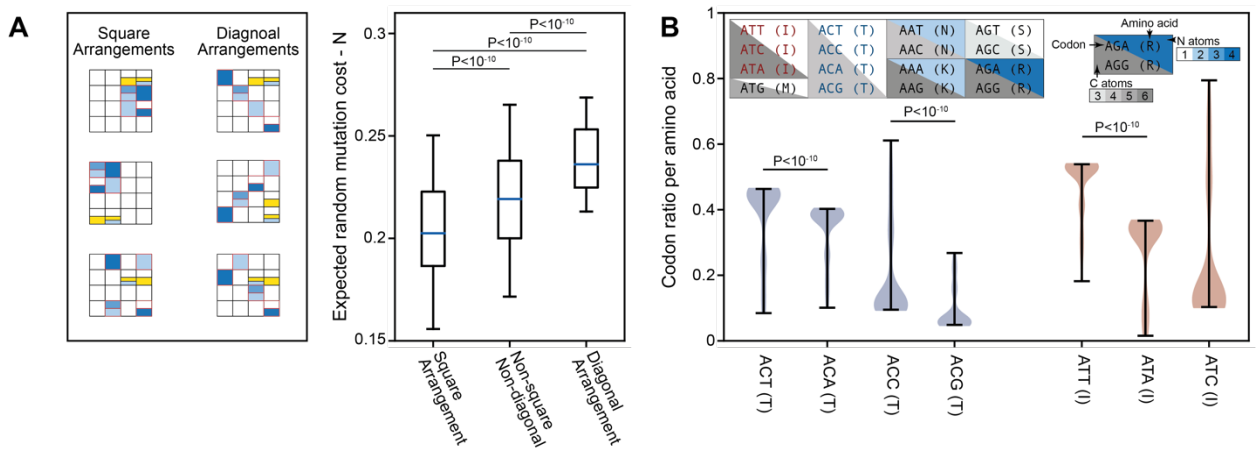
336 To explore whether this nutrient conservation optimization mechanism of the genetic code
337 extends across organisms, we performed a similar calculation using codon abundances from 39
338 organisms across all domains of life, including all human protein-coding sequences, and tested a
339 range of transition:transversion rates (Methods). We find that, similarly to marine microbes, the
340 genetic code features optimization in terms of resource utilization for all tested organisms,
341 manifested by a significant minimization of the combined ERMC of nitrogen and carbon in all
342 transition:transversion rates tested (P<0.01, **Fig. 3C**). Moreover, we find significant optimization,
343 albeit of a lower magnitude, even in the theoretical case where all codon abundances are the
344 same (**Fig. S14B**). The codon abundances of a great majority of organisms also demonstrate
345 significant minimization of ERMC in nitrogen (**Fig. S14C**) and carbon (**Fig. S14D**), given a random
346 mutation, for a wide range of transition:transversion rates. These results indicate that resource

347 optimization in the genetic code transcends taxonomy, codon choices, and mutation rates. It
348 shows that the genetic code may have structural properties that make it robust in terms of
349 resource consumption.

350

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

352 We next wished to examine the organizing principles that drive the resource optimization evident
353 in the standard genetic code. We observed that codons of the nitrogen-rich amino acids histidine,
354 glutamine, asparagine, lysine and arginine span only two nucleotides in their first position and two
355 in their second position. We define this organization to be a 'square' arrangement (**Fig. 4A**;
356 Methods), and hypothesize that, as compared to other arrangements, it amplifies nitrogen
357 conservation. Specifically, in the square arrangement, codons coding for some amino acids
358 (alanine, valine, phenylalanine, and several leucine and serine codons) require at least two
359 mutations to increase the number of nitrogen atoms in the resulting amino acid. This is in contrast
360 to other hypothetical arrangements, including a 'diagonal' arrangement in which nitrogen-rich
361 amino acid codons span all possible nucleotides in the first and second positions (**Fig. 4A**;
362 Methods). We suggest that the diagonal arrangement would be nutrient-wasteful, as in such
363 arrangements a single mutation both in the first or second codon position, could increase the
364 nitrogen content of a protein. To rigorously test this hypothesis, we generated 10,000 random
365 genetic codes, with 220 arrangements happening to embody a square structure, and 127 a
366 diagonal one. We found that, when compared to all other possible arrangements, square
367 arrangements present a significantly lower $ERMC_N$ (**Fig. 4A**; Mann-Whitney *U* test $P<10^{-10}$) while
368 diagonal arrangements exhibit a significantly higher $ERMC_N$ (**Fig. 4A**; Mann-Whitney *U* test $P<10^{-}$
369 ^{10}). This result demonstrates that resource optimization in the genetic code is driven by structural
370 principles, perhaps underlying the optimization observed across kingdoms.



371

372 ***Figure 4. Structural properties and codon usage bias underlying optimality in the genetic code.***
373 *(A) Box plot (line, median; box, IQR; whiskers, 5th and 95th percentiles) of* $ERMC_N$ *of square*
374 *arrangements (left) and diagonal arrangements (right, Methods), as compared to all other*
375 *arrangements (center) out of 10,000 randomized arrangements of the code. P, Mann-Whitney U test.*
376 *(B) Violin plot of codon usage among 187 species of Prochlorococcus and Synechococcus showing*
377 *significant preference of threonine codons ACT and ACC as compared to ACA and ACG, and of*
378 *isoleucine codon ATT as compared to ACA. P, Wilcoxon signed-rank test.*

379

380 Finally, we hypothesized that codon usage for a single amino acid may also be biased due to the
381 differential cost of a random mutation for each codon. We therefore examined all amino acids
382 coded by codons with adenine in their first position, focusing on codon usage of the amino acid
383 threonine. We note that a C-to-G transversion in the second position for codons ACT and ACC
384 yields serine (AGT and AGC, respectively), and that the same mutation for codons ACA and ACG
385 yields arginine (AGA and AGG, respectively; **Fig. 4B**, inset). Arginine has higher carbon and
386 nitrogen content than serine. We thus hypothesized that for a cell to conserve nutrients in case of
387 a random mutation, codons ACT and ACC should have a higher abundance than codons ACA
388 and ACG, respectively, given a known GC-bias that is a genomic quantity. To test this hypothesis,
389 we examine codon usage in 187 *Prochlorococcus* and *Synechococcus* strains, and show a
390 significantly higher use of ACT as compared to ACA (**Fig. 4B**; Wilcoxon signed-rank test $P<10^{-20}$)
391 and ACC as compared to ACG (**Fig. 4B**; Wilcoxon signed-rank test $P<10^{-20}$). Similarly, isoleucine

392 codon ATT has higher abundance as compared to ATA (**Fig. 4B**; Wilcoxon signed-rank test $P<10^{-20}$).
393
394 These results demonstrate that resource conservation is a central driving force in selection
395 processes that may guide codon usage and thereby affect not only protein sequence but also
396 cellular translation efficiency.

397 **Discussion**

398 In this work, we characterize and quantify the selective forces exerted by nutrient availability on
399 protein coding genes in marine environments. As metrics for purifying selection, we calculate
400 pN/pS from marine metagenomic samples and dN/dS from single-cell genome assemblies. We
401 control for potential confounders and show that a substantial fraction of the variance in these
402 metrics could be attributed to environmental factors while highlighting a strong association with
403 nitrate concentrations. We further show that the variance in pN/pS and dN/dS that is attributable
404 to environmental factors increases with the potential impact of a gene on the cellular nutritional
405 budget, with higher EEV in highly expressed and extracellular protein-coding genes. Using single
406 nucleotide polymorphisms from across marine samples, we characterize the typical mutations in
407 high- versus low-nitrate marine environments and show that the former drive incorporation of
408 additional nitrogen-rich amino acids and fewer carbon-rich amino acids to protein sequences.
409 Finally, we show that a significant portion of DNA mutations may not result in increased nutrient
410 incorporation into protein sequences, due to the pattern of codon arrangements in the genetic
411 code. We provide evidence that the standard genetic code, shared among most lifeforms, along
412 with organismal codon choices, is robust against the incorporation of additional atoms of nitrogen
413 and carbon given a random mutation. We quantitatively demonstrate that this nutrient
414 conservation is not confounded by known error minimization mechanisms of the genetic code and
415 it is of similar extent in terms of robustness to the effect of point mutations. Notably, we show that
416 the standard genetic code surpasses a great majority of randomly arranged codes in conservation

417 of nitrogen and carbon, across multiple taxa from all domains of life and across multiple codon
418 choices and mutation rates.

419 It should be noted that in addition to resource-driven selection there are likely additional selective
420 forces that contribute to the observed association of pN/pS (and of dN/dS) with the environment.
421 Genomic GC-content is associated with exposure to UV radiation[43] and nitrogen availability[4,44] yet
422 nonsynonymous mutations, especially in codons encoding structurally and functionally important
423 protein positions, are likely to be less affected by GC bias. In addition, codon usage bias has been
424 shown to be associated with the environment[45], and clearly we cannot rule out the association of
425 the environment with other selective forces operating on synonymous mutations that have not
426 been reported yet.

427 In light of these results, we hypothesize that resource driven selection is equally exerted on all
428 parts of the protein-coding gene. This sets it apart from selection to maintain the structural integrity
429 of a protein or the function of its active site, which occur predominantly in structurally important
430 regions[46]. Thus, accounting for resource-driven selection may improve the identification of
431 alternative translation start sites, alternatively spliced introns, or readthrough stop codons, as
432 intermittently translated regions of the protein may be under weaker resource-driven selection as
433 compared to constitutively translated ones[47].

434 We show that the genetic code optimizes nutrient conservation, calling upon an examination of
435 code evolution theories. Hypotheses regarding the origin of the genetic code include
436 stereochemical affinity between a codon or anticodon and their amino acid[48,49]; a frozen accident
437 theory[16], relying on the fact that the code is highly immutable; a co-evolution of the genetic code
438 with the emergence of amino acid-producing biosynthetic pathways[14]; and an early fixation of an
439 optimal genetic code, suggesting that the code evolved under selection for error minimization[50].
440 Our results are in line with the latter theory of optimality, and suggest that the genetic code may
441 also have been optimized for nutrient conservation. This implies that the genetic code can be

442 viewed as a buffer between the evolutionary forces of mutation and selection, the former occurring
443 in DNA sequences and the latter predominantly in proteins. Simply put, the genetic code
444 determines which mutations the selective forces 'see'. In the case of nutrient conservation, many
445 DNA mutations do not result in the incorporation of additional nutrients into proteins, and are thus
446 not selected against, which may allow more 'freedom' to explore the mutation-space.

447 We show that the codon choices of organisms across the tree of life, along with the structure of
448 the genetic code, confer robustness against mutations that add nitrogen to the protein sequence.
449 It is possible that amino acid and codon usage in organisms have evolved to lower nutrient
450 consumption in case of a random mutation, informed by the structure of the code. Interestingly,
451 this is also true for heterotrophic eukaryotes which are net N-producers. We hypothesize that
452 while an organism may be a net N-producer, it is possible that conservation of resources still
453 results in a selective advantage, in times of low nutrient availability and hunger. In multicellular
454 organisms, we hypothesize that resource conservation may be essential during development,
455 where rates of growth are conserved and a deviation from them may be selected against.

456 Since the genetic code is common, with only a few changes, to virtually all organisms across the
457 tree of life, we can assume that it was already fixed in its current form, or one very similar, in the
458 last universal common ancestor. If, as previously postulated[50], the genetic code evolved through
459 selection, it may reflect the selective pressures on organisms in the beginning of life on earth.
460 While we do not know the nature of nutrient cycling in the primordial ocean, we can hypothesize
461 that scarcity of nitrogen and carbon that are common now may have also prevailed alongside
462 early lifeforms. If so, an organism harboring a nitrogen- and carbon-efficient genetic code would
463 have had a selective advantage over its peers, especially in the absence of fully evolved DNA
464 mutation repair mechanisms.

465 Overall, by using publicly available data on ocean microbes and their corresponding
466 environmental measurements, we were able to discern a ubiquitous purifying selective pressure

467 which shapes marine microbial life, and may have even shaped the structure of the genetic code
468 that was since preserved for billions of years. With the advent of new multi-omic data from
469 environmental studies, we will be able to better divulge the intricate relationships of microbes with
470 a rapidly changing global environment.

471 **Methods**

472 **Marine microbiome samples**

473 Marine samples collected with Tara oceans[3], bioGEOTRACES[27], the Hawaii Ocean Timeseries
474 (HOT) and the Bermuda Atlantic Timeseries Series (BATS) [27] were downloaded from ENA with
475 accessions ENA:PRJEB1787 (TARA oceans prokaryotic fraction), ENA:PRJNA385854
476 (bioGEOTRACES) and ENA:PRJNA385855 (HOT/BATS), each sample with a minimum of 5
477 million reads. For all samples for which a size parameter was available, we selected the 0.22µm+
478 size category (typically between 0.22 and 5µm). This includes TARA ocean samples[3] (prokaryote
479 size) and data pertaining to expression[33]. In samples from the bioGEOTRACES[27], HOT[27] and
480 BATS[27] sampling stations, we used the available size fraction, larger than 0.2µm.

481 **Mapping of Illumina reads to reference gene sequences**

482 Samples were mapped to nucleotide sequences from the Ocean Microbiome Reference Gene
483 Catalog (OM-RGC) [3] using bowtie2 with parameters `--sensitive -a 20 --quiet -p 8` and saved as
484 a bam file using the 'samtools view' command. As gene sequences are relatively short, reads
485 from both ends of the metagenomic sequencing samples were mapped separately, and reunited
486 prior to variant calling.

487 **Determining metagenomic read assignment probability**

488 We determined the probability of assignment of metagenomic reads to marine microbial genes
489 using the Iterative Coverage-based Read-assignment Algorithm (ICRA)[51] with parameters
490 max_mismatch=12, consider_lengths=True, epsilon=1e-6, max_iterations=30, min_bins=4,
491 max_bins=100, min_reads=10, dense_region_coverage=60, length_minimum=300,
492 length_maximum=2e5, use_theta=False. To prevent spurious mapping, alignments were
493 considered for downstream analysis only if the probability of alignment was higher than 0.9.

494

495 **Variant calling**

496 Alignments from both ends of all sequencing runs pertaining to the same sample were united
497 using the samtools cat command and sorted using the samtools sort command, with default
498 parameters.

499 To facilitate variant calling in tractable timescales, each filtered, untied and sorted .bam file was
500 split into chunks, each encompassing 10,000 reference sequences (out of about 40 million
501 reference sequences). For each such batch of reference sequences, we called variants across
502 all samples using the following command: bcftools mpileup --threads 4 -a FORMAT/AD -Q 15 -L
503 1000 -d 100000 -m 2 -f <OM-RGC fasta> <bam filenames> | bcftools call --threads 4 -Ov -mv -o
504 <output vcf>, where <OM-RGC fasta> is the fasta file of OM-RGC nucleotide sequences, <bam
505 filenames> are the filenames of all .bam files pertaining to the reference sequence chunk in
506 question, and <output vcf> is the output .vcf file pertaining to that same chunk.

507 To ensure that sequencing errors do not skew our results, we considered single nucleotide
508 variants as SNPs for downstream analysis only if they had a GATK quality score of at least 30,
509 and an allele frequency of at least 1%[52]. While the GATK quality requirements filter out
510 sequencing errors randomly distributed across the genes, the 1% criterion eliminates random
511 sequencing errors that accumulate in the same position when depth of coverage reaches very
512 high numbers.

513 For a sample mapped to a reference gene to be considered for downstream analysis, we
514 demanded that at least 60% of SNPs called along the length of the reference gene for that sample
515 would be supported by at least 4 reads, thereby enabling accurate calculation of pN/pS ratios.
516 For a gene to be considered for downstream analysis, we demanded for that gene to have at least
517 one SNP common to 20 or more samples.

518

519 **Calculation of pN/pS in single genes**

520 While comparing SNP patterns across samples, it is instrumental to avoid biases due to
521 differences in coverage. We therefore downsampled the read coverage depth to the minimum
522 depth across all samples, for each position in a sample that was supported by more than 4 reads.
523 For positions that had minimum support of fewer than 4 reads, no subsampling was performed.
524 Subsampling was performed by drawing from a multinomial distribution, with *n* trials and variant
525 probabilities *p*, where *n* was set to the calculated minimum depth and *p* set to the relative
526 abundance of each variant in the given sample.

527 The ratio of non-synonymous and synonymous polymorphisms was calculated by considering all
528 called SNPs in every gene. First, we calculate a consensus sequence for each gene by taking,
529 for each SNP position, the variant that was overall more common across all samples (after the
530 subsampling performed above). We counted, for each gene, the number of non-synonymous and
531 synonymous sites across the consensus sequence. For each SNP position in each sample, we
532 counted the number of synonymous and nonsynonymous substitutions. As more than one variant
533 can exist in a single sample, we considered the relative abundance of synonymous to
534 nonsynonymous substitutions dictated by the different variants. For example, if the reference
535 codon was CAC, coding for histidine, one variant, a C-to-G transversion in the third position
536 abundant at 50%, led to a nonsynonymous mutation that resulted in glutamine (CAG) while
537 another variant in the same sample and the same position was a synonymous C-to-T transition,
538 we counted 0.5 synonymous substitutions and 0.5 nonsynonymous substitutions. We followed by
539 calculating the pN/pS ratio:

540 $$pN/pS = \frac{nonsynonymous\ substitutions}{nonsynonymous\ sites} / \frac{synonymous\ substitutions}{synonymous\ sites}$$

541

542

543 **Aggregation of calculated metrics using KEGG and eggNOG orthologies**

544 Functional assignments to KEGG KOs and eggNOG OGs for all OM-RGC genes were computed
545 using eggNOG-mapper v2 based on the eggNOG v5.0 database[53,54]. For each functional
546 assignment in each sample, all OM-RGC genes assigned with the same functional assignment
547 were concatenated and treated as one long genomic sequence per the calculation of pN/pS ratios.
548 To reduce noise in pN/pS calculation, we considered only KOs and OGs that had at least 5 genes
549 per sample, in at least 50 samples.

550

551 **Benchmarking data of assembled genomes**

552 Benchmarking data was employed in the form of uncultivated single cells of the dominant and
553 representative lineages of the surface ocean (SAR-11, SAR-86 and Prochlorococcus) from
554 different parts of the ocean. We used genomes varying at the strain level[31], with potentially
555 minimal batch effects, as those have been analyzed in a single laboratory.

556

557 **Calculation of dN/dS in single genes**

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

559

560 1. We extracted all gene sequences pertaining to each ortholog from all strains within each
561 clade. We determined orthologous genes using the annotation that was published
562 alongside the single-cell data. For each clade, we chose the annotation type that had the
563 highest coverage of genes, i.e., prokka[55] for Prochlorococcus, and Swissprot[56] for SAR11
564 and SAR86.
565 2. To account for the possibility that gene sequences may be misannotated resulting in the
566 grouping of more than one orthologous group under the same name, we performed

567 additional curation by clustering genes using Ward hierarchical clustering[57] on protein
568 sequences using normalized edit distance (Levenshtein) as a dissimilarity metric. We cut
569 off clusters that had dissimilarity lower than 30%, to ensure structural similarity.

570 3. For each of the resulting clusters that had more than 20 members (i.e., genes
571 corresponding to strains within the clade) we generated a multiple sequence alignment
572 (MSA) of protein sequences using MAFFT[58] , with the following command:
573 mafft --quiet --maxiterate 1000 --localpair <input_fasta> > <output_msa>
574 Following that, we computed a codon alignment using the generated MSA and nucleotide
575 sequences, with the biopython AlignIO package.

576 4. Using the MSA, we computed a phylogenetic tree of the sequences using raxml[59]
577 (raxmlHPC command line version) with model GTRGAMMA; rooted it using the -f I
578 modifier; and computed the ancestral sequence using the -f A modifier.

579 5. For each leaf node (strain ortholog) in the phylogenetic tree of gene sequences, we
580 calculated dN/dS with respect to the tree root, using the biopython cal_dn_ds function with
581 the "ML" model (Goldman and Yang 1994) and default parameters.

582

583 **Environmental variables**

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

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

592 bioGEOTRACES metadata was compiled from CTD sensor data and discrete sample data from
593 the GEOTRACES intermediate data product v.2[60]. HOT metadata was downloaded from the ftp
594 server of the University of Hawai'i at Manoa (ftp://ftp.soest.hawaii.edu/hot/) and BATS metadata
595 was downloaded from the Bermuda Institute of Ocean Sciences (http://bats.bios.edu/bats-data/).
596 As GEOTRACES/HOT/BATS ocean water samples are not linked to specific biological samples
597 as is the case with Tara oceans samples, we considered only water samples from the exact same
598 geographic location, within a day from biological sample collection time, and within 5% difference
599 in depth of collection, and chose the closest sample in terms of time and depth of collection. As
600 all these measurements of environmental conditions are highly correlated with each other (**Fig.**
601 **S2I**), we utilized this correlation structure to impute missing values using the EM algorithm [61].

602

603 **Linear mixed models**

604 We used a linear mixed model (LMM) with variance components, commonly used in population
605 genetics[30] (Methods), which allows to estimate the fraction of variance in pN/pS ratios (dependent
606 variable) that is explained by the environment (random effect) while controlling for the complex
607 correlation structure between the environmental parameters, in a data-driven manner.

608 *Generative model.* Consider a collection of $M_i, where\ i \in \{1,2\}$ features (i.e., $M_{1,2}$number of
609 KEGG and eggNOG orthologs respectively), each measured across N samples. We get as input
610 an $(M_i \times N)$ matrix O^i, where ${O_{nm}}^i$ is the pN/pS of ortholog m in sample n. Let ${y^i}_m =$
611 $({O_{m1}}^i, \ldots, {O_{mN}}^i)$ be a $N \times 1$ vector representing the pN/pS in ortholog m, according to grouping
612 $i,$ across N samples (e.g., pN/pS in KO K02274 across N samples). Let W be a $(N \times q)$
613 normalized matrix of environmental measurements. This included the depth of the sample, water
614 temperature and salinity, as well as concentration of the key molecules nitrate, nitrite, oxygen,
615 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 $u_m \sim N(0, \sigma^2{}_{u_m} I)$ 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

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

where $\sigma^2_K = M_i \sigma^2{}_{u_m}, K = \frac{1}{M_i} WW^T$. 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

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

Estimating effective population size

To approximate effective population size (N_e) we use fourfold degenerate synonymous diversity. Calculation was adapted from Schloissnig et al.[28]:

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

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 the number of times nucleotide B_j was seen in the fourfold degenerate position i, and c_i is the coverage of position i.

Controlling for the effects of time and effective population size

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

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

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 (as a proxy for effective population size) is the second fixed effect with β_{2_i} as its coefficient. $W$is a normalized matrix of environmental factors, u_m and ϵ_m are independent random variables distributed as $u_m \sim N(0\,, \sigma^2{}_{u^n{}_m} I\,)$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 $\sigma^2_{\epsilon_m}$.

LMM on samples from the epipelagic zone

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

Linear model for dN/dS rates

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

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

where β_{0_m} is the intercept, W is a normalized matrix of environmental factors and u_m is a vector of 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.

661 To account for the effect of time as potential confounder we constructed the following linear model
662 for dN/dS (the dependent variable) with the environmental factors and dS (as a proxy for time) as
663 independent variables.

664 $$dN/dS_m = \beta_{0_m} + \beta_{1_m} dS_m + Wu_m + \epsilon_{2_m}, \qquad (7)$$

665 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
666 is a normalized matrix of environmental factors and u_m is a vector of environmental coefficients
667 that were available with these samples (phosphate, depth, nitrate + nitrite). $\epsilon_{2_m} \sim N(0, \sigma^2{}_{\epsilon_{2_m}} I)$.
668 $m \in \{1, \ldots, M\}$, where M is the number of genes tested. In this model we only used depth,
669 NO_2+NO_3 and phosphate, as these were the environmental factors available for the benchmarking
670 data.
671 To account for potential underlying structure in terms of phylogeny, we applied phylogenetic
672 generalized least squares[62] (PGLS) regression in addition to ordinary least squares (OLS)
673 regression to model dN/dS. Phylogenetic regression models are used when there is an underlying
674 phylogenetic clustering structure of the observations. In this case samples may be lowly or highly
675 phylogenetic diverged leading to residuals that are not *i.i.d.*, which can affect the fit of the model.
676 On the other hand, the caveat of PGLS is an underestimation of the R^2 as compared to linear
677 models[63,64]. In our data PGLS and OLS essentially gave the same results and we therefore only
678 report the latter.
679

680 **KEGG KO expression data as a ranking metric**

681 Using expression data from 4,092 KEGG KOs collected by Kolody et al.[33], we ranked the KO
682 genes in our marine samples in the following way. We first represent the expression data in
683 relative abundance space (normalize each sample by its read counts). Next, for each focal KO i,
684 where $i \in \{1, \ldots, I\}$, we sum across its different instances within each sample. The input is an
685 $m \times n$ matrix, where m is the number of different instances of focal KO i, and n is the number of

686 samples. The output is a vector of length n: $(x_1, \ldots, x_n)$.Finally, we average the expression levels

687 $(x_1, \ldots, x_n)$ across samples:$\frac{1}{n}\sum_{j=1}^{n} x_j$ and rank the KOs based on the calculated average

688 expression. Notably, we limited the scope of our analysis to samples collected only in small

689 fraction filters (0.22 μm).

690

691 **Determination of extracellular genes**

692 To determine extracellular gene groups, we searched the eggNOG v.5 OG database for the words

693 'secreted' or both words 'extracellular' and 'protein' in their description. We demanded that the

694 words 'autoinducer', 'expression', 'role' and 'hypothetical' are not in the description to prevent

695 instances where (a) the OG in question describes a hypothetical protein; and (b) where the OG

696 produces a secreted particle but is not secreted by itself, as is the case with autoinducer producing

697 genes. To ensure robust pN/pS calculations, descriptions encompassing 10 OGs or more were

698 assigned a group name, while descriptions encompassing less than 10 OGs were all grouped

699 together in one group.

700

701 **Calculation of mutation flux in divergent nitrate concentrations**

702 To calculate the differential rate of nutrient incorporation between environments, we compare, for

703 each pair of codons *a* and *b*, the ratio of mutations that result in *a* becoming *b* and the ratio of

704 reverse mutations, that result in *b* becoming *a*. We term this quantity mutation flux. To prevent

705 simplex-related confounding effects, mutation flux is estimated using the log-odds ratio between

706 the two (e.g., $log\ (p(AAA \rightarrow AAC)/p(AAC \rightarrow AAA))$. We created a matrix $H^{(U)}$ as follows: Consider

707 a set U of all genes for which SNP measurements exist and a subset $T \subseteq U$ of this set across

708 Ksamples. Let $G^{(T)}{}_j = (V, E)$ be a codon graph for subset T and sample j, where $v \in V$ is a codon

709 (e.g., CUU coding for Alanine) and $(v, v') \in E$ if and only if v and v' are one mutation apart (e.g.,

CUU for Alanine and CAU for Histidine). Let ${w^{(T)}}_j : (v, v') \rightarrow [0,1]$ be a weight function where ${w^{(T)}}_j(v, v') = \frac{Number\ of\ mutations\ turning\ v\ to\ v\prime\ in\ subset\ T\ and\ sample\ j}{Number\ of\ occurrences\ of\ codon\ v\ in\ subset\ T\ and\ sample\ j}$. 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 H per each sample and compared the codon mutation frequencies between the 40 lowest- and 40 highest-nitrate samples. Despite significant differences in codon mutation frequencies between low-nitrate and high-nitrate samples, some of the difference could be driven by the simplex properties of the sum-normalized codon mutation frequencies, and some could be attributed to the different rates of synonymous mutations between the high- and low-nitrate groups, which, combined with simplex properties, may affect observed nonsynonymous mutation rates. To address simplex properties, we employed a centered log-ratio (CLR) normalization on *H*. The CLR transformation is a mapping, per codon composition, from the simplex to a Euclidean vector subspace. This log transforms each value and then centers them around zero as given below:

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

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

To address differences in rates of the different types of mutations, for each mutation (v, v') in the CLR normalized matrix $H'^{(U)}$ we calculated the log odds ratio between the mutation and its reverse mutation. 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.

Calculation of expected random mutation cost per genetic code

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 $c_e: V \times V \rightarrow Z$ be a cost function for element e, where:

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

When testing the random mutation cost on hydrophobicity:

$$c_{hyd}(v, v') = 0\ if\ v\ and\ v'\ are\ both\ hydrophilic\ or\ hydrophobic, 1\ otherwise\ .$$

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

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

and the ERMC cost as:

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

where

$$c^{+}{}_e = max(0, c_e).$$

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 wished to calculate mutation rates in sites that were under minimal selection. To this end, we estimated $P(mut(v, v'))$ 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*, *Synechococcus* and Human genomes using published transition:transversion rates[5,65].

3. We calculate *c* using information on the amino acids which each codon codes for.

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

759 **Confounding effects between cost functions for the structure of the genetic code**

760 To confirm that our elemental cost function is not confounded by traditional properties of amino
761 acids such as the polar requirement (PR) and hydropathy index[38–41], we calculated the expected
762 random mutation cost (ERMC), per genetic code, using these common cost functions across 1
763 million simulated alternative codes. To this end, we randomized the first and second position of
764 all codons, while maintaining that the two sets of first and second positions in which the stop
765 codons reside are separated by a single transition mutation. We next calculated a contingency
766 table for each pair of cost functions for both nitrogen and carbon (i.e., $ERMC_N(V)$:$ERMC_{PR}(V)$,
767 $ERMC_N(V)$:$ERMC_{hyd}(V)$, (V):$ERMC_{PR}(V)$, $ERMC_C(V)$:$ERMC_{hyd}(V)$). We assign each code to one
768 of four bins in the following way: (1) surpassing the standard genetic code in both cost functions
769 (e.g., nitrogen and PR), (2) surpassing the standard genetic code only in element e cost (e.g. only
770 nitrogen), (3) surpassing the standard genetic code only in the traditional cost function (e.g., PR),
771 (4) not surpassing the standard genetic code in neither. Finally, we applied the Chi-square test of
772 independence with two degrees of freedom to each contingency table.

773 Finally, we examined the subset of genetic codes (out of the 1M hypothetical codes) that have a
774 lower ERMC than the standard genetic code in terms of PR $(ERMC_{PR})$ or hydropathy $(ERMC_{hyd})$,
775 and tested whether this subset is also optimized for nitrogen and carbon. To this end, we created
776 a null model in which we compute EMRC only accounting for hydropathy or PR robustness and
777 an expanded, nested model that adds minimization of nitrogen and carbon utilization. Under the
778 null, we expect that in the subset of codes optimizing for hydropathy and PR, nitrogen and carbon
779 utilization will also be minimized. Under the alternative, if such nutrient optimization indeed exists

780 separately from structural optimization, we expect that nitrogen carbon and utilization will not be
781 optimized. Exact P-values are defined to be the fraction of shuffled codes with ERMC smaller or
782 equal to the standard genetic code (one for each element - carbon and nitrogen) in this subset of
783 codes that have a lower ERMC than the standard genetic code in terms of PR ($ERMC_{PR}$) or
784 hydropathy ($ERMC_{hyd}$).

785

786 **Determination of 'square' and 'diagonal' arrangements of the nitrogen genetic code**

787 We define a 'square' arrangement of the codons coding for nitrogen-rich amino acids histidine,
788 glutamine, asparagine, lysine and arginine as one where their codons span only two nucleotides
789 in the first position and two nucleotides in the second position. In the standard genetic code, these
790 amino acids are coded by CAN, CGN, AAN, AGR, following a square configuration. In contrast,
791 a 'diagonal' arrangement of the codons coding for these amino acids is one where they span all
792 possible nucleotides in the first position and all possible nucleotides in the second position. For
793 example, a genetic code where TTY codes for histidine, TTR for glutamine, CCN and AAR for
794 arginine, GGY for asparagine and GGR for lysine constitutes a 'diagonal' arrangement of nitrogen
795 amino acids. In each of the alternative genetic codes we generated, we tested whether either of
796 these conditions hold and, if so, designated the code as 'square' or 'diagonal' accordingly.

797

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

799 We downloaded *Prochlorococcus* and *Synechococcus* protein-coding gene sequences (where
800 available) from the Joint Genome Institute (https://genome.jgi.doe.gov/portal/) following
801 accession numbers published by Berube et al.[5]. To estimate codon relative abundance $P(v)$, we
802 counted and sum-normalized codons in all protein-coding genes for each species. To estimate

803 codon mutation rate $P(mut(v, v'))$, we used the published transition:transversion rate of 2:1 for
804 *Prochlorococcus* and *Synechococcus* [42].

805

806 **Multiple taxa ERMC calculation**

807 To calculate ERMC for 39 taxa across multiple transition:transversion rates, we downloaded
808 codon usage and GC-content data collected by Athey et al.[66]. We used codon usage counts to
809 estimate $P(v)$and 11 transition:transversion rates (1:5, 1:4, 1:3, 1:2, 2:3, 1:1, 3:2, 2:1, 3:1, 4:1,
810 5:1) to estimate $P(mut(v, v'))$.

811

References

1. Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 6506–6511 (2018).
2. del Giorgio, P. A. & Duarte, C. M. Respiration in the open ocean. *Nature* **420**, 379–384 (2002).
3. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
4. Mende, D. R. *et al.* Environmental drivers of a microbial genomic transition zone in the ocean's interior. *Nature Microbiology* vol. 2 1367–1373 (2017).
5. Berube, P. M., Rasmussen, A., Braakman, R., Stepanauskas, R. & Chisholm, S. W. Emergence of trait variability through the lens of nitrogen assimilation in Prochlorococcus. *Elife* **8**, (2019).
6. Druffel, E. R. M., Griffin, S., Coppola, A. I. & Walker, B. D. Radiocarbon in dissolved organic carbon of the Atlantic Ocean. *Geophys. Res. Lett.* **43**, 5279–5286 (2016).
7. Grzymski, J. J. & Dussaq, A. M. The significance of nitrogen cost minimization in proteomes of marine microorganisms. *ISME J.* **6**, 71–80 (2012).
8. Akashi, H. & Gojobori, T. 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).
9. Koonin, E. V. & Novozhilov, A. S. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* **61**, 99–111 (2009).
10. Hinegardner, R. T. & Engelberg, J. RATIONALE FOR A UNIVERSAL GENETIC CODE. *Science* **142**, 1083–1085 (1963).
11. Woese, C. R., Hinegardner, R. T. & Engelberg, J. Universality in the Genetic Code. *Science* vol. 144 1030–1031 (1964).

837 12. Woese, C. R. Order in the genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **54**, 71–75 (1965).
838 13. Gamow, G. Possible Relation between Deoxyribonucleic Acid and Protein Structures.
839 *Nature* **173**, 318–318 (1954).
840 14. Wong, J. T. A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **72**,
841 1909–1912 (1975).
842 15. Freeland, S. J. & Hurst, L. D. The genetic code is one in a million. *J. Mol. Evol.* **47**, 238–248
843 (1998).
844 16. Crick, F. H. The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379 (1968).
845 17. Massey, S. E. A neutral origin for error minimization in the genetic code. *J. Mol. Evol.* **67**,
846 510–516 (2008).
847 18. Massey, S. E. Genetic code evolution reveals the neutral emergence of mutational
848 robustness, and information as an evolutionary constraint. *Life* **5**, 1301–1332 (2015).
849 19. Massey, S. E. The neutral emergence of error minimized genetic codes superior to the
850 standard genetic code. *J. Theor. Biol.* **408**, 237–242 (2016).
851 20. Novozhilov, A. S. & Koonin, E. V. Exceptional error minimization in putative primordial
852 genetic codes. *Biol. Direct* **4**, 44 (2009).
853 21. Novozhilov, A. S., Wolf, Y. I. & Koonin, E. V. Evolution of the genetic code: partial
854 optimization of a random code for robustness to translation error in a rugged fitness
855 landscape. *Biol. Direct* **2**, 24 (2007).
856 22. Salinas, D. G., Gallardo, M. O. & Osorio, M. I. Local conditions for global stability in the
857 space of codons of the genetic code. *Biosystems.* **150**, 73–77 (2016).
858 23. Torabi, N., Goodarzi, H. & Shateri Najafabadi, H. The case for an error minimizing set of
859 coding amino acids. *J. Theor. Biol.* **244**, 737–744 (2007).
860 24. Zhu, W. & Freeland, S. The standard genetic code enhances adaptive evolution of proteins.
861 *J. Theor. Biol.* **239**, 63–70 (2006).
862 25. Woese, C. R., Dugre, D. H., Saxinger, W. C. & Dugre, S. A. The molecular basis for the

863 genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **55**, 966–974 (1966).
864 26. Haig, D. & Hurst, L. D. A quantitative measure of error minimization in the genetic code. *J.*
865 *Mol. Evol.* **33**, 412–417 (1991).
866 27. Biller, S. J. *et al.* Marine microbial metagenomes sampled across space and time. *Sci Data*
867 **5**, 180176 (2018).
868 28. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature*
869 **493**, 45–50 (2013).
870 29. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila.
871 *Nature* **351**, 652–654 (1991).
872 30. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population
873 stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).
874 31. Pachiadaki, M. G. *et al.* Charting the Complexity of the Marine Microbiome through Single-
875 Cell Genomics. *Cell* vol. 179 1623–1635.e11 (2019).
876 32. Koonin, E. V. Are There Laws of Genome Evolution? *PLoS Comput. Biol.* **7**, e1002173
877 (2011).
878 33. Kolody, B. C. *et al.* Diel transcriptional response of a California Current plankton
879 microbiome to light, low iron, and enduring viral infection. *ISME J.* (2019)
880 doi:10.1038/s41396-019-0472-2.
881 34. Anderson, T. R. & le B. Williams, P. J. A one-dimensional model of dissolved organic
882 carbon cycling in the water column incorporating combined biological-photochemical
883 decomposition. *Global Biogeochemical Cycles* vol. 13 337–349 (1999).
884 35. Milo, R. What is the total number of protein molecules per cell volume? A call to rethink
885 some published values. *Bioessays* **35**, 1050–1055 (2013).
886 36. Taniguchi, Y. Quantifying E. coli Proteome and Transcriptome with Single-molecule
887 Sensitivity in Single Cells. *Seibutsu Butsuri* vol. 51 136–137 (2011).
888 37. Woese, C. R. Evolution of the genetic code. *Naturwissenschaften* **60**, 447–459 (1973).

889 38. de Oliveira, L. L., de Oliveira, P. S. L. & Tinós, R. A multiobjective approach to the genetic
890 code adaptability problem. *BMC Bioinformatics* **16**, 52 (2015).
891 39. Di Giulio, M. & Medugno, M. The level and landscape of optimization in the origin of the
892 genetic code. *J. Mol. Evol.* **52**, 372–382 (2001).
893 40. Higgs, P. G. A four-column theory for the origin of the genetic code: tracing the evolutionary
894 pathways that gave rise to an optimized code. *Biol. Direct* **4**, 16 (2009).
895 41. Sengupta, S. & Higgs, P. G. Pathways of Genetic Code Evolution in Ancient and Modern
896 Organisms. *J. Mol. Evol.* **80**, 229–243 (2015).
897 42. Urbach, E., Scanlan, D. J., Distel, D. L., Waterbury, J. B. & Chisholm, S. W. Rapid
898 Diversification of Marine Picophytoplankton with Dissimilar Light-Harvesting Structures
899 Inferred from Sequences of Prochlorococcus and Synechococcus (Cyanobacteria). *Journal
900 of Molecular Evolution* vol. 46 188–201 (1998).
901 43. Singer, C. E. & Ames, B. N. Sunlight ultraviolet and bacterial DNA base ratios. *Science*
902 **170**, 822–825 (1970).
903 44. Mcewan, C. E. A., Gatherer, D. & Mcewan, N. R. Nitrogen-fixing aerobic bacteria have
904 higher genomic GC content than non-fixing species within the same genus. *Hereditas* **128**,
905 173–178 (1998).
906 45. Botzman, M. & Margalit, H. Variation in global codon usage bias among prokaryotic
907 organisms is associated with their lifestyles. *Genome Biol.* **12**, R109 (2011).
908 46. Liu, J., Zhang, Y., Lei, X. & Zhang, Z. Natural selection of protein structural and functional
909 properties: a SNP perspective. *Genome Biology* vol. 9 R69 (2008).
910 47. Read, R. W. *et al.* Nitrogen cost minimization is promoted by structural changes in the
911 transcriptome of N-deprived Prochlorococcus cells. *ISME J.* **11**, 2267–2278 (2017).
912 48. Hopfield, J. J. Origin of the genetic code: a testable hypothesis based on tRNA structure,
913 sequence, and kinetic proofreading. *Proc. Natl. Acad. Sci. U. S. A.* **75**, 4334–4338 (1978).
914 49. Yarus, M., Widmann, J. J. & Knight, R. RNA–Amino Acid Binding: A Stereochemical Era for

915 the Genetic Code. *J. Mol. Evol.* **69**, 406 (2009).
916 50. Freeland, S. J., Knight, R. D., Landweber, L. F. & Hurst, L. D. Early fixation of an optimal
917 genetic code. *Mol. Biol. Evol.* **17**, 511–518 (2000).
918 51. Zeevi, D. *et al.* Structural variation in the gut microbiome associates with host health.
919 *Nature* **568**, 43–48 (2019).
920 52. 1000 Genomes Project Consortium *et al.* A map of human genome variation from
921 population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
922 53. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology
923 Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
924 54. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically
925 annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids*
926 *Res.* **47**, D309–D314 (2019).
927 55. Seemann, T. Prokka: rapid prokaryotic genome annotation. Bio-Informatics 30: 2068--2069.
928 (2014).
929 56. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its
930 supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
931 57. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.*
932 **58**, 236–244 (1963).
933 58. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
934 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
935 59. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
936 phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
937 60. Schlitzer, R. *et al.* The GEOTRACES Intermediate Data Product 2017. *Chem. Geol.* **493**,
938 210–223 (2018).
939 61. Honaker, J., King, G., Blackwell, M. & Others. Amelia II: A program for missing data. *J.*
940 *Stat. Softw.* **45**, 1–47 (2011).

941 62. Symonds, M. R. E. & Blomberg, S. P. A Primer on Phylogenetic Generalised Least
942 Squares. *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary*
943 *Biology* 105–130 (2014) doi:10.1007/978-3-662-43550-2_5.
944 63. Westoby, M., Leishman, M. R. & Lord, J. M. On misinterpreting thephylogenetic correction'.
945 *J. Ecol.* **83**, 531–534 (1995).
946 64. Rheindt, F. E. & Grafe, T. U. Rapidly evolving traits and the comparative method: how
947 important is testing for phylogenetic signal? *Evol. Ecol.* (2004).
948 65. Wang, J., Raskin, L., Samuels, D. C., Shyr, Y. & Guo, Y. Genome measures used for
949 quality control are dependent on gene function and ancestry. *Bioinformatics* **31**, 318–323
950 (2015).
951 66. Athey, J. *et al.* A new and updated resource for codon usage tables. *BMC Bioinformatics*
952 **18**, 391 (2017).
953 67. Charlesworth, B. Effective population size and patterns of molecular evolution and
954 variation. *Nat. Rev. Genet.* **10**, 195–205 (2009).
955 68. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes.
956 *Science* vol. 290 1151–1155 (2000).
957 69. Hughes, A. L., Friedman, R., Rivailler, P. & French, J. O. Synonymous and
958 nonsynonymous polymorphisms versus divergences in bacterial genomes. *Mol. Biol. Evol.*
959 **25**, 2199–2209 (2008).
960 70. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids*
961 *Res.* **28**, 27–30 (2000).

962

963

964 **Supplementary Information**

965 **Supplementary Note 1. The association between the environment and pN/pS is robust to**
966 **time and effective population size.**

967 The magnitude of purifying selection, approximated by pN/pS ratios, may increase as a function
968 of both time and effective population size[67,68]. These two mechanisms might also be associated
969 with environmental conditions and we thus sought to control for them as potential confounders.
970 Time may affect the magnitude of pN/pS ratios via random genetic drift. Effective population size
971 may affect selection via competition, as in large populations, any slightly deleterious mutation is
972 rapidly selected against. We used the ratio of synonymous substitutions (i.e., pS) calculated from
973 our data as an approximation of dS, representing neutral polymorphism rate, as a proxy for time[69].
974 To approximate effective population size, we relied on population genetic theory suggesting that
975 effective population size is scaled to diversity at the silent sites when assuming constant mutation
976 rate between populations. As our data is composed of closely related sequences and hence
977 similar mutation rate, we used fourfold-degenerate synonymous diversity (SD) as a proxy for the
978 effective population size[28] (Methods).

979 We first tested whether SD is associated with environmental conditions, and compared the SD of
980 populations from different oceanic regions with different environmental parameters. We found that
981 pN/pS was more strongly associated with most environmental parameters than SD (Kolmogorov-
982 Smirnov test $P<10^{-20}$ for depth, nitrate, oxygen, phosphate, silicate and temperature ; $P<10^{-5}$ for
983 NO_2+NO_3 concentrations; non-significant for salinity; **Fig. S3**) suggesting that selection on protein
984 sequences may be more strongly affected by the environment than effective population sizes are.
985 We further found that in KOs with the highest EEV these differences are the most pronounced
986 (Wilcoxon signed-rank test $P<10^{-15}$; **Fig. S15**).

987 As LMMs also allow controlling for potential confounders as fixed effects, we next added our
988 approximations of time and the effective population size to our LMM in order to account for their

989 effect on the magnitude of pN/pS ratios. Accounting for these potential confounders as fixed
990 effects would enable a a more accurate quantification of the environmental effect. Similar to the
991 results of our original LMM, across both KEGG and eggNOG orthologs, we found that a significant
992 fraction of the variance in pN/pS can be attributed to the environment, even after controlling for
993 these confounders (**Fig. S4C,D**; Mann-Whitney *U* test $P<10^{-16}$) and a high correlation between
994 EEV estimated by the two models (R^2=0.947 and 0.929 for KEGG and eggNOG, respectively;
995 **Fig. S16**).

996

997 **Supplementary Note 2. The association between the environment and pN/pS is not**
998 **confounded by depth.**

999 To demonstrate that our results are not confounded by depth, the environmental variable with the
1000 second strongest correlation to pN/pS (after nitrate), we recalculated our LMM only on samples
1001 from the epipelagic zone (Methods). Similar to the original results of the LMM, we show significant
1002 variance explained in pN/pS in this environmental niche (**Fig. S5**; Mann-Whitney *U* test $P<10^{-20}$).

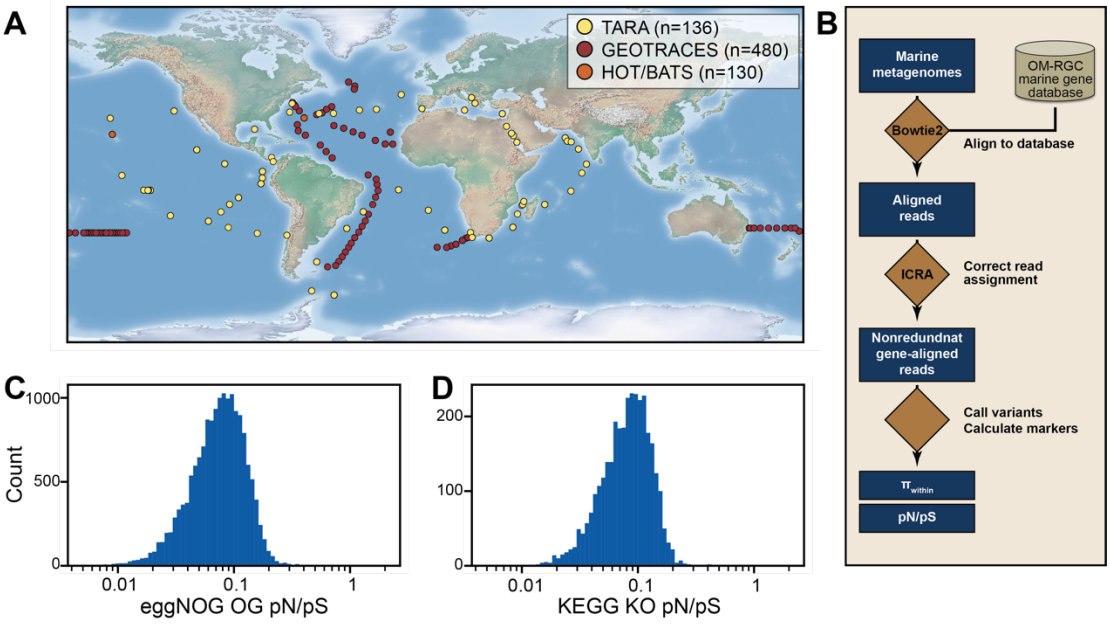
1003

1004 **Supplementary Note 3. The association between the environment and dN/dS in specific**
1005 **lineages is robust to time.**

1006 To account for the same potential confounders as in the original LMM model model, we added
1007 the rate of synonymous substitutions (dS) to the model, as a measure of time (Methods). Similar
1008 to the results of our original model, we found that a significant portion of the variance in dN/dS
1009 ratios is explained by the environment, even after controlling for time as a confounder, across the
1010 three bacterioplankton lineages tested (SAR-11, SAR-86 and Prochlorococcus; Wilcoxon signed-
1011 rank test $P<10^{-16}$; **Fig. S6**). These results validate our findings using pN/pS calculated from
1012 metagenomic sequences and demonstrate the association between selective pressure and

environmental conditions, even in a single clade. This indicates that associations of pN/pS or dN/dS with environmental variables are not confounded by organismal properties which may differ in tropism across environments, but rather exhibit a consistent trend even within a single clade.

1017 **Supplementary Figures**



1018

1019 **Figure S1.** (A) Geographical overview of the samples used in this study. (B) Illustration of our
1020 computational pipeline. (C,D) Histogram of pN/pS rates for eggNOG orthologous groups (OG[53]; C)
1021 and KEGG orthologs (KO[70]; D) across all marine samples.

1022

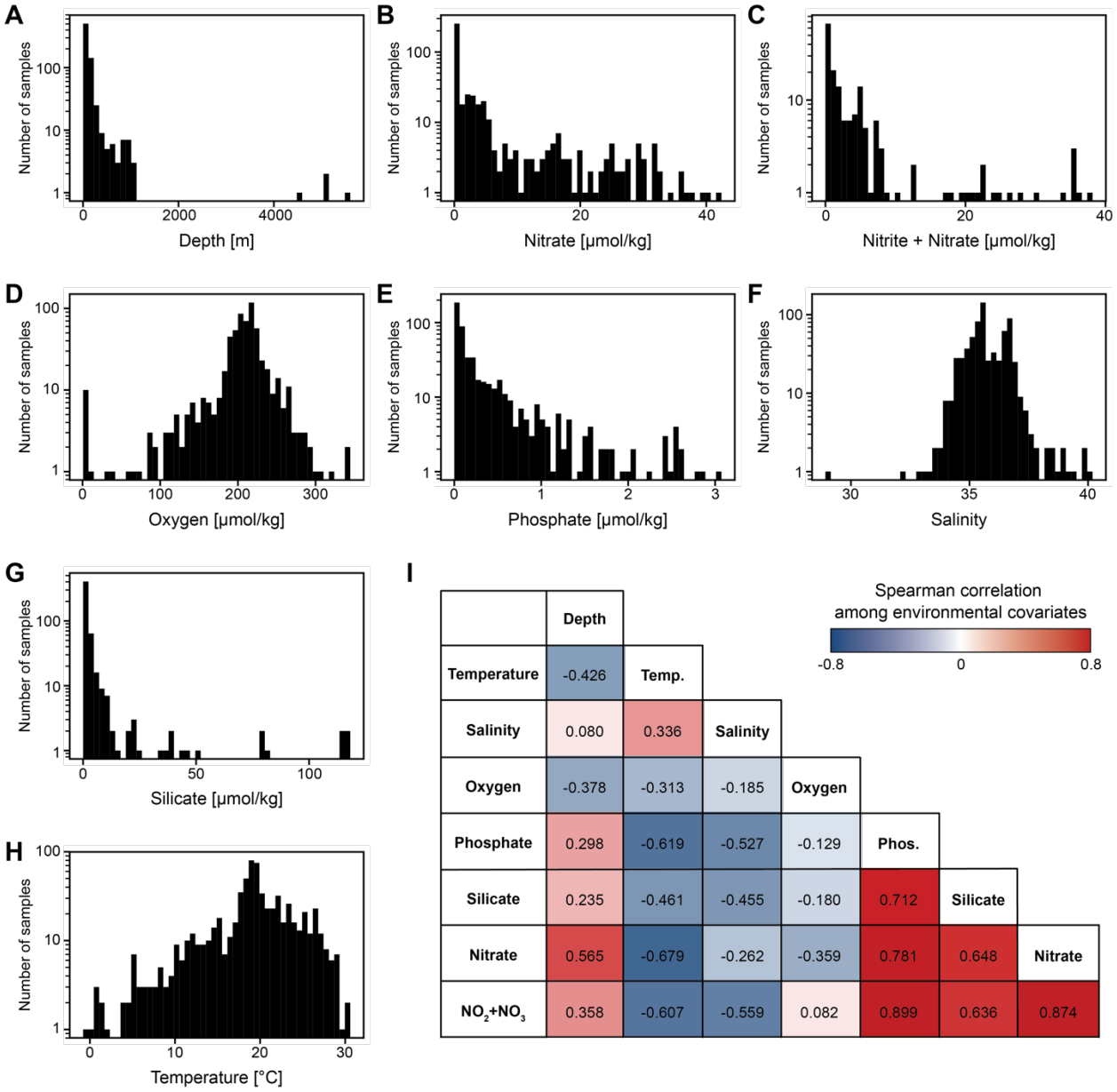


Figure S2. (A-H) Distribution of measurements taken alongside marine microbial samples for depth (A), nitrate (B), nitrate and nitrite (C), oxygen (D), phosphate (E), salinity (F), silicate (G) and temperature (H). (I) Spearman correlation coefficients between all pairs of environmental measurements across all available samples.

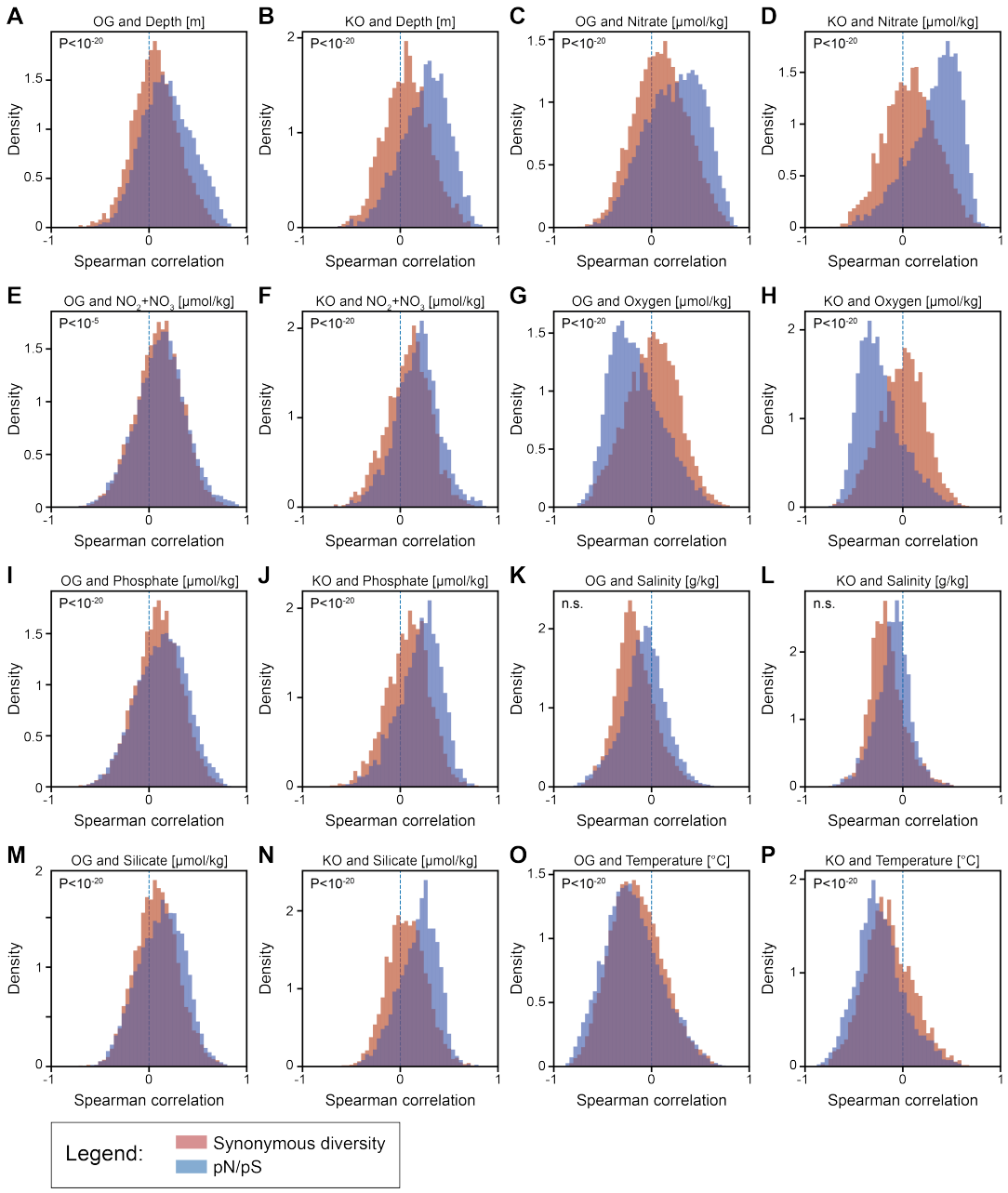


Figure S3. (A-P) Histograms of Spearman correlations between synonymous diversity (red) or pN/pS (blue) and environmental variables, for both KEGG KOs and eggNOG OGs. Panels A, C, E, G, I, K, M and O depict correlations between OG calculated parameters and depth, nitrate, nitrite and nitrite, oxygen, 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.

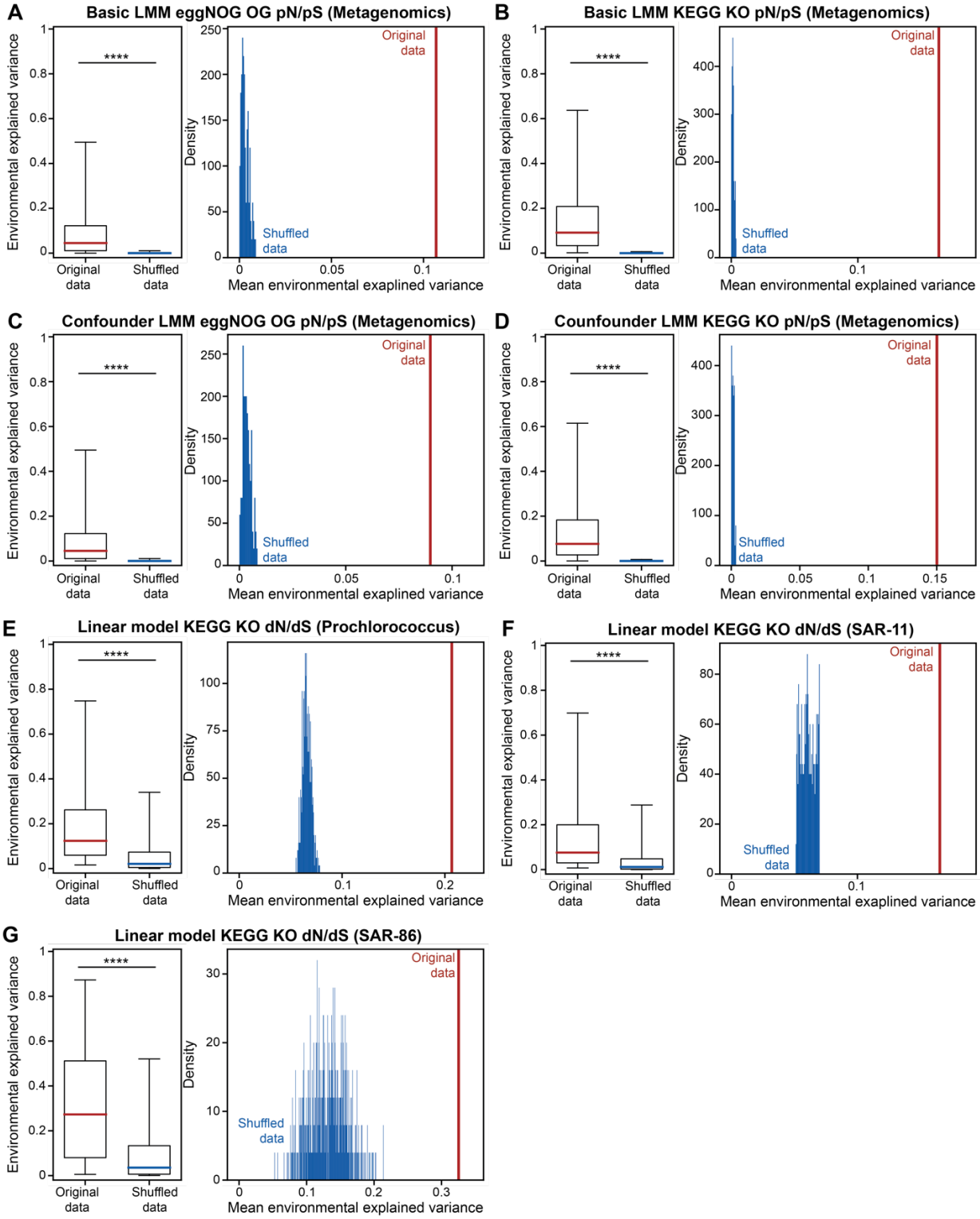
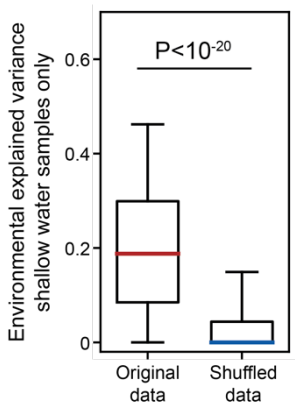


Figure S4. (A) Left, box plot (line, median; box, IQR; whiskers, 5th and 95th percentiles) of the variance of eggNOG OG pN/pS rates that was explained by the environment in a basic linear mixed model (LMM) setting (Methods) as compared to the same data with shuffled labels; Right, mean

1039 variance explained in unshuffled data (red) as compared to a histogram (blue) of mean variance
1040 explained in 500 executions with shuffled data. (B) Same as A, for KEGG KO pN/pS. (C) Same as A,
1041 for variance in eggNOG OG pN/pS in a LMM corrected for confounding factors of time and effective
1042 population size. (D) Same as C, for KEGG KO pN/pS. (E) Box plot (line, median; box, IQR; whiskers,
1043 5th and 95th percentiles) of the variance of KEGG KO dN/dS in genus *Prochlorococcus* that was
1044 explained by the environment in a linear model. (F) Same as E, for the SAR-11 clade. (G) Same as E,
1045 for the SAR-86 clade. ****, Wilcoxon signed-rank test $P<10^{-16}$.

1046



1047

1048 **Figure S5.** Box plot (line, median; box, IQR; whiskers, 5th and 95th percentiles) of the variance
1049 of KEGG KO pN/pS ratios that was explained by the environment in epipelagic samples (up to
1050 76m) as compared to the same data with shuffled labels.

1051

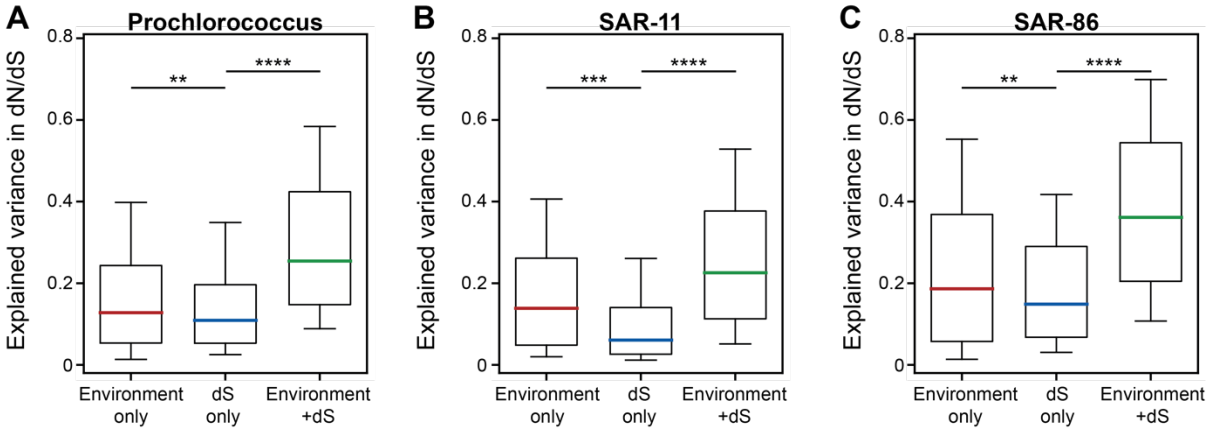


Figure S6. Box plots (line, median; box, IQR; whiskers, 10th and 90th percentiles) of explained variance in dN/dS using linear models considering the environment only (left, red), dS only (middle, blue), or both (right, green). (A) Prochlorococcus, (B) SAR-11 clade and (C) SAR-86 clade. **, Wilcoxon signed-rank test $P<0.01$; ***, $P<10^{-5}$; ****, $P<10^{-16}$.

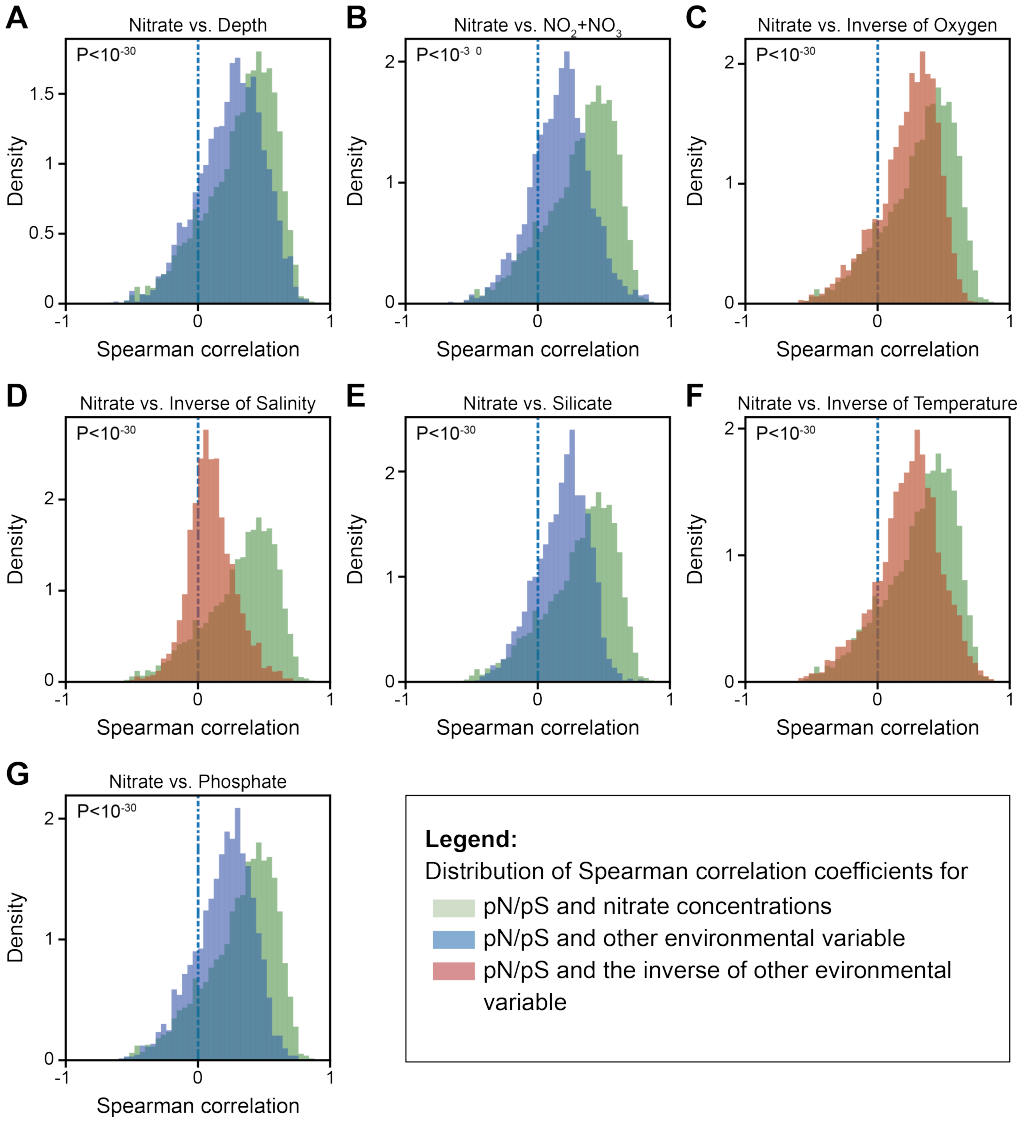


Figure S7. Histograms of Spearman correlations between environmental variables (blue) or their inverses (red) and pN/pS ratios of KEGG KOs as compared to a Spearman correlation between the concentration of nitrate and pN/pS ratios (green). P, one sided Kolmogorov-Smirnov test.

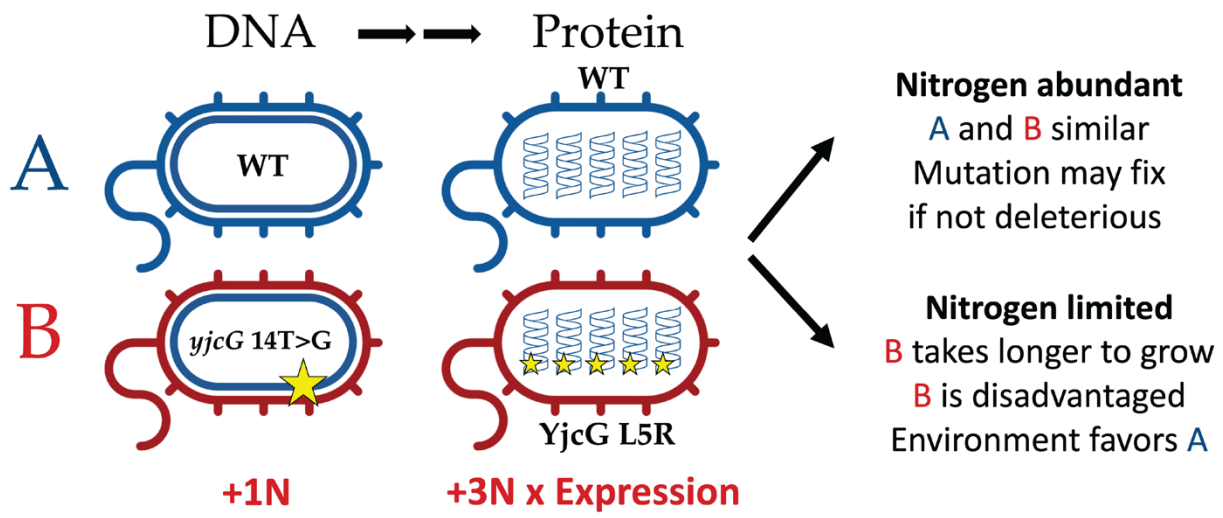


Figure S8. Illustration of the potential environmental effect of nitrate on protein-coding sequences.

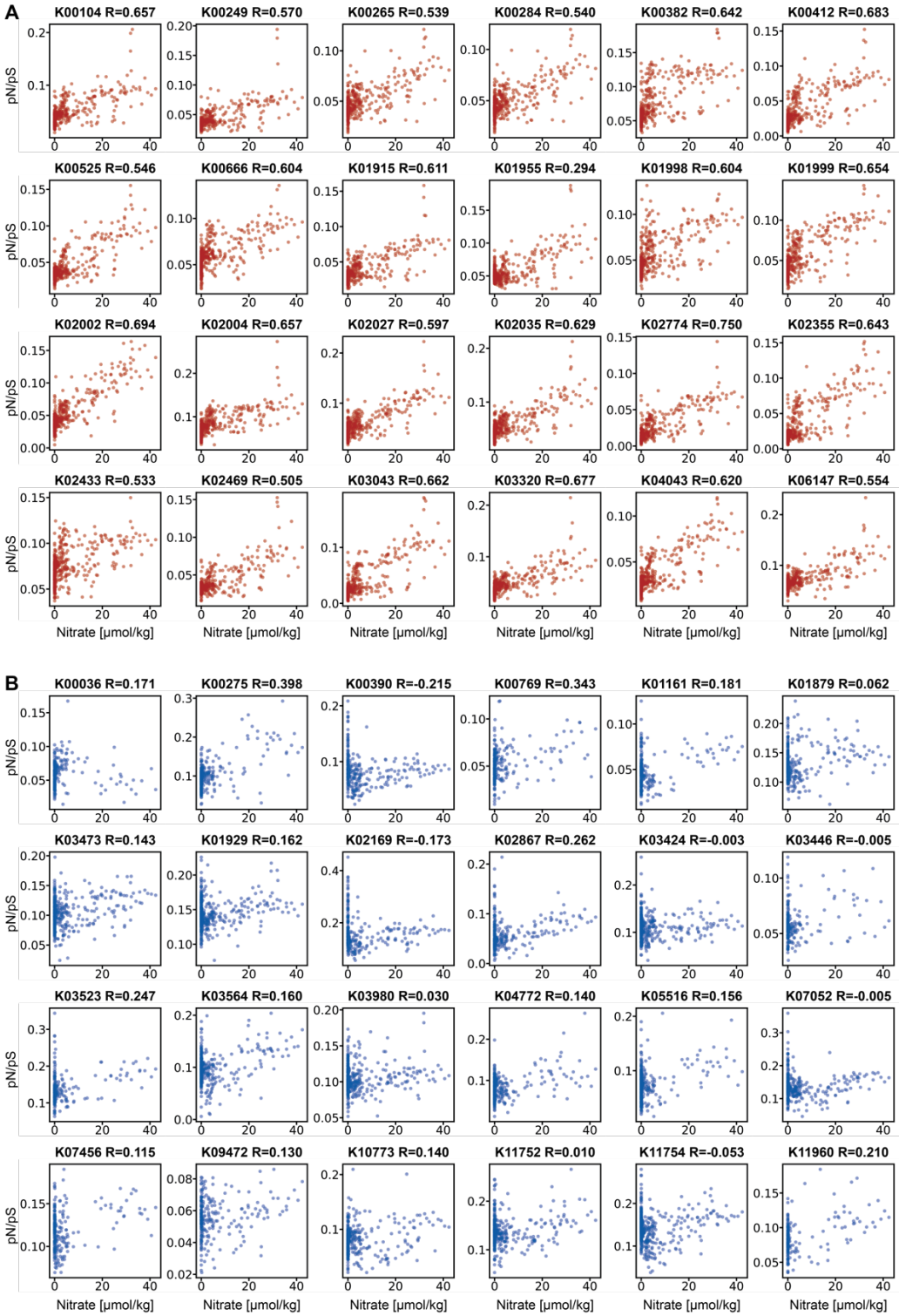
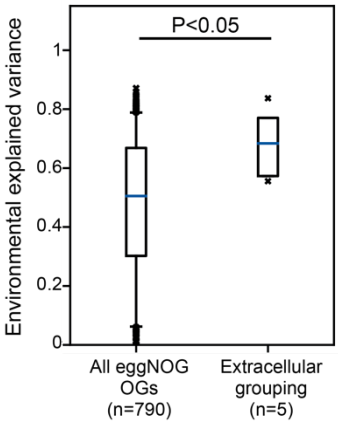


Figure S9. (top, red) 24 highly expressed KOs, exhibiting high correlation between pN/pS and nitrate concentrations. (bottom, blue) 24 lowly expressed KOs, exhibiting low correlation between pN/pS and nitrate concentrations. R, Spearman rho.



1070

1071 **Figure S10.** Box plot (line, median; box, IQR; whiskers, 5th and 95th percentiles) of variance in
1072 pN/pS explained by the environment in extracellular gene groups versus all eggNOG OGs.

1073

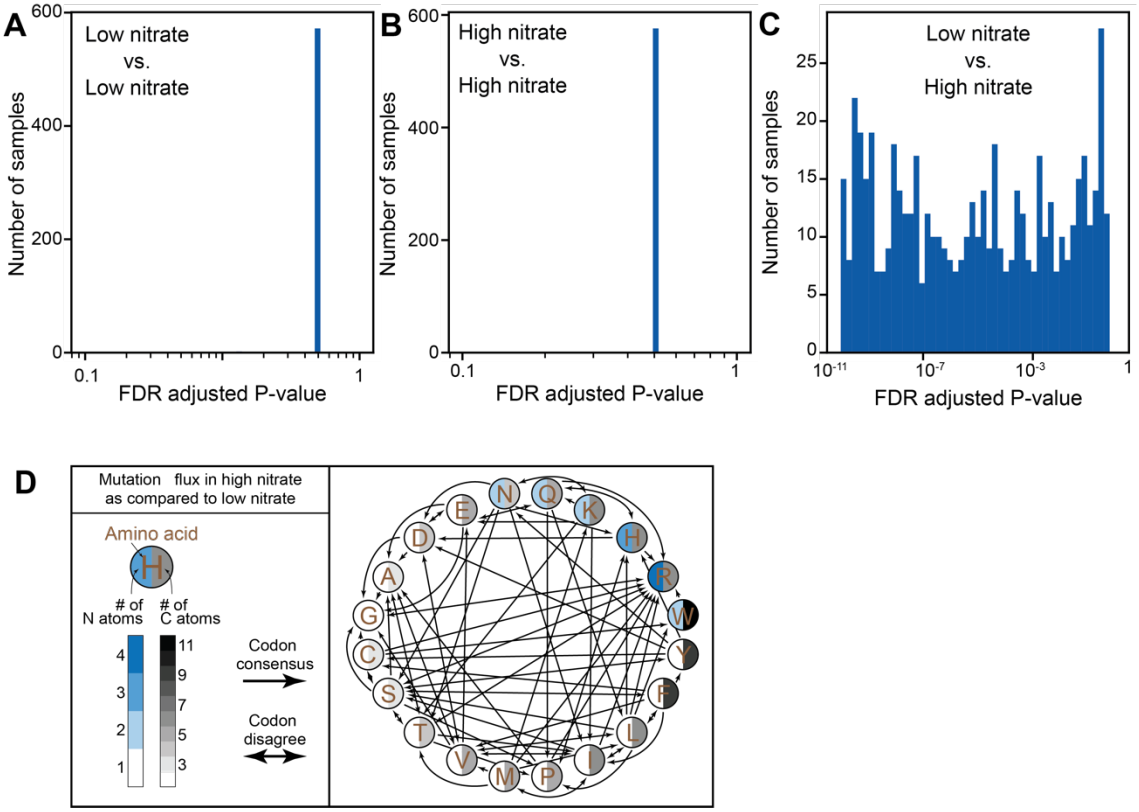


Figure S11. (A-C) Histograms of the distribution of Mann-Whitney *U* P-values of codon-to-codon mutations compared between (A) 40 low-nitrate samples and 40 other low-nitrate samples selected randomly out of the 80 lowest-nitrate samples; (B) 40 high-nitrate samples and 40 other high-nitrate samples 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 nitrogen and carbon content. Two-headed arrows mark cases where different codons of the same amino acid have opposite mutation fluxes.

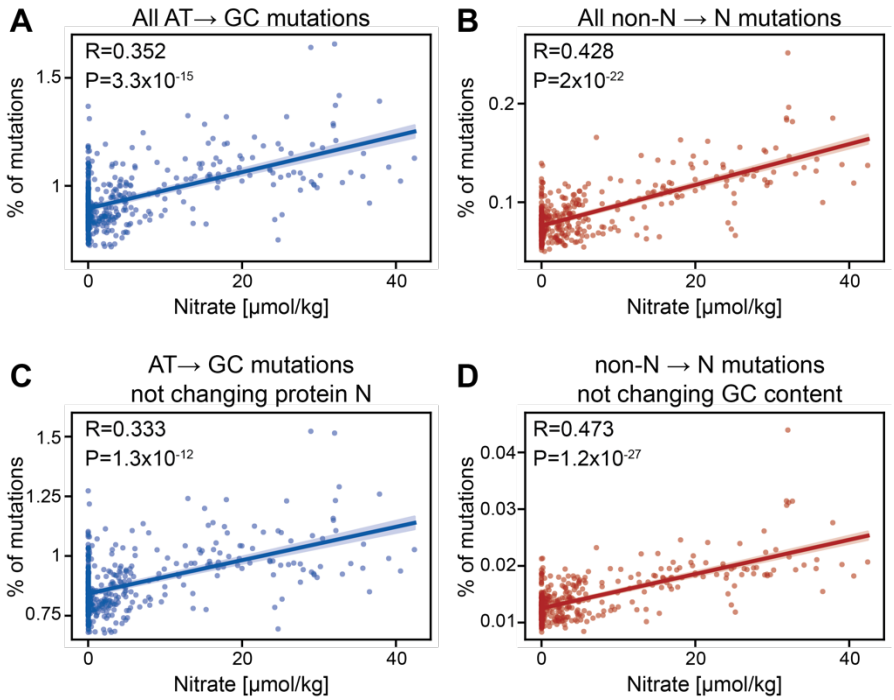


Figure S12. Scatter plots of nitrate concentrations (x-axis) versus percent of mutations (y-axis) for (A) all AT to GC mutations; (B) All mutations in coding sequences changing an amino acid with no nitrogen in its side-chain to one with nitrogen in the side-chain (non-N to N mutations); (C) AT to GC 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 for fit.

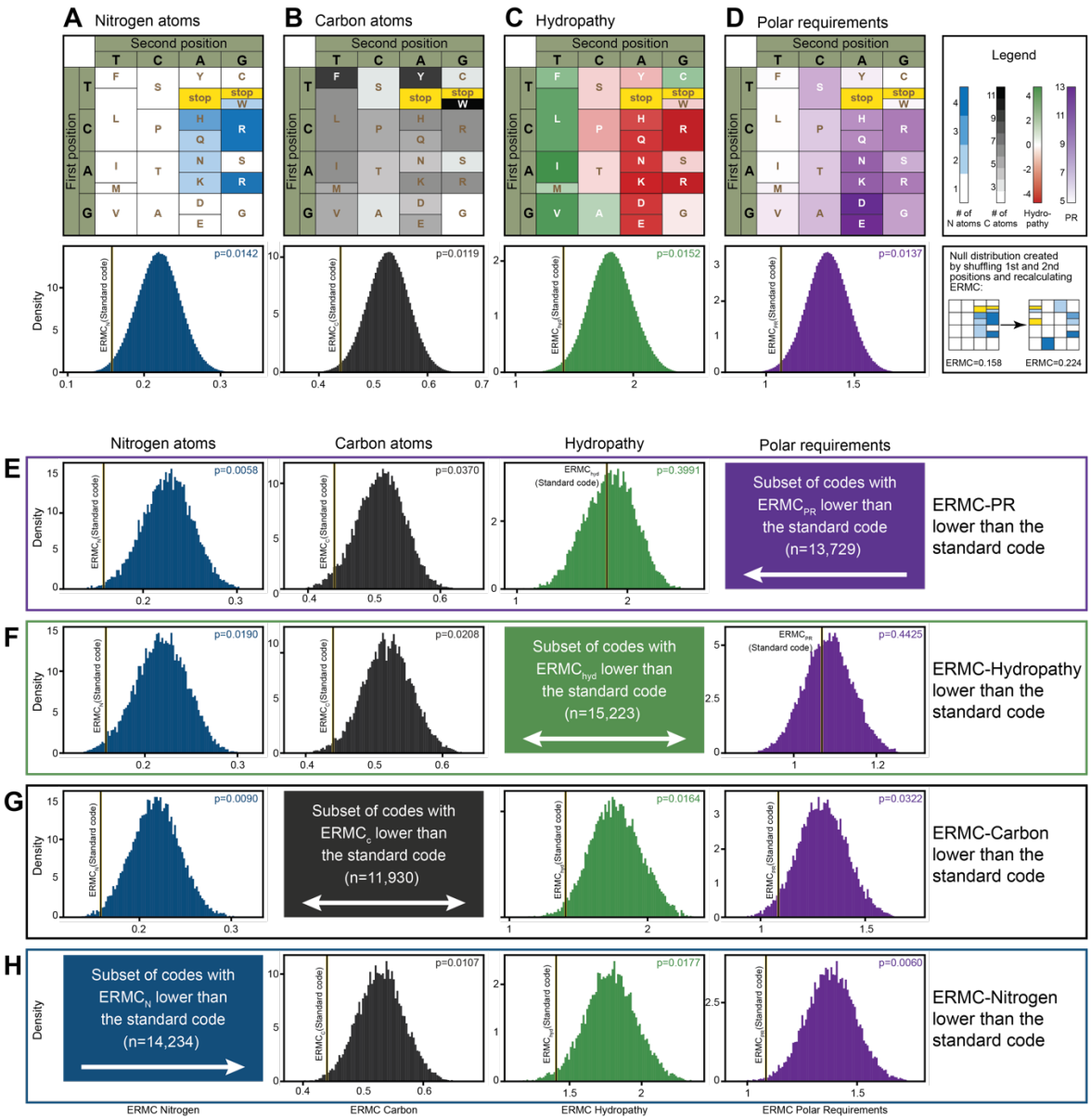
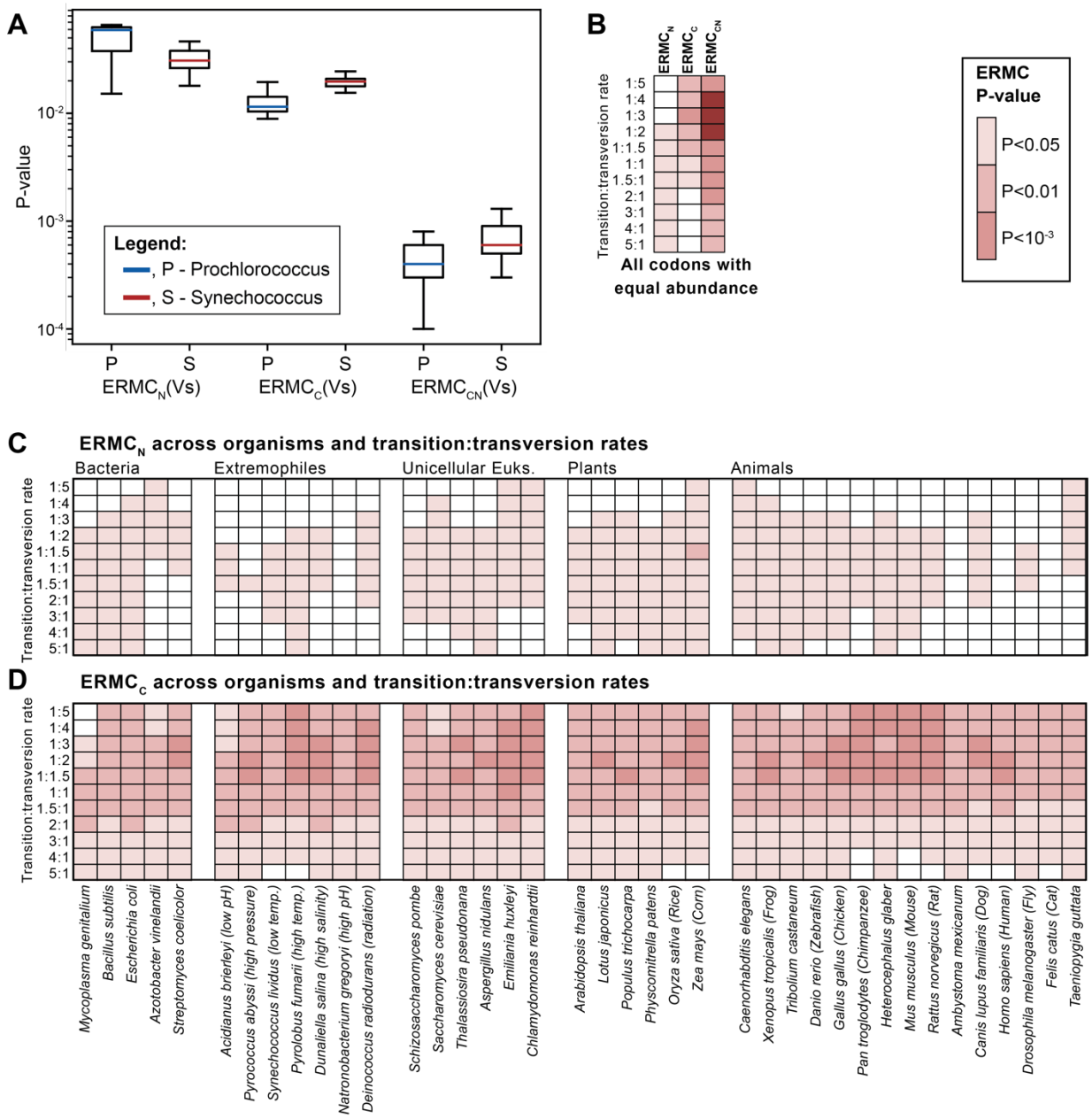


Figure S13. Upper panel: (Top) Nitrogen content (A; blue color scale), carbon content (B; black), hydropathy (C; green-red) and polar requirement (D; purple) of different amino acids depicted along their positions in the standard genetic code. (Bottom) Histograms of the expected random mutation cost (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) the polar requirement, (F) hydropathy, (G) carbon and (H) nitrogen. Vertical black-yellow line, ERMC of the standard genetic code for each property. P, permutation test.

1101



1102

1103 **Figure S14.** (A) Box plots (line, median; box, IQR; whiskers, 5th and 95th percentiles) of P-values
1104 for the ERMC of the standard genetic code for nitrogen (left), carbon (center) and both (right)
1105 across 187 Prochlorococcus (P, blue) and Synechococcus (S, red) strains. (B) Heat map of ERMC
1106 P-values for nitrogen, carbon and both, for a theoretical case in which all codons are of the same
1107 abundance. (C,D) Same as Fig. 3C for $ERMC_N$ (D) and $ERMC_C$ (E) P-values.

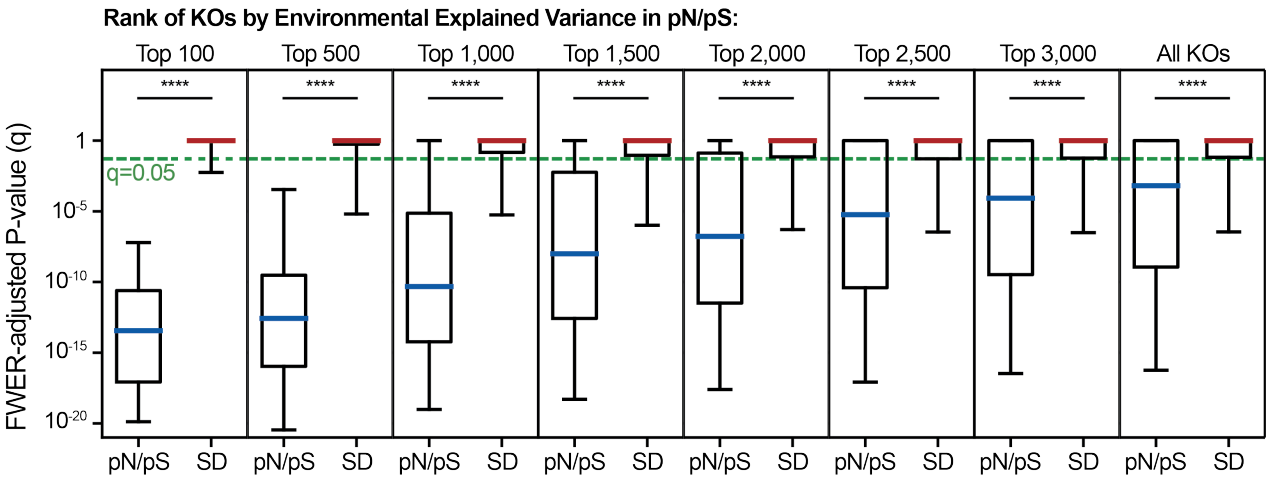
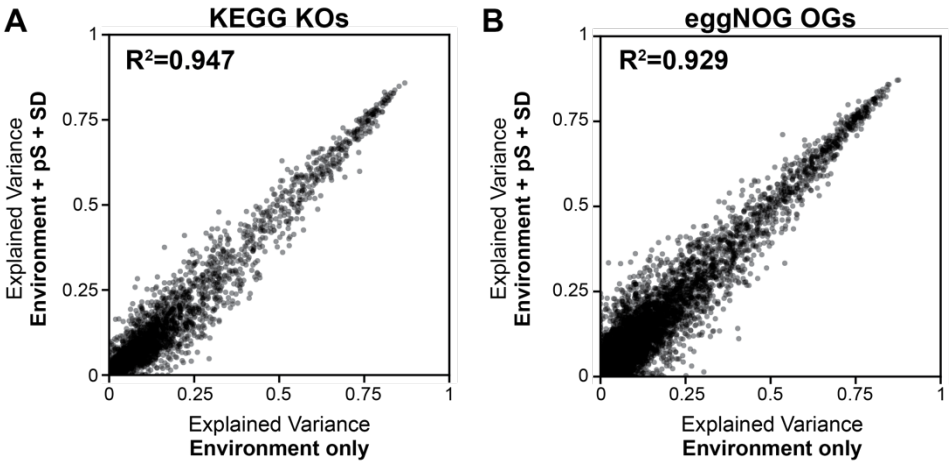


Figure S15. Box plots (line, median; box, IQR; whiskers, 10th and 90th percentiles) of P-values (t-test, multiple hypothesis adjustment) between samples with high-N (more than 19.6 μmol/kg) and low-N (0 μmol/kg), for KEGG KOs ordered by their EEV in the LMM, for both pN/pS (blue) and synonymous diversity (red). ****, Wilcoxon signed-rank test $P<10^{-15}$.



1114

1115 **Figure S16.** (A,B) scatter plot of explained variance in a linear mixed model considering only
1116 environmental factors as random effects (x-axis) versus a model considering environmental
1117 factors as random effects plus dS (as a proxy for time) and synonymous diversity (as a proxy for
1118 effective population size) as fixed effects (y-axis) in all (A) KEGG KOs and (B) eggNOG OGs.

1119

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	

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.