# 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

Ocean microbes are responsible for about 50% of primary production on Earth, and are strongly affected by environmental resource availability. However, selective forces resulting from environmental conditions are not well understood. We studied selection by examining single-nucleotide variants in the marine environment, and discovered strong purifying selective forces exerted across marine microbial genes. We present evidence indicating that this selection is driven by the environment, and especially by nitrogen availability. We further corroborate that nutrient availability drives this 'resource-driven' selection by showing stronger selection on highly expressed and extracellular genes, that are more resource-consuming. Finally, we show that the standard genetic code, along with amino acid abundances, facilitates nutrient conservation by providing robustness to mutations that increase nitrogen and carbon consumption. Notably, this robustness generalizes to multiple taxa across all domains of life, including the Human genome, and manifests in the code structure itself. Overall, we uncover overwhelmingly strong purifying selective pressure across marine microbial life that may have contributed to the structure of our genetic code.

## Introduction

Ocean microbes, the largest group of organisms on the planet[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 marine 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, and despite their importance in global energy flux and nutrient cycling, 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 lower incorporation of nitrogen-rich side chains into proteins, a strong A+T bias in nucleotide sequences, and smaller genome sizes, suggesting that nitrogen conservation is a strong selective force[7]. An opposite trend was shown for carbon[4]. Previous studies[7,8] identified a purifying selective pressure associated with resource conservation, which we term ‘resource-driven’ selection. Such ‘resource-driven’ selection against incorporation of nutrients in a resource-limited environment may be further propagated by the high effective population sizes observed in the open ocean, where even slightly deleterious mutations are rapidly selected against[9]. Notwithstanding, the ways in which resource-driven selection manifests in protein-coding sequences are not fully elucidated.

48 To illuminate mechanisms through which resource-driven selection affects protein-coding genes,
49 we amalgamated measurements of environmental conditions with publicly available marine
50 metagenomic data from oceanic habitats across the globe (Fig. 1A)[3,10]. We analyzed purifying
51 selection in 746 such marine samples by devising a tailored computational pipeline examining
52 single nucleotide polymorphisms. This enabled us to systematically associate purifying selection
53 with related environmental measurements. We revealed a strong purifying selective pressure,
54 which seems to be acting in a similar fashion across most marine microbial genes. This purifying
55 selection is associated with environmental nutrient concentrations, specifically nitrate. We further
56 show that resource-consuming genes, which are highly expressed or code for extracellular
57 proteins, are under stronger resource-driven selection as compared to other, less resource-
58 consuming genes. We analyze mutations in nitrate-rich as compared to nitrate-poor waters and
59 show that this selection is likely characterized by specific amino acid preferences depending on
60 environmental conditions. Finally, we demonstrate that the distribution of amino acids, along with
61 the structure of the genetic code, provides robustness against random mutations that increase
62 carbon and nitrogen incorporation into protein sequences. We extend this observation to codon
63 distributions across many diverse life forms, and suggest that nutrient conservation is encoded in
64 the standard genetic code, which is robust to mutations that result in higher nitrogen and carbon
65 utilization.

66 **Results**

67

68 **Single nucleotide polymorphisms in marine microbial genes reveal strong purifying**
69 **selection**

70 To better understand the underlying mechanisms governing resource-driven selection acting on
71 marine microbes, we sought to characterize, at the single nucleotide level, how coding sequences
72 of marine microbes are affected by resource availability in their environment. To this end, we
73 devised a computational pipeline that calculates metrics of selection from marine metagenomic
74 samples (Fig. 1B). We downloaded metagenomic sequencing data from 746 samples from the
75 Tara oceans[3] (n=136); bioGEOTRACES[10] (n=480); Hawaii Ocean Time-series[10] (HOT; n=68);
76 and Bermuda Atlantic TimeSeries[10] (BATS; n=62) expeditions (Fig. 1A; Methods). We aligned
77 these reads to the Ocean Microbiome Reference Gene Catalog (OM-RGC)[3] and searched for
78 single nucleotide polymorphisms (SNPs) in genes that had sufficient high-quality coverage (Fig.
79 1B; Methods). Overall, we found 71,921,864 high-confidence SNPs in 1,590,843 genes.

80 Next, to quantify purifying selection on different gene functions, we annotated genes from the OM-
81 RGC database to inform their functional group membership with either KEGG orthology (KO) or
82 eggNOG orthologous group (OG; Methods). We then calculated, using called SNPs per
83 orthologous group in each of the samples, the ratio of non-synonymous to synonymous
84 polymorphisms[11,12] (pN/pS; Methods). We used pN/pS rated to approximate purifying selection at
85 the population level as dN/dS ratios, which are typically used to characterize these stresses[12],
86 were not applicable in this setting (Methods). pN/pS quantifies the rate of nonsynonymous
87 polymorphisms (pN), which lead to a change in the resulting amino acid, normalized to the rate
88 of synonymous polymorphisms (pS), which maintain the coded amino acid.

As pN/pS is a proxy for the magnitude of purifying selection exerted on protein-coding sequences, we sought to utilize it to evaluate the selective forces acting on marine microbial functions. The rate of non-synonymous to synonymous polymorphisms was on average, across samples, 0.074 (CI [0.072, 0.075]) across all OGs and 0.079 (CI [0.077, 0.080]) across all KOs, similar to previously reported pN/pS ratios across different microbial genomes in the human microbiome[11] (Fig. 1C,D). With values close to zero indicating very strong selection against amino acid changes, these findings imply that purifying selection is on the same scale in free-living marine organisms as compared to host-associated microbes in the human gut. While gut microbes are expected to be under strong purifying selection in order to keep functioning in the host-associated niche[13], the source of this strong purifying selection on ocean microbes is not well understood.

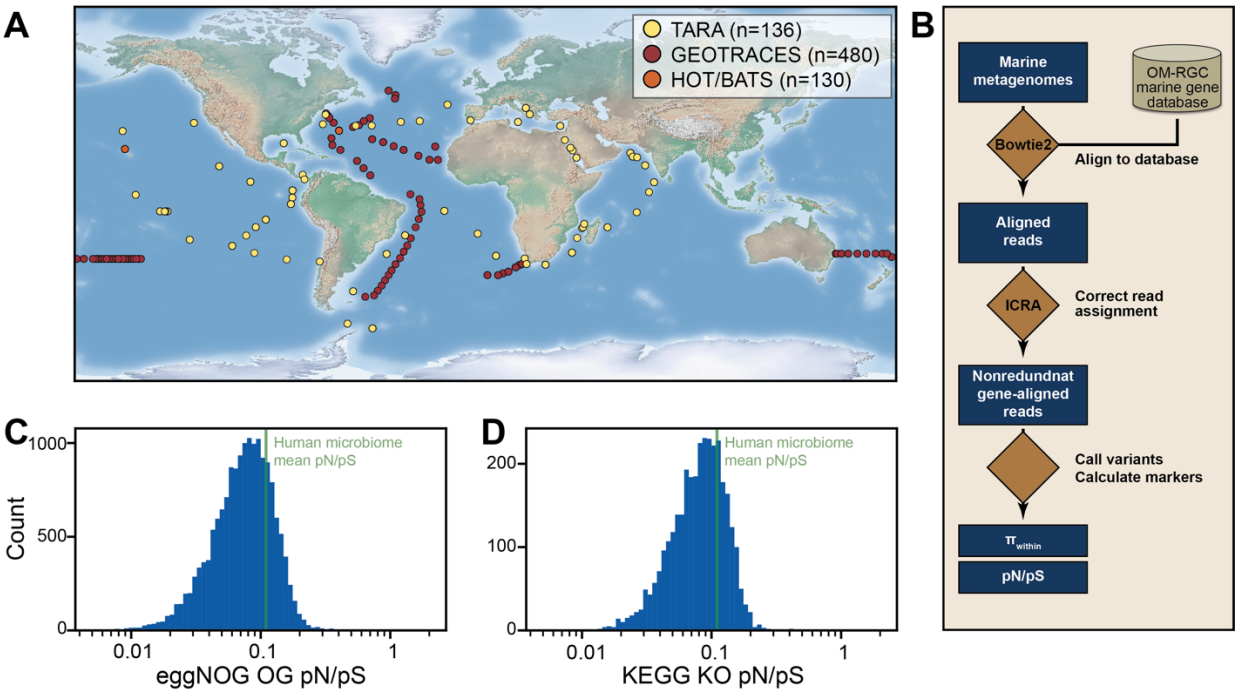



**Figure 1. Calculation of evolutionary metrics from marine metagenomic samples.** (A) Geographical overview of the samples used in this study. (B) Illustration of our computational pipeline. (C,D) Histogram of pN/pS rates for eggNOG orthologous groups (OG; C) and KEGG orthologs (KO; D) across all marine samples.

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

105 To quantify the effects of environmental conditions on selective forces acting on marine microbial
106 genes, we extracted measurements regarding the environment in which each sample was taken.
107 This included the depth of the sample, water temperature and salinity, as well as concentration of
108 the key molecules nitrate, nitrite, oxygen, phosphate and silicate (Fig. S1A-H; Methods). All these
109 measurements of environmental conditions are highly correlated with each other (Fig. S1I), and
110 also presented consistent correlation patterns with pN/pS of many KEGG and eggNOG orthologs
111 (Fig. S2), with low pN/pS at shallow depths and low nitrate concentrations. We therefore sought
112 to estimate the overall variance explained by the environment while accounting for these
113 covariations. To this end, we used a linear mixed model (LMM) with variance components,
114 commonly used in population genetics [14] (Methods). We defined the environmental covariates as
115 random effects in order to quantify the fraction of variance in pN/pS that is explained by resource
116 availability (i.e., environmental explained variance; EEV; Methods). Across both KEGG and
117 eggNOG orthologs we found that a substantial fraction of the variance in pN/pS can be attributed
118 to the environment, where across all orthologs this effect is significantly bigger than zero (Fig.
119 2A,B; Mann-Whitney *U* test $P < 10^{-16}$).

120 Different environmental niches may harbor different taxa with different trophic interactions that
121 could lead to differences in selective pressures. Thus, we sought to ensure that these
122 associations between pN/pS and the environment are not confounded by organismal differences
123 across different depths and nitrate concentrations. To this end, we analyzed genes belonging
124 exclusively to the genus *Synechococcus* (Methods). We calculated pN/pS across all of the coding
125 sequences combined and found a significant positive correlation with nitrate concentrations
126 ($P<10^{-20}$; Fig. S3A). To further validate that this correlation does not stem from the different
127 effective population sizes in the gradient of environmental nitrogen, we divided the samples into
128 five identically sized groups, based on environmental nitrate concentrations while constraining the

129 scope of each group to genes present in at least half of the samples. We show that an association
130 of pN/pS with nitrate extends to specific niches across the nitrocline, i.e., in concentrations higher
131 than 1 μmol/kg ($P<0.05$; Fig. S3B). These results demonstrate the existence of pN/pS gradient
132 as a function of nitrate concentrations, even in a single taxonomic group in a specific
133 environmental niche. This indicates that correlations of pN/pS with environmental variables are
134 not driven exclusively be organismal properties or by differences in trophic conditions across
135 environments, but rather exhibit a significant trend even within a single taxon, in specific niches.

136 Genes are not stand-alone entities, but are rather coded as a sequence in the genome of a
137 microbe. We therefore accounted for a potential non-random association structure between
138 pN/pS rates of different orthologs. To this end, we used a different setting of the LMM, now
139 including both the environmental covariates and pN/pS rates of all other orthologs. Even after
140 accounting for potential non-random association structure between pN/pS rates of different
141 orthologs due to clonal reproduction, the environmental effect was still significantly bigger than
142 zero ($P<0.05$). In particular, there is an overlap of over 40% in orthologs that are top ranked in
143 terms of EEV between these two model settings. Overall, we observed a very strong association
144 between environmental measurements and the magnitude of purifying selection exerted on most
145 orthologous gene groups. We also observed an association with environmental parameters
146 across many functional categories, including 'housekeeping' genes that are important for survival
147 in any given niche, and wished to further elucidate potential mechanisms that can explain it.

148

149 **Environment-associated selection is stronger in resource-consuming genes**

150 Previous studies[4,7] suggested that random mutations that lead to incorporation of additional
151 nutrients to the protein sequences of microbes result in a selective disadvantage. This implies
152 that genes whose protein sequences consume more resources would be under stronger
153 selection. Specifically, highly expressed genes would consume more resources and will therefore

154 be under stronger purifying selection[7,8,15]. In these genes, a mutation leading to incorporation of
155 additional resources would be magnified, as one highly transcribed DNA sequence could translate
156 to thousands of proteins, each consuming more resources. To quantitatively corroborate this
157 hypothesis and examine whether these underlying forces are reflected in the above associations
158 with pN/pS rates, we used an additional expression dataset for marine microbial genes [16] to rank
159 KEGG orthologs by their mean expression (Methods). We found that the most highly expressed
160 genes had a significantly higher fraction of the variance in their pN/pS explained by the
161 environment, as compared to the least expressed ones (Fig. 2C; Mann-Whitney *U* test $P<10^{-9}$).
162 We found a significant difference between the gradient of pN/pS rates, as a function of depth
163 (highly correlated with nitrate, nitrite and oxygen; Fig S1I), in highly expressed genes, as
164 compared to least expressed ones (Fig. S4A; Mann-Whitney *U* test $P<10^{-5}$), where the former
165 increased more sharply with depth. A few notable examples for genes with high expression and
166 environmental explained variance, as compared to other KEGG KOs, are the β subunit of RNA
167 polymerase (K03043; EEV = 0.82, 0.33 in the first and second LMM settings, respectively; ranked
168 first in both model settings), the β' subunit (K03046; EEV = 0.8, 0.28; ranked in the top five in
169 both settings) and a peptide/nickel transporter involved in quorum sensing (K02035; EEV = 0.81,
170 0.33; ranked second in both settings).

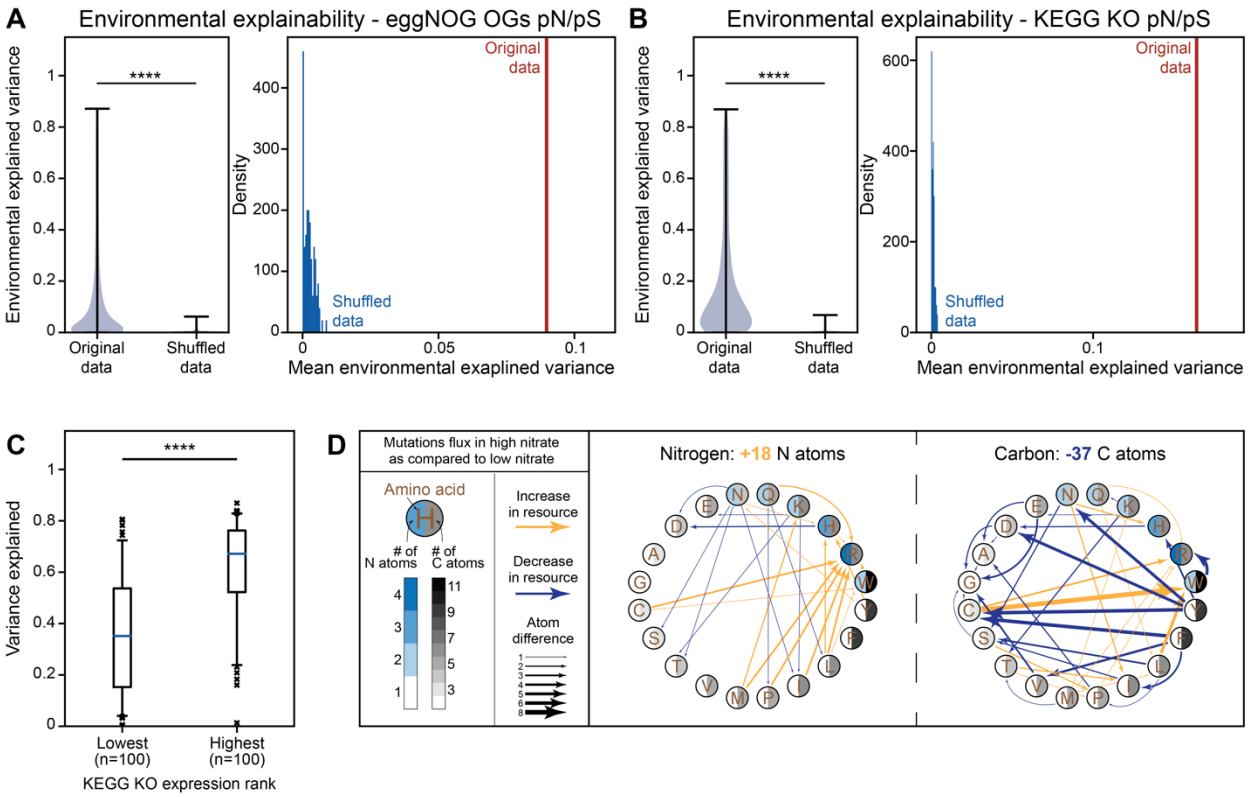



**Figure 2. Analysis of pN/pS rates reveals strong resource-driven selection.** (A) Left, violin plot of the variance of eggNOG OG pN/pS rates that was explained by the environment as compared to the same data with shuffled labels; Right, mean variance explained in unshuffled data (red) as compared to a histogram (blue) of mean variance explained in 100 executions with shuffled data. (B) Same as A, for KEGG KO pN/pS. (C) Box plots (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. (D) 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. ****, Mann-Whitney *U* $p<10^{-9}$.

We additionally hypothesized that genes coding for extracellular proteins will have a similar pattern as in this case, the resources excreted from the cell cannot be recycled. We found the same pattern of significantly higher EEV in extracellular protein-coding genes as compared to other gene groups (Methods; Fig. S4B; $P<0.05$). Overall, our results indicate that these genes exhibit higher ‘resource sensitivity’, manifested by higher variance explained by the environment,

188 potentially due to their high expression levels. This finding provides a data-driven evolutionary
189 perspective to theory and experiments showing lower resource incorporation in highly expressed
190 genes[7,8]. In summary, the variation in resource-consuming genes, (i.e., highly expressed and
191 extracellular protein-coding), further strengthens our results regarding the breadth of resource-
192 driven selection.

193

194 **Resource-driven selection exerts a strong effect on protein-coding sequences**

195 We next sought to quantify the effects of this resource-driven selection on protein-coding
196 sequences. To this end, we compared the codon mutation frequencies in low- and high-nitrate
197 samples, after accounting for simplex-related confounders (Methods), and found significant
198 differences in codon mutation frequencies (Fig. S5A-C). We sought to examine the typical change
199 in nutrient consumption in varying nitrate concentrations. We thus defined mutation flux, as the
200 ratio between a codon mutation and its reverse, and estimated it using the log-odds ratio between
201 the two (e.g., $log\ (p(AAA \rightarrow AAC)/p(AAC \rightarrow AAA))$). Notably, across all the mutations significantly
202 more prevalent in samples from high nitrate environments (Methods), averaged across amino
203 acids, we find a significant total increase in nitrogen (Fig. 2D; 18 N atoms summed across all
204 significant amino acid changes, P=0.0082), decrease in carbon (Fig 2D; -37 atoms, P=0.0165),
205 decrease in sulfur (-6 atoms, P=0.009), a significant decrease in molecular weight (-508.91 g/mol,
206 P=0.0193) and a non-significant decrease in oxygen (-6 atoms, P=0.1505). These results indicate
207 that the lower the nitrate concentrations are, the stronger the selection against mutations leading
208 to higher nitrogen incorporation in protein sequences.

209 While nitrate concentrations increase with depth (Fig. S1I), dissolved organic carbon
210 concentrations typically decrease[6]. Our results are supported by previous observations regarding
211 genomic and proteomic changes associated with environmental concentrations of nitrate[4].
212 Mutations in nitrate rich and, typically, carbon poor environments were shown to drive an increase

213 in genomic GC content, accompanied by higher rates of nitrogen incorporation and lower rates of
214 carbon incorporation into protein sequences. Here we show, in high resolution, the typical
215 mutations that underlie this phenomenon (Fig. 2D; Fig. S5D). As we base our analysis on pN/pS
216 rates, a proxy for the magnitude and direction of selection exerted on coding sequences, we
217 suggest that the differences observed in gene GC content across varying nitrate concentrations
218 are inseparable from changes to the proteome, and are possibly the result of resource-driven
219 selection exerted on these coding sequences.

220

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

222 The standard genetic code is known to be highly efficient in minimizing the effects of
223 mistranslation errors and point mutations[17–20]. This optimality is prominent among theories
224 regarding the origin of the genetic code[21–24]. According to the theory of error minimization,
225 selection to minimize the adverse effect of point mutations and translation errors was the principal
226 factor governing the evolution of the genetic code[25–32]. As a quantitative exploration of this theory
227 requires a well-defined cost function, a few measures of amino acid fitness were previously
228 suggested (e.g., PR scale, Hydropathy index) based on stereochemical theories and hydropathy
229 properties[33–36]. As we have observed strong patterns of selection for specific amino acids in
230 nutrient-limited environments, we hypothesized that resource conservation may also be a factor
231 in code error minimization.

232 Specifically, we hypothesized that the strong resource-driven selection, whose signature is visible
233 on protein-coding sequences across marine functional groups, may also have resulted in a
234 resource-optimized genetic code, such that the expected cost of a random mutation, in terms of
235 added resources, is minimized. To rigorously test this hypothesis, we first defined a cost function
236 for each element $e$ (e.g., carbon, nitrogen), such that the 'tariff' of a single mutation is the
237 difference in the number of atoms before and after the mutation. As an example, a missense

238 mutation from codon CCA to codon CGA results in an amino acid change from proline to arginine,
239 and an increase of 3 nitrogen atoms and one carbon atom, setting the nitrogen cost of such
240 mutation to 3 and the carbon cost to 1 (Fig. 3A).

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

244 $$ERMC_e(V_S) = \sum_{v,v' \in V_S} P(v)\mathrm{P}(mut(v,v'))\mathrm{c}_e^+(v,v')$$

245 Where $P(v)$ is the abundance of codon $v$, calculated from all marine samples; $\mathrm{P}(mut(v,v'))$ is the
246 probability of mutation from codon $v$ to codon $v'$, set to be the relative abundance of the single
247 nucleotide mutation driving this codon change (e.g. for mutation from GCA to CCA, we use the
248 abundance of G-to-C transversions), calculated from all mutations observed in fourfold
249 degenerate codons to avoid sites under strong selection; and $\mathrm{c}_e^+(v,v')$ is the 'tariff' of a single
250 mutation if an atom of element $e$ has been added to the post-mutation amino acid (Methods). For
251 the standard genetic code, and typical codon abundances and mutation rates calculated from
252 marine microbes, we report an ERMC of 0.440, 0.158 and 0.163 for carbon, nitrogen and oxygen
253 respectively, corresponding to an average increase of this number of atoms per random mutation.

254 To check if the genetic code, along with codon abundances and mutation rates in marine
255 microbes, is indeed robust to resource-consuming mutations, we compared it to other hypothetical
256 codes. To this end, we simulated alternative genetic codes by randomizing the first and second
257 position in all codons, while constraining the permutation in stop codons (Methods), creating a
258 null distribution of ERMC. We found that the standard genetic code, common to most life forms,
259 is parsimonious in terms of carbon and nitrogen utilization, given a random mutation, manifested
260 by minimization of the ERMC for nitrogen (Fig 3B; $ERMC_N$ P=0.0153) and carbon (Fig. 3B; $ERMC_C$

261 P=0.0109), while in the ERMC for oxygen we did not find a significant trend (Fig. 3B; P=0.8659).
262 Remarkably, only two out of 10,000 randomized genetic codes were more resource-robust than
263 the standard genetic code in conservation of nitrogen and carbon together (P=0.0002).
264 Nonetheless, these alternative codes are less conservative than the standard code in maintaining
265 hydrophobicity and hydrophilicity of amino acids given a random mutation (Methods; Fig. S6).
266 They may therefore lead to proteins that are more susceptible to structural changes in the event
267 of a random mutation, as was postulated previously in theories governing the evolution of the
268 genetic code.

269 Finally, we sought to confirm that our elemental cost function is not confounded by traditional
270 properties of amino acids such as the polar requirement (PR) and hydropathy index[33–36]. To this
271 end, we calculated the ERMC using these common cost functions for the standard genetic code
272 and for simulated alternative ones, and compared the overlap between traditional cost functions
273 and our elemental cost functions (Methods). For both cost functions, we found that optimality in
274 terms of carbon or nitrogen utilization implies lack of optimality in polar requirement or in
275 hydropathy (Table S1 nitrogen-PR, $P<10^{-16}$; nitrogen-hydropathy, $P<10^{-4}$; carbon-PR, $P<10^{-7}$;
276 carbon-hydropathy, $P<10^{-20}$). This result indicates that carbon and nitrogen conservation in the
277 genetic code is not confounded by previously reported optimization properties such as hydropathy
278 and PR. Altogether, these results indicate resource optimization in marine microbes is driven by
279 the structure of the genetic code, alongside specific amino acid choices.

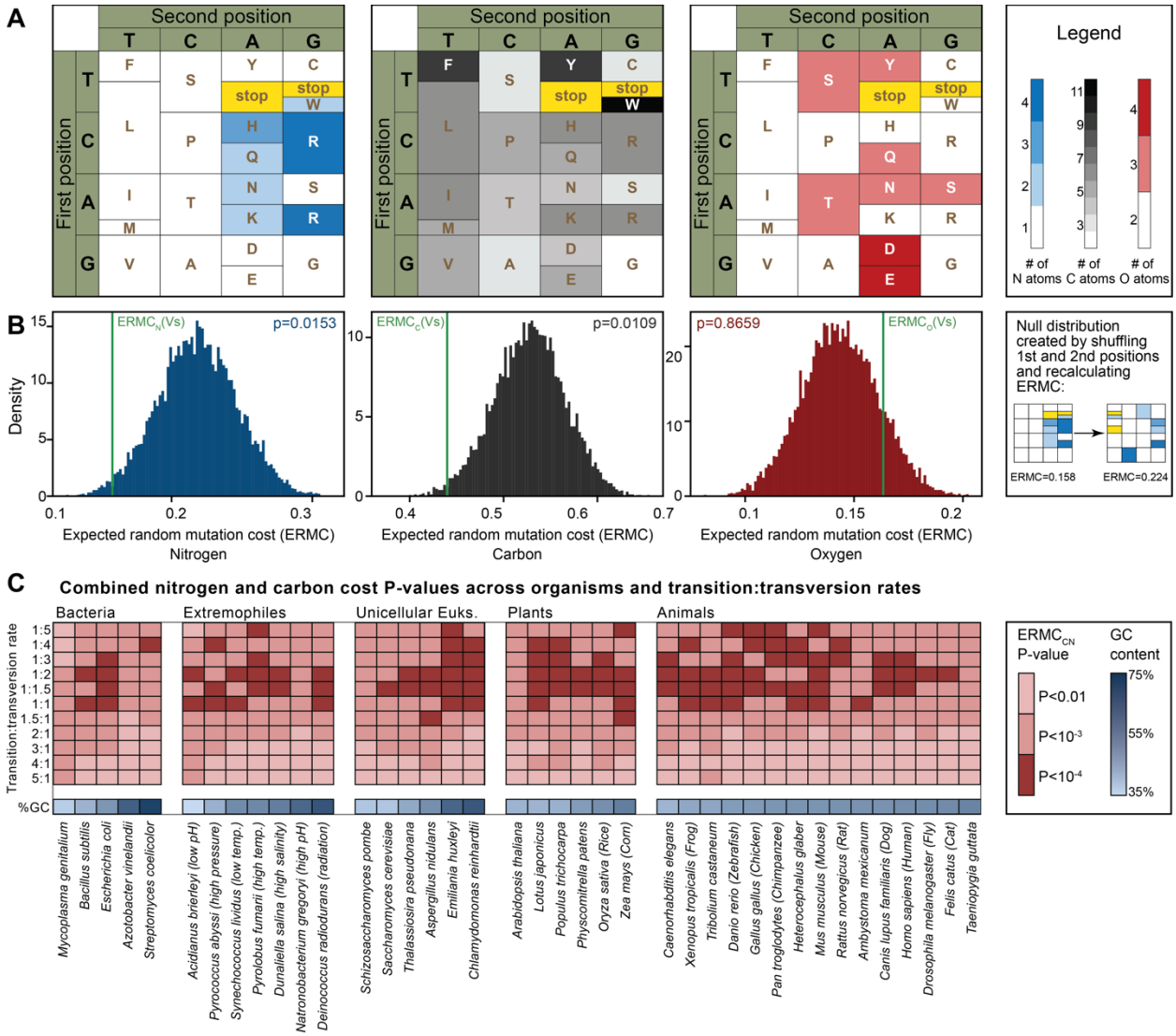



**Figure 3. Resource-conservation is facilitated by the genetic code.** (A) Nitrogen (left), carbon (center) and oxygen (right) content of different amino acids depicted along their positions in the standard genetic code. (B) Histograms of the expected random mutation cost (ERMC), in 10,000 random permutations of the genetic code for nitrogen (left, blue), carbon (center, black) and oxygen (right, red). Green 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.

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

290 To show that the robustness of the genetic code in terms of resource-consumption was not limited
291 to our dataset and analytic approach, we calculated the ERMC of 187 species of genera
292 *Prochlorococcus* and *Synechococcus*. We calculated codon abundances and mutation rates
293 using prior knowledge of both protein-coding sequences[5] and the published
294 transition:transversion rate of 2:1[37] (Methods). By testing the ERMC of the standard genetic code
295 against a null distribution generated, as before, given these known parameters rather than ones
296 inferred from marine samples, we were able to reveal significant conservation of carbon, nitrogen,
297 and both elements combined (Fig. S7A; $ERMC_C$ mean P=0.013, P=0.020; $ERMC_N$ mean
298 P=0.049, P=0.032; $ERMC_{CN}$ P=0.0004, P=0.0007 for *Prochlorococcus* and *Synechococcus*,
299 respectively). To account for inaccuracies and variation in the known parameters, we next
300 calculated the ERMC null distribution for a wide range of transition:transversion rates. We show
301 that the ERMC of the standard genetic code remains significantly conserved for nitrogen, carbon,
302 and both elements combined for most physiological transition:transversion rates (Fig. S7B),
303 indicating that the structure of the genetic code and codon abundances of organisms are the
304 driving force behind genetic code optimization.

305 To explore whether this optimality in the genetic code in terms of nutrient conservation extends
306 across different lifeforms, we performed a similar calculation using codon abundances from 39
307 organisms across all domains of life, including all human protein-coding sequences, and tested a
308 range of transition:transversion rates (Methods). We find that, similarly to marine microbes, the
309 genetic code features optimization in terms of resource utilization for all tested organisms,
310 manifested by a significant minimization of the combined ERMC of nitrogen and carbon in all
311 transition:transversion rates tested (P<0.01, Fig. 3C). Moreover, we find significant optimization,
312 albeit of a lower magnitude, even in the theoretical case where all codon abundances are the
313 same (Fig. S7C). The codon abundances of a great majority of organisms also demonstrate

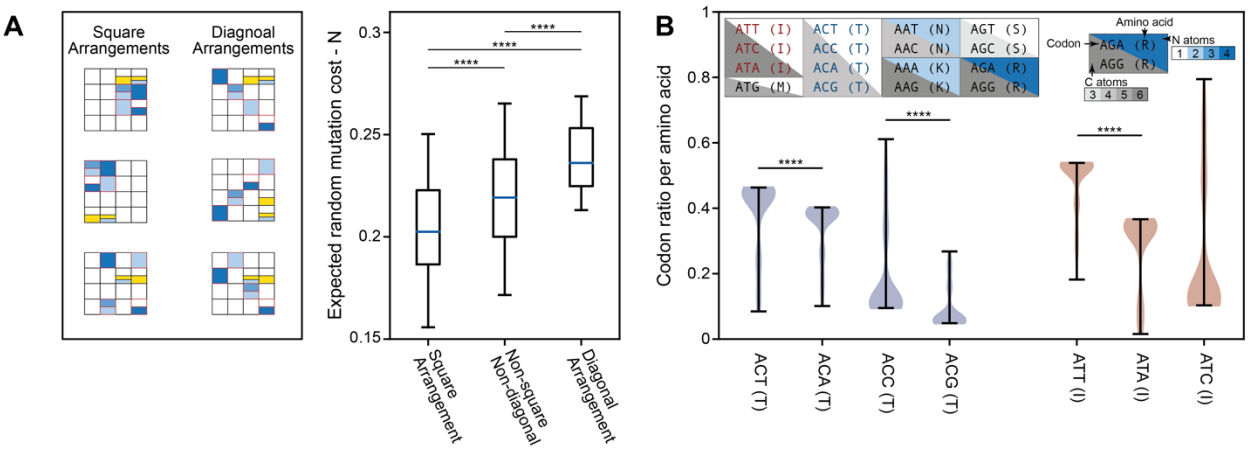
314 significant minimization of ERMC in nitrogen (Fig. S7D) and carbon (Fig. S7E), given a random
315 mutation, for a wide range of transition:transversion rates. These results indicate that resource
316 optimization in the genetic code transcends taxonomy, codon choices, and mutation rates. It
317 shows that the genetic code may have structural properties that make it robust in terms of
318 resource-consumption. It is also possible that amino acid and codon usage in organisms has
319 evolved to lower nutrient consumption in case of a random mutation, informed by the structure of
320 the code.

321

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

323 We next wished to examine the organizing principles that drive the strong resource optimization
324 evident in the standard genetic code. We observed that codons of the nitrogen-rich amino acids
325 histidine, glutamine, asparagine, lysine and arginine span only two nucleotides in their first
326 position and two in their second position. We define this organization to be a 'square' arrangement
327 (Fig. 4A; Methods), and hypothesize that, as compared with other arrangements, it amplifies
328 nitrogen conservation. Specifically, in the square arrangement, codons coding for some amino
329 acids (alanine, valine, phenylalanine, and several leucine and serine codons) require at least two
330 mutations to increase the number of nitrogen atoms in the resulting amino acid. This is in contrast
331 to other hypothetical arrangements, including a 'diagonal' arrangement in which nitrogen-rich
332 amino acid codons span all possible nucleotides in the first and second positions (Fig 4A;
333 Methods). We suggest that the diagonal arrangement would be nutrient-wasteful, as in these
334 arrangements a single mutation could increase the nitrogen content of a protein sequence in more
335 than one way. To rigorously test this hypothesis, we generated 10,000 random genetic codes,
336 with 220 arrangements happening to embody a square structure, and 127 a diagonal one. We
337 found that, when compared to all other possible arrangements, square arrangements present a
338 significantly lower ERMC (Fig 4A; Mann-Whitney $U$ $P<10^{-10}$) while diagonal arrangements exhibit

339 a significantly higher ERMC (Fig. 4A; Mann-Whitney *U* $P<10^{-10}$). This result demonstrates that
340 resource optimization in the genetic code is driven by structural principles, perhaps underlying
341 the significant optimization observed across kingdoms.



342

343 **Figure 4. Structural properties and codon usage bias underlying optimality in the genetic code.**
344 (A) Box plots (line, median; box, IQR; whiskers, 5th and 95th percentiles) of $ERMC_N$ of square
345 arrangements (left) and diagonal arrangements (right, Methods), as compared to all other
346 arrangements (center) out of 10,000 randomized arrangements of the code. (B) Violin plot of codon
347 usage among 187 species of *Prochlorococcus* and *Synechococcus* showing significant preference of
348 threonine codons ACT and ACC as compared to ACA and ACG, and of isoleucine codon ATT as
349 compared to ACA. ****, $P<10^{-10}$.

350

351 Finally, we hypothesized that codon usage for a single amino acid may also be biased due to
352 differential cost of a random mutation for each codon. We therefore examined all amino acids
353 coded by codons with adenine in their first position, focusing on codon usage of the amino acid
354 threonine. We note that a C-to-G transversion in the second position for codons ACT and ACC
355 yields serine (AGT and AGC, respectively), and that the same mutation for codons ACA and ACG
356 yields arginine (AGA and AGG, respectively; Fig. 4B, inset). As arginine has higher carbon and
357 nitrogen content than serine, and lysine a higher carbon content than asparagine, we
358 hypothesized that following a nutrient-conservative model, codons ACT and ACC will have a
359 higher abundance than codons ACA and ACG, respectively, given a known genomic GC content.

360 We thus examined codon usage in 187 *Prochlorococcus* and *Synechococcus* strains, and show
361 a significantly higher use of ACT as compared to ACA (Fig 4B; Wilcoxon signed-rank test $P<10^{-}$
362 $^{20}$) and ACC as compared to ACG (Fig 4B; Wilcoxon signed-rank test $P<10^{-20}$). Similarly,
363 Isoleucine codon ATT has higher abundance as compared to ATA (Fig 4B; Wilcoxon signed-rank
364 test $P<10^{-20}$). These results demonstrate that resource conservation is a central driving force in
365 selection processes guiding codon usage and may affect not only protein sequence but also
366 cellular translation efficiency.

367

368 **Discussion**

369 In this work, we use the proxy of pN/pS rates to show strong purifying selection acting upon
370 protein-coding genes in the marine environment. We demonstrate that a substantial fraction of
371 the variance in pN/pS rates could be attributed to environmental factors, and highlight a strong
372 association of these rates with nitrate concentrations. We show that the variance in pN/pS rates,
373 across resource-consuming genes (i.e., highly expressed and extracellular protein-coding), can
374 be attributed to environmental factors, suggesting that stronger resource-driven selection is
375 exerted upon them. Using single nucleotide polymorphisms from across marine samples, we
376 characterize the typical mutations in nitrate-rich versus nitrate-poor environments and show that
377 these drive incorporation of additional nitrogen-rich amino acids and fewer carbon-rich amino
378 acids to protein sequences. Finally, we provide evidence that the standard genetic code, shared
379 among most lifeforms, facilitates resource conservation, demonstrating that along with codon
380 choices, it is conservative in incorporating additional atoms of nitrogen and carbon given a random
381 mutation. Notably, we show, across tens of thousands of simulated genetic codes, that the
382 standard genetic code surpasses almost all other random simulated codes in conservation of
383 nitrogen and carbon, across multiple taxa from all domains of Life and across multiple codon
384 choices and mutation rates.

385 Hypotheses regarding the origin of the genetic code include stereochemical affinity between a
386 codon or anticodon and their amino acid[38,39]; a frozen accident theory[24], relying on the fact that
387 the code is highly immutable; a co-evolution of the genetic code with the emergence of amino
388 acid-producing biosynthetic pathways[22]; and an early fixation of an optimal genetic code,
389 suggesting that the code evolved under selection for error minimization[40]. Our observations are
390 in line with the latter theory of optimality, and suggest that the genetic code may have been
391 optimized also for nutrient conservation. While we do not know the nature of nutrient cycling in
392 the primordial ocean, we hypothesize that scarcity of nitrogen and carbon that are common now
393 may have also prevailed alongside early lifeforms. Thus, an organism harboring a nitrogen- and
394 carbon-efficient genetic code would have had a selective advantage over its peers, especially in
395 the absence of fully evolved DNA mutation repair mechanisms.

396 We note that while we observe resource optimization for nitrogen and carbon conservation in the
397 standard genetic code, oxygen conservation is not optimized. We offer several hypotheses
398 regarding this lack of optimization. First, it is possible that since oxygen is highly abundant in
399 organic molecules, it is less of a limiting factor as compared to carbon and nitrogen and therefore
400 its optimization does not confer a selective advantage. Another option is that oxygen-rich amino
401 acids may function in cellular processes that are independent of protein synthesis and are
402 therefore more readily available and thus not optimized. Prominent examples for this hypothesis
403 are aspartate, which also performs an important function in the malate-aspartate shunt, and
404 glutamate, which plays a role in countless cellular processes.

405 Overall, by using publicly available data on ocean microbes and their corresponding
406 environmental measures, we were able to discern strong purifying selective pressure which
407 shapes marine microbial life, and may have even shaped the structure of the genetic code that
408 was since preserved for billions of years. With the advent of new multi-omic data from

environmental studies, we will be able to better divulge the intricate relationships of microbes with a rapidly changing global environment.

411 **Methods**

412

413 **Marine microbiome samples**

414 Marine samples collected with Tara oceans[3], bioGEOTRACES[10], the Hawaii Ocean Timeseries
415 (HOT) and the Bermuda Atlantic Timeseries Series (BATS)[10] were downloaded from ENA with
416 accessions ENA:PRJEB1787 (TARA oceans prokaryotic fraction), ENA:PRJNA385854
417 (bioGEOTRACES) and ENA:PRJNA385855 (HOT/BATS), each sample with a minimum of 5
418 million reads.

419

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

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

426

427 **Determining metagenomic read assignment probability**

428 We determined the probability of assignment of metagenomic reads to marine microbial genes
429 using the Iterative Coverage-based Read-assignment Algorithm (ICRA)[43] with parameters
430 `max_mismatch=12, consider_lengths=True, epsilon=1e-6, max_iterations=30, min_bins=4,`
431 `max_bins=100, min_reads=10, dense_region_coverage=60, length_minimum=300,`
432 `length_maximum=2e5, use_theta=False`. To prevent spurious mapping, alignments were
433 considered for downstream analysis only if the probability of alignment was higher than 0.9.

434 **Variant calling**

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

438 To facilitate variant calling in tractable timescales, each filtered, untied and sorted .bam file was
439 split into chunks, each encompassing 10,000 reference sequences (out of about 40 million
440 reference sequences). For each such batch of reference sequences, we called variants across
441 all samples using the following command: bcftools mpileup --threads 4 -a FORMAT/AD -Q 15 -L
442 1000 -d 100000 -m 2 -f <OM-RGC fasta> <bam filenames> | bcftools call --threads 4 -Ov -mv -o
443 <output vcf>, where <OM-RGC fasta> is the fasta file of OM-RGC nucleotide sequences, <bam
444 filenames> are the filenames of all .bam files pertaining to the reference sequence chunk in
445 question, and <output vcf> is the output .vcf file pertaining to that same chunk.

446 Single nucleotide variants were considered as SNPs if they had an allele frequency if at least 1%
447 [44], were supported by at least 4 reads across samples, and had a GATK quality score of at least
448 30.

449 For a sample mapped to a reference gene to be considered for downstream analysis, we
450 demanded that at least 60% of SNPs called along the length of the reference gene for that sample
451 would be supported by at least 4 reads, thereby enabling accurate calculation of pN/pS rates. For
452 a gene to be considered for downstream analysis, we demanded for that gene to have at least
453 one SNP common to 20 or more samples.

454

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

456 While comparing SNP patterns across samples, it is instrumental to avoid biases due to
457 differences in coverage. We therefore downsampled the read coverage depth to the minimum

458 depth across all samples, for each position in a sample that was supported by more than 4 reads.
459 For positions that had minimum support of fewer than 4 reads, no subsampling was performed.
460 Subsampling was performed by drawing from a multinomial distribution, with *n* trials and variant
461 probabilities *p*, where *n* was set to the calculated minimum depth and *p* set to the relative
462 abundance of each variant in the given sample.

463 The expected ratio of non-synonymous and synonymous substitutions was calculated by
464 considering all called SNPs in every gene. First, we calculate a consensus sequence for each
465 gene by taking, for each SNP position, the variant that was overall more common across all
466 samples (after the subsampling performed above). We counted, for each gene, the number of
467 non-synonymous and synonymous sites across the consensus sequence. For each SNP position
468 in each sample, we counted the number of synonymous and nonsynonymous substitutions. As
469 more than one variant can exist in a single sample, we considered the relative abundance of
470 synonymous to nonsynonymous substitutions dictated by the different variants. For example, if
471 the reference codon was CAC, coding for histidine, one variant, a C-to-G transversion in the third
472 position abundant at 50%, led to a nonsynonymous mutation that resulted in glutamine (CAG)
473 while another variant in the same sample and the same position was a synonymous C-to-T
474 transition, we counted 0.5 synonymous substitutions and 0.5 nonsynonymous substitutions. We
475 followed by calculating the pN/pS ratio:

476 $$pN/pS = \frac{nonsynonymous\ substitutions}{nonsynonymous\ sites} / \frac{synonymous\ substitutions}{synonymous\ sites}$$

477 pN/pS characterizes selective constraints at the population level, as opposed to dN/dS that
478 characterizes it between individual species [12] and can thus be standardized to a specific time
479 interval and used as an absolute metric. Nonetheless, dN/dS ratios are not applicable in our
480 study since polymorphic sites derived from short read sequencing impede having haplotypes
481 which are a prerequisite for calculating dN/dS.

482

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

484 Functional assignments to KEGG KOs and eggNOG OGs for all OM-RGC genes were computed
485 using eggNOG-mapper v2 based on the eggNOG v5.0 database[41,45]. For each functional
486 assignment in each sample, all OM-RGC genes assigned with the same functional assignment
487 were concatenated and treated as one long genomic sequence per the calculation of pN/pS ratios.
488 To reduce noise in pN/pS calculation, we considered only KOs and OGs that had at least 5 genes
489 per sample, in at least 50 samples.

490

491 **Environmental variables**

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

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

500 bioGEOTRACES metadata was compiled from CTD sensor data and discrete sample data from
501 the GEOTRACES intermediate data product v.2[46]. HOT metadata was downloaded from the ftp
502 server of the University of Hawai'i at Manoa (ftp://ftp.soest.hawaii.edu/hot/) and BATS metadata
503 was downloaded from the Bermuda Institute of Ocean Sciences (http://bats.bios.edu/bats-data/).
504 As GEOTRACES/HOT/BATS ocean water samples are not linked to specific biological samples
505 as is the case with Tara oceans samples, we considered only water samples from the exact same

geographic location, within a day from biological sample collection time, and within 5% difference in depth of collection, and chose the closest sample in terms of time and depth of collection. As all these measurements of environmental conditions are highly correlated with each other (Fig. S1I), we utilized this correlation structure to impute missing values using the EM algorithm[47].

**Linear mixed models**

***Generative model*** Consider a collection of $M_i, where\ i \in \{1,2\}$ features (i.e., $M_{1,2}$number of KEGG and eggNOG orthologs respectively), each measured across $K$ samples. We get as input an $(M_i \times K)$ matrix $O^{\ i}$, where ${O_{kj}}^i$ is the pN/pS of ortholog $j$ in sample $k$. Let ${y^i}_m = ({O_{m1}}^i, \ldots, {O_{mK}}^i)$ be a $K \times 1$ vector representing the pN/pS in ortholog $m$, according to grouping $i$, across $K$ samples (e.g., pN/pS in KEGG KO K02274 across *K* samples). Let $W$ be a $(K \times q)$ normalized matrix of environmental measurements. This included the depth of the sample, water temperature and salinity, as well as concentration of the key molecules nitrate, nitrite, oxygen, phosphate and silicate.

With these notations, we assume the following generative linear model

$$ {y^i}_m = W u_m + \epsilon_m \qquad (1) $$

Where $u_m$ and $\epsilon_m$ 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 $\sigma^2_{\epsilon_m}$.

It is easy to verify that an equivalent mathematical representation of model (1) is given by

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

528 where $\sigma_K^2 = M^i \sigma^2{}_{u_m}, K = \frac{1}{M^i} WW^T$. We will refer to $K$ as the environmental kinship matrix, which

529 represents the similarity, in terms of environmental covariates, between every pair of samples

530 across grouping $i$ (i.e., represent the correlation structure to the data).

531

532 **Environmental explained variance**: The explained variance of a specific feature $y^i{}_m$ by the

533 environmental measurements

534 $$\chi^i{}_m = \frac{\sigma_K^2}{\sigma_K^2 + \sigma_{\epsilon_m}^2} \qquad (3)$$

535 In the second model setting, we wished to account for a potential non-random association

536 structure between pN/pS rates of different orthologs. To this end, we included both the

537 environmental covariates and pN/pS rates of all orthologs not inferred as two different sets of

538 variance components.

539 In this setting, Let $W_1$ be an $(K \times q)$ normalized matrix of environmental measurements, as before

540 and $W_2$ be an $(K \times M_i - 1)$ normalized matrix of pN/pS measurements according to grouping $i$

541 (i.e., KEGG or eggNOG orthologs) across $K$ samples, where for each $y^i{}_m$ we exclude the pN/pS

542 rates of the focal ortholog $m$.

543 With these notations, we assume the following model

544 $$y^i{}_m = W_1 u_{1m} + W_2 u_{2m} + \epsilon_m \qquad (4)$$

545 It is easy to verify that an equivalent mathematical representation of model (1) is given by

546 $$y^i{}_m \sim N(0, \sigma_{K_1}^2 K_1 + \sigma_{K_2}^2 K_2 + \sigma_{\epsilon_m}^2 I) \qquad (5)$$

547 where $\sigma^2{}_{K_1} = M^i \sigma^2{}_{u_{1m}}, \sigma^2{}_{K_2} = M^i \sigma^2{}_{u_{2m}}, K_1 = \frac{1}{M^i} W_1 W_1{}^T, K_2 = \frac{1}{M^i} W_2 W_2{}^T. K_1$ and $K_2$ represent

548 the similarity, in terms of environmental covariates and pN/pS rates, between every pair of

549 samples across grouping $i$ (i.e., represent the correlation structure to the data).

550 In this setting, the environmental explained variance is:

551 $$\chi^{i}{}_{m} = \frac{\sigma^2{}_{K_1} + \sigma^2{}_{K_2}}{\sigma^2_{K_1} + \sigma^2_{K_2} + \sigma^2_{\epsilon_m}} \quad (6)$$

552

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

554 Using expression data from 4,092 KEGG KOs collected by Kolody et al. [16], we ranked the KO
555 genes in our marine samples in the following way. We first represent the expression data in
556 relative abundance space (normalize each sample by its read counts). Next, for each KO $i$, where
557 $i \in \{1, \ldots, I\}$ , we sum across different instances of this focal KO. The input is an $m \times n$ matrix,
558 where $m$ is the number of different instances of focal KO $i$, and $n$ is the number of samples. The
559 output is a vector of length $n$: $(x_1, \ldots, x_n)$. Finally, we average the expression levels
560 $(x_1, \ldots, x_n)$ across samples: $\frac{1}{n}\sum_{j=1}^{n} x_j$ and rank the KOs based on the calculated average
561 expression. Notably, we limited the scope of our analysis to samples collected only in small
562 fraction filters (0.22 $\mu m$).

563

564 **Determination of *Synechococcus*-specific pN/pS rates**

565 We identified genes from the OM-RGC database belonging exclusively to genus *Synechococcus*
566 using eggNOG-mapper. We filtered out genes that were present in fewer than 20% of all samples.
567 For each of the samples, we calculated pN/pS on all gene sequences combined. We further
568 divided the samples into five identically sized groups, based on environmental nitrate
569 concentrations, and for each group filtered out genes that were present in fewer than 50% of all
570 samples in the group. We next calculated pN/pS on all the gene sequences that remained in each
571 of the groups.

572

573 **Determination of extracellular genes**

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

582

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

584 We created matrix $H^{(U)}$ as follows: Consider a set $U$ of all genes for which SNP measurements
585 exist and a subset $T \subseteq U$ of this set across $K$samples. Let ${G^{(T)}}_j = (V, E)$ be a codon graph for
586 subset $T$ and sample $j$, where $v \in V$ is a codon (e.g., CUU coding for Alanine) and $(v, v') \in E$ if
587 and only if $v$ and $v'$ are one mutation apart (e.g., CUU for Alanine and CAU for Histidine). Let
588 ${w^{(T)}}_j : (v, v') \rightarrow [0,1]$ be a weight function where ${w^{(T)}}_j(v, v') =$
589 $\frac{Number\ of\ mutations\ turning\ v\ to\ v\prime\ in\ subset\ T\ and\ sample\ j}{Number\ of\ occurrences\ of\ codon\ v\ in\ subset\ T\ and\ sample\ j}$. Let $H^{(T)}$ be a matrix of dimension $(|E| \times K)$,
590 where ${H^{(T)}}_{(v,v\prime),j} = {w^{(T)}}_j(v, v')$.

591 We next sum-normalized H per each sample and compared the codon mutation frequencies
592 between the 40 lowest- and 40 highest-nitrate samples. Despite significant differences in codon
593 mutation frequencies between low-nitrate and high-nitrate samples, some of the difference could
594 be driven by the simplex properties of the sum-normalized codon mutation frequencies, and some
595 could be attributed to the different rates of synonymous mutations between the high- and low-
596 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 ${F^{(U)}}_{(v,v'),j} = {H^{(U)}}_{(v,v'),j} - {H^{(U)}}_{(v',v),j}$. 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 $\mathrm{P}(mut(v, v'))$ the probability of a single mutation from codon $v$ to $v'$. Let $c_e: V \times V \rightarrow \boldsymbol{Z}$ 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:

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

618 And the ERMC cost as:

619 $$ERMC_e(V) = \sum_{v,v' \in V} P(v)\mathrm{P}(mut(v,v'))c^{+}{}_{e}(v,v')$$

620 Where

621 $$c^{+}{}_{e} = max(0, c_e)$$

622 We estimate $ERMC_e(V)$ as follows:

623 1. We define $P(v)$ as the median abundance of all codons $v'$ coding for the same amino acid
624 as $v$.

625 2. We wished to calculate mutation rates in sites that were under minimal selection. To this
626 end, we estimated $\mathrm{P}(mut(v,v'))$ by calculating, from fourfold-degenerate synonymous
627 mutation sites the average abundance of each single nucleotide mutation (e.g. A to C)
628 across all genes in which there are called SNPs in all ocean samples, excluding stop
629 codons. We then estimate $\mathrm{P}(mut(v,v'))$ using the relative abundances of all pairs of single
630 nucleotide mutations. We estimate $\mathrm{P}(mut(v,v'))$ for *Prochlorococcus*, *Synechococcus*
631 and Human genomes using published transition:transversion rates[5,48].

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

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

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

638 To confirm that our elemental cost function is not confounded by traditional properties of amino
639 acids such as the polar requirement (PR) and hydropathy index[33–36], we calculated the expected

640 random mutation cost (ERMC), per genetic code, using these common cost functions across 1
641 million simulated alternative codes. To this end, we randomized the first and second position of
642 all codons, while maintaining that the two sets of first and second positions in which the stop
643 codons reside are separated by a single transition mutation. We next calculated a contingency
644 table for each pair of cost functions for both nitrogen and carbon (i.e., $ERMC_N(V)$:$ERMC_{PR}(V)$,
645 $ERMC_N(V)$:$ERMC_{hydropathy}(V)$, $(V)$:$ERMC_{PR}(V)$, $ERMC_C(V)$:$ERMC_{hydropathy}(V)$). We assign
646 each code to one of four bins in the following way: (1) surpassing the standard genetic code in
647 both cost functions (e.g., nitrogen and PR), (2) surpassing the standard genetic code only in
648 element $e$ cost (e.g. only nitrogen), (3) surpassing the standard genetic code only in the traditional
649 cost function (e.g., PR), (4) not surpassing the standard genetic code in neither. Finally, we
650 applied the Chi-square test of independence with two degrees of freedom to each contingency
651 table.

652

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

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

664

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

666 We downloaded *Prochlorococcus* and *Synechococcus* protein-coding gene sequences (where
667 available) from the Joint Genome Institute (https://genome.jgi.doe.gov/portal/) following
668 accession numbers published by Berube et al. [5]. To estimate codon relative abundance $P(v)$, we
669 counted and sum-normalized codons in all protein-coding genes for each species. To estimate
670 codon mutation rate $P\ \ (mut(v, v'))$ we used the published transition:transversion rate of 2:1 for
671 *Prochlorococcus* and *Synechococcus* [37].

672

673 **Multiple taxa ERMC calculation**

674 To calculate ERMC for 39 taxa across multiple transition:transversion rates, we downloaded
675 codon usage and GC-content data collected by Athey et al.[49]. We used codon usage counts to
676 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,
677 5:1) to estimate $\mathrm{P}(mut(v, v'))$.

678 **References**


679 1. Morris, R. M. *et al.* SAR11 clade dominates ocean surface bacterioplankton communities.
680 *Nature* **420**, 806–810 (2002).

681 2. del Giorgio, P. A. & Duarte, C. M. Respiration in the open ocean. *Nature* **420**, 379–384
682 (2002).

683 3. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**,
684 1261359 (2015).

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

687 5. Berube, P. M., Rasmussen, A., Braakman, R., Stepanauskas, R. & Chisholm, S. W.
688 Emergence of trait variability through the lens of nitrogen assimilation in Prochlorococcus.
689 *Elife* **8**, (2019).

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

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

694 8. Akashi, H. & Gojobori, T. Metabolic efficiency and amino acid composition in the proteomes
695 of Escherichia coli and Bacillus subtilis. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 3695–3700
696 (2002).

697 9. Batut, B., Knibbe, C., Marais, G. & Daubin, V. Reductive genome evolution at both ends of
698 the bacterial population size spectrum. *Nat. Rev. Microbiol.* **12**, 841–850 (2014).

699 10. Biller, S. J. *et al.* Marine microbial metagenomes sampled across space and time. *Sci Data*
700 **5**, 180176 (2018).

701 11. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature*
702 **493**, 45–50 (2013).

703 12. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila.
704 *Nature* **351**, 652–654 (1991).

705 13. Garud, N. R., Good, B. H., Hallatschek, O. & Pollard, K. S. Evolutionary dynamics of
706 bacteria in the gut microbiome within and across hosts. *PLoS Biol.* **17**, e3000102 (2019).

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

709 15. Koonin, E. V. Are There Laws of Genome Evolution? *PLoS Comput. Biol.* **7**, e1002173
710 (2011).

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

714 17. Koonin, E. V. & Novozhilov, A. S. Origin and evolution of the genetic code: the universal

enigma. *IUBMB Life* **61**, 99–111 (2009).

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

19. Woese, C. R., Hinegardner, R. T. & Engelberg, J. Universality in the Genetic Code. *Science* **144**, 1030–1031 (1964).

20. Woese, C. R. Order in the genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **54**, 71–75 (1965).

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

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

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

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

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

26. Massey, S. E. Genetic code evolution reveals the neutral emergence of mutational robustness, and information as an evolutionary constraint. *Life* **5**, 1301–1332 (2015).

27. Massey, S. E. The neutral emergence of error minimized genetic codes superior to the standard genetic code. *J. Theor. Biol.* **408**, 237–242 (2016).

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

29. Novozhilov, A. S., Wolf, Y. I. & Koonin, E. V. Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. *Biol. Direct* **2**, 24 (2007).

30. Salinas, D. G., Gallardo, M. O. & Osorio, M. I. Local conditions for global stability in the space of codons of the genetic code. *Biosystems.* **150**, 73–77 (2016).

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

32. Zhu, W. & Freeland, S. The standard genetic code enhances adaptive evolution of proteins. *J. Theor. Biol.* **239**, 63–70 (2006).

33. de Oliveira, L. L., de Oliveira, P. S. L. & Tinós, R. A multiobjective approach to the genetic code adaptability problem. *BMC Bioinformatics* **16**, 52 (2015).

34. Di Giulio, M. & Medugno, M. The level and landscape of optimization in the origin of the genetic code. *J. Mol. Evol.* **52**, 372–382 (2001).

35. Higgs, P. G. A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. *Biol. Direct* **4**, 16 (2009).

36. Sengupta, S. & Higgs, P. G. Pathways of Genetic Code Evolution in Ancient and Modern Organisms. *J. Mol. Evol.* **80**, 229–243 (2015).

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

38. Hopfield, J. J. Origin of the genetic code: a testable hypothesis based on tRNA structure, sequence, and kinetic proofreading. *Proc. Natl. Acad. Sci. U. S. A.* **75**, 4334–4338 (1978).

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

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

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

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

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

44. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

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

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

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

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

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

## Supplementary Material

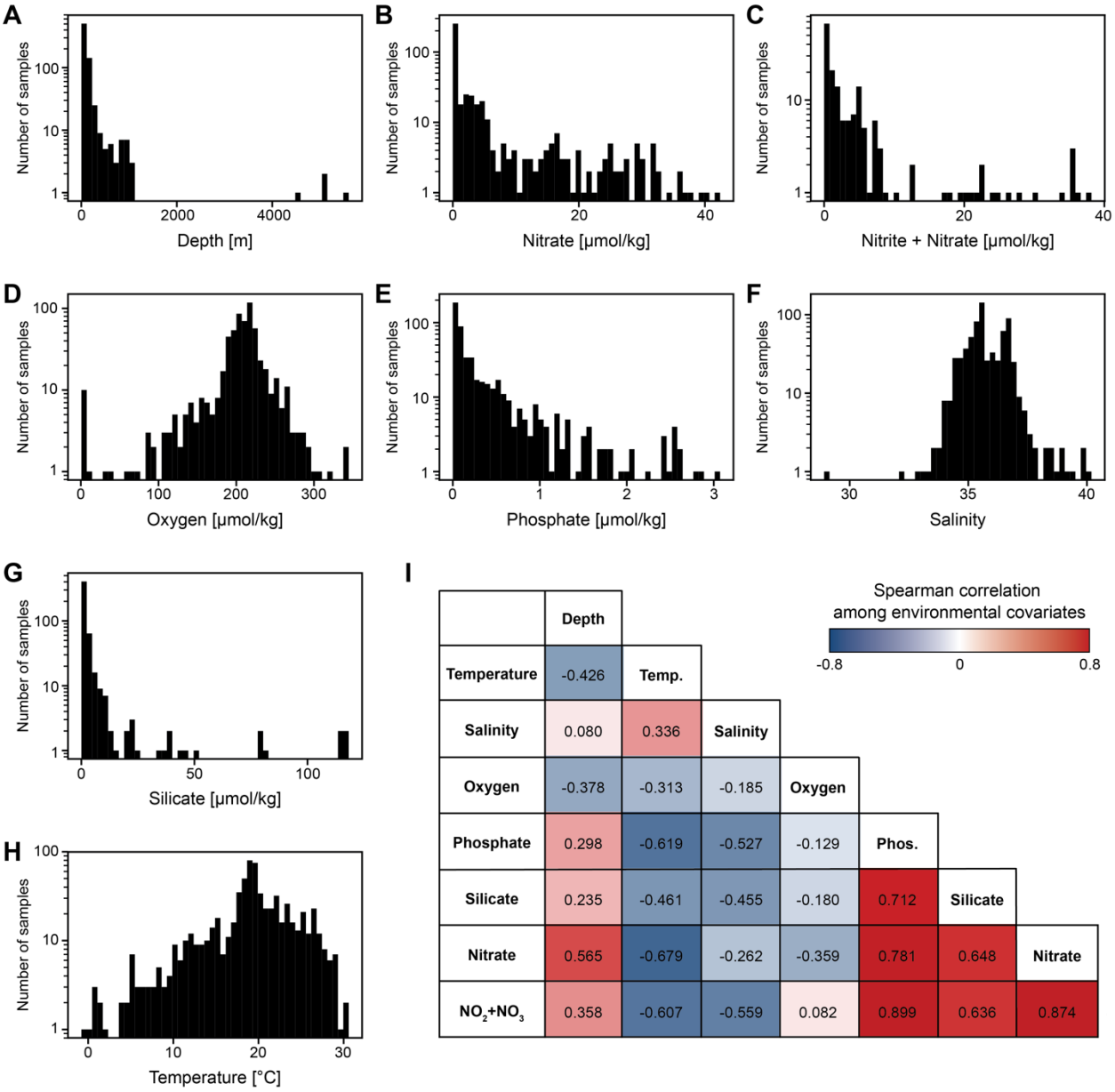



**Figure S1.** (A-H) Distribution of measurements taken alongside marine microbial samples for depth (A), nitrate (B), nitrate and nitrite (C), oxygen (D), phosphate (E), salinity (F), silicate (G) and temperature (H). (I) Spearman correlation coefficients between all pairs of environmental measurements across all available samples.

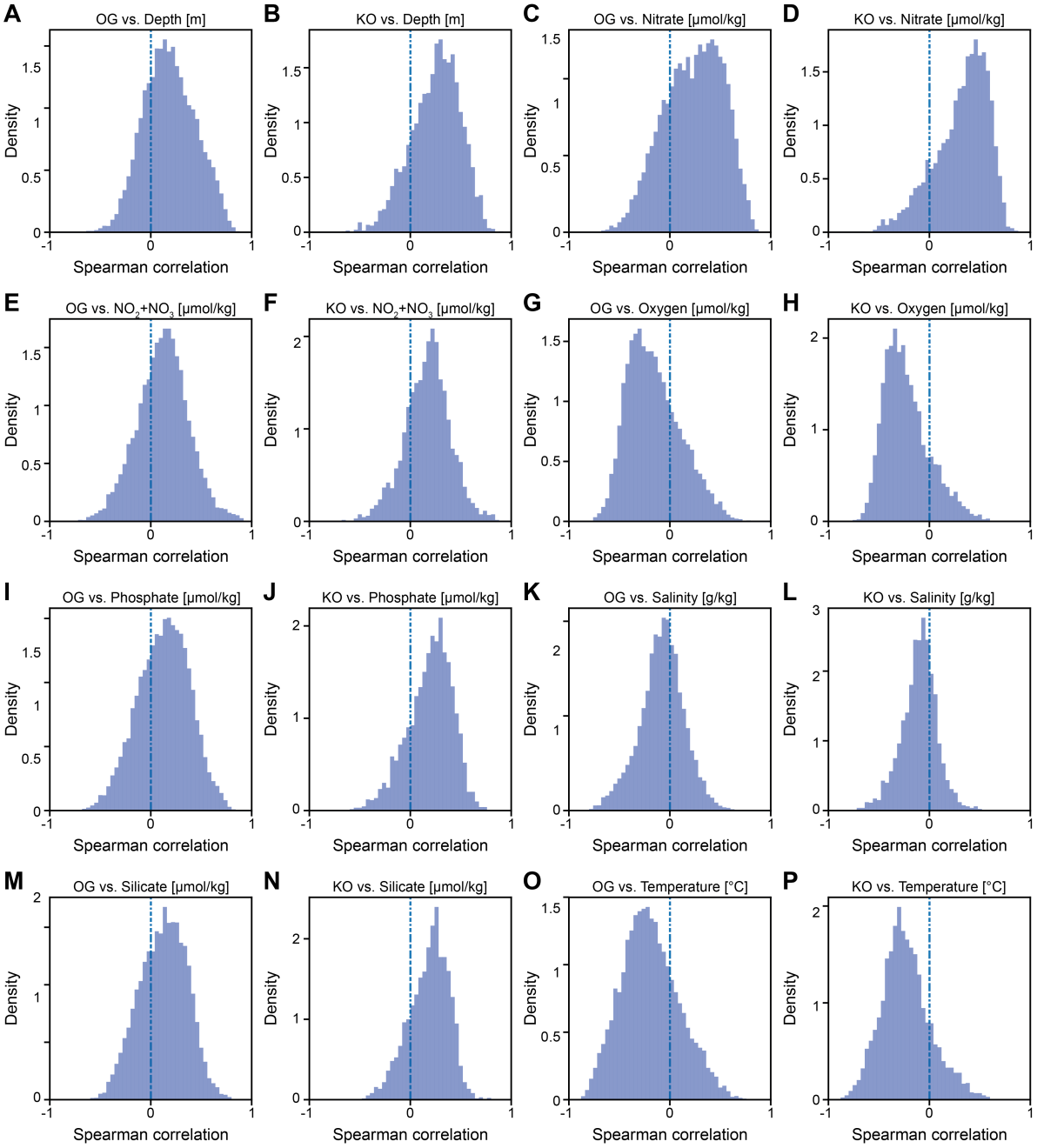



**Figure S2.** (A-P) Histograms of Spearman correlations between $\pi_{within}$ (red) or pN/pS rates (blue) and environmental variables, for both KEGG KOs and eggNOG OGs. Panels A, C, E, G, I, K, M and O depict correlations between OG calculated parameters and depth, nitrate, nitrite and nitrite, oxygen, phosphate, salinity, silicate and temperature, respectively. Panels B, D, F, H, J, I, N and P depict correlations between KO calculated parameters and depth, nitrate, nitrite and nitrite, oxygen, phosphate, salinity, silicate and temperature, respectively.

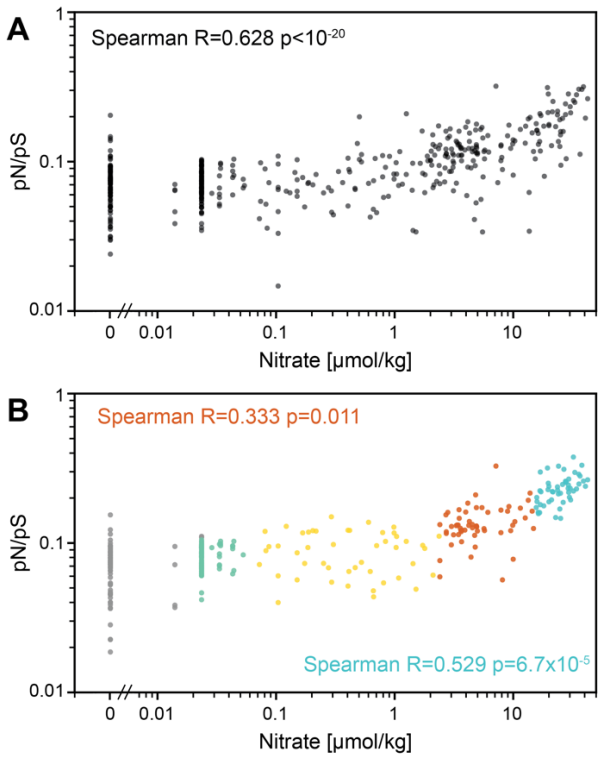



**Figure S3**. (A,B) Scatter plot of the association of pN/pS genes from genus *Synechococcus* with environmental concentrations of nitrate (A) for all synechococcus genes and (B) for genes present in over 50% in nitrate strata.

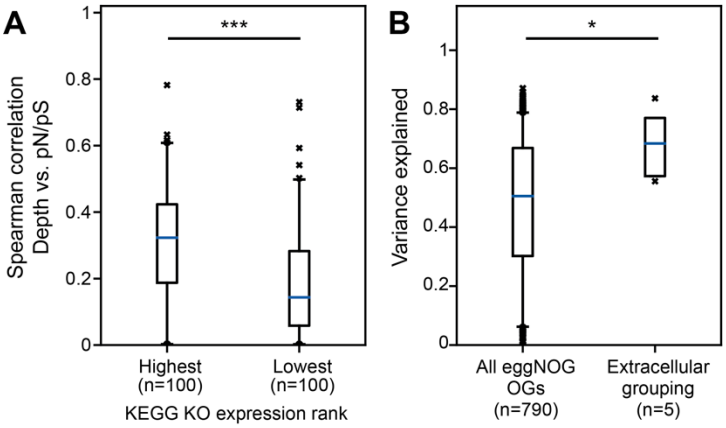



**Figure S4**. (A, B) Box plots (line, median; box, IQR; whiskers, 5th and 95th percentiles) of (A) Spearman correlation coefficients between depth and pN/pS in highly expressed (left) and lowly expressed (right) KEGG KOs. (B) Spearman correlation coefficients between depth and pN/pS in highly expressed (left) and lowly expressed (right) KEGG KOs. Variance explained by the environment in extracellular gene groups versus all eggNOG OGs (Methods). *, $P<0.05$; ***, $P<10^{-5}$.

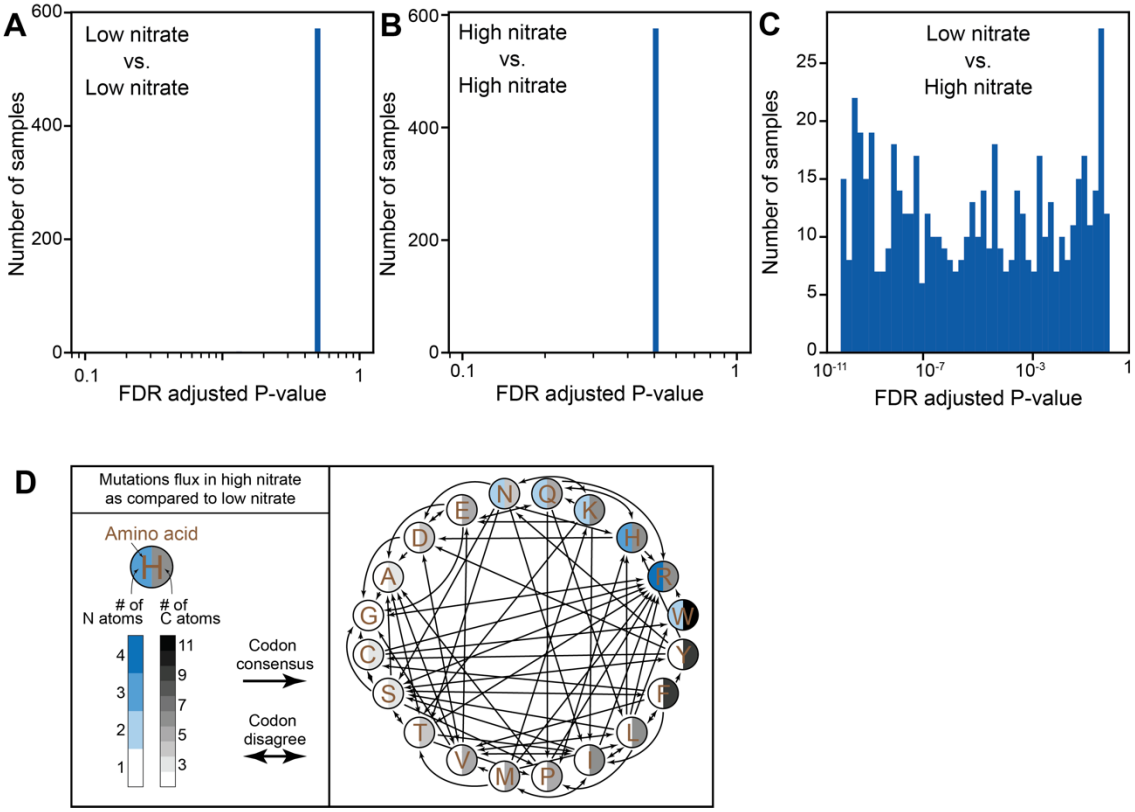



**Figure S5**. (A-C) Histograms of the distribution of P-values of codon-to-codon mutations compared between (A) 40 low-nitrate samples and 40 other low-nitrate samples selected randomly out of the 80 lowest-nitrate samples; (B) 40 high-nitrate samples and 40 other high-nitrate samples selected randomly 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 mutations common in high versus low environmental nitrate concentrations. Two edged arrows mark amino-acids in which some codon mutations were more common in high-nitrate but the opposite codon mutation was more common in low-nitrate.

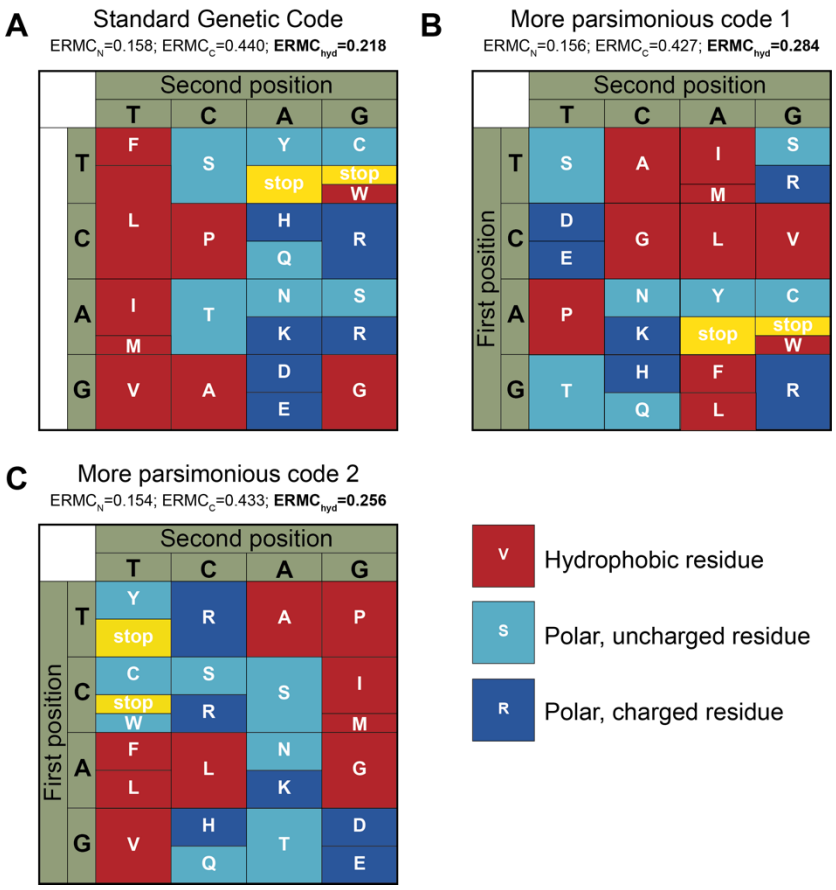



**Figure S6.** (A-C) Hydrophobic or hydrophilic properties of different amino acids depicted across their positions in (A) the standard genetic code, (B+C) two permutations of the standard genetic code that were more conservative in terms of nitrogen and carbon ERMC.

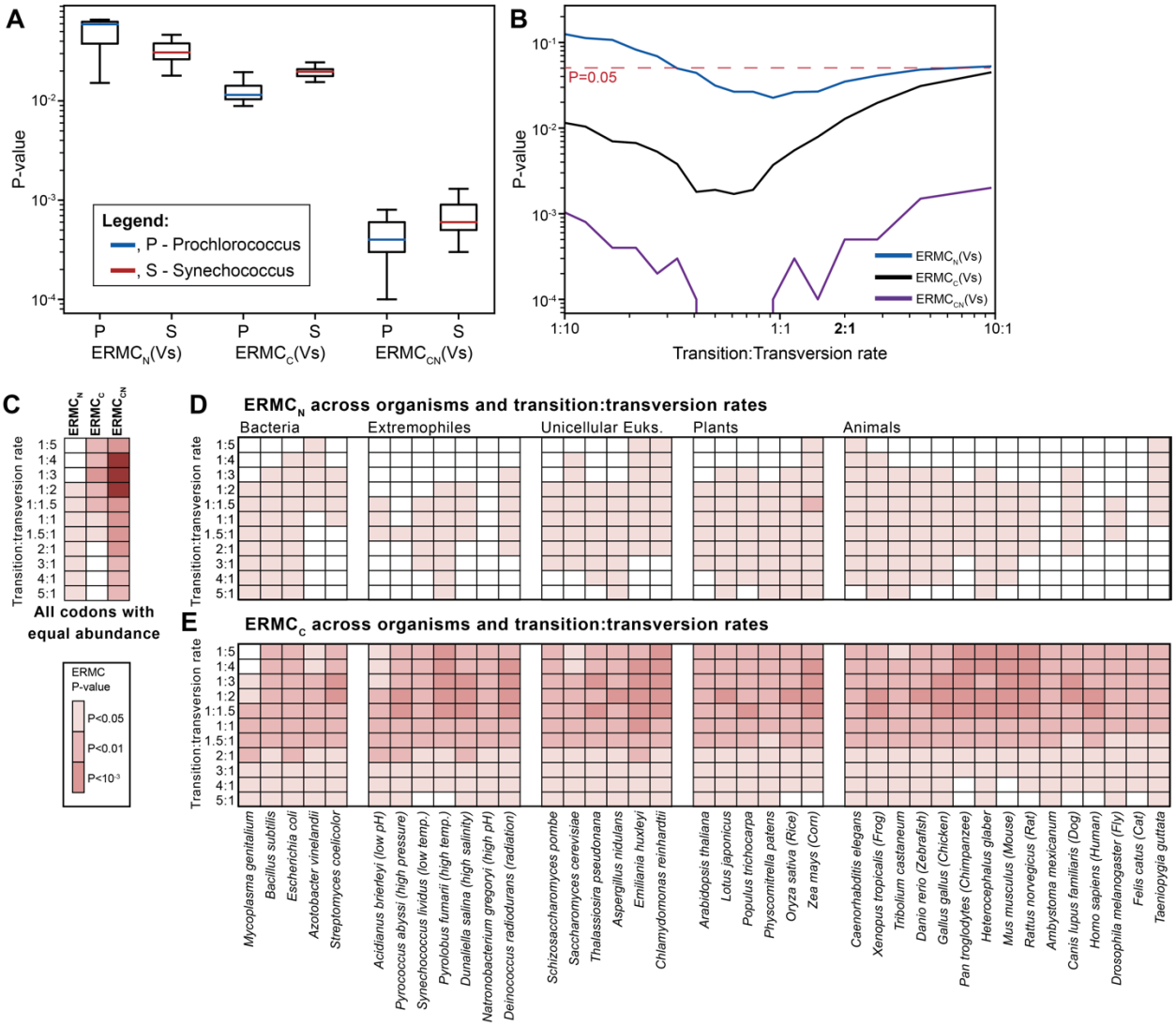



**Figure S7.** (A) Box plots (line, median; box, IQR; whiskers, 5th and 95th percentiles) of P-values for the ERMC of the standard genetic code for nitrogen (left), carbon (center) and both (right) across 187 *Prochlorococcus* (P, blue) and *Synechococcus* (S, red) strains. (B) P-values for the ERMC of the standard genetic code for nitrogen (blue), carbon (black) and both (purple) across a wide range of transition:transversion rates, calculated using combined codon abundance of 187 *Prochlorococcus* and *Synechococcus* strains. (C) Heat map of ERMC P-values for nitrogen, carbon and both, for a theoretical case in which all codons are of the same abundance. (D,E) Same as Fig. 3C for $ERMC_N$ (D) and $ERMC_C$ (E) P-values.

| 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.