Preferred synonymous codons are translated more accurately: Proteomic evidence, among-species variation, and mechanistic basis

Mengyi Sun and Jianzhi Zhang

Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109, USA

*Correspondence to
Jianzhi Zhang
Department of Ecology and Evolutionary Biology
University of Michigan
4018 Biological Sciences Building
1105 North University Avenue
Ann Arbor, MI 48109, USA
Phone: 734-763-0527
Email: jianzhi@umich.edu

Running title: Preferred codons are decoded more accurately

Keywords: *Escherichia coli*, codon usage, tRNA, translational accuracy, translational efficiency, evolution

ABSTRACT

A commonly stated cause of the widespread phenomenon of unequal uses of synonymous codons is their differential translational accuracies. However, this long-standing translational accuracy hypothesis (TAH) of codon usage bias has had no direct evidence beyond anecdotes. Analyzing proteomic data from *Escherichia coli*, we observe higher translational accuracies of more frequently used synonymous codons, offering direct, global evidence for the TAH. The experimentally measured codon-specific translational accuracies validate a sequence-based proxy; this proxy provides support for the TAH from the vast majority of over 1000 taxa surveyed in all domains of life. We find that the relative translational accuracies of synonymous codons vary substantially among taxa and are strongly correlated with the amounts of cognate tRNAs relative to those of near-cognate tRNAs. These and other observations suggest a model in which selections for translational efficiency and accuracy drive codon usage bias and its coevolution with the tRNA pool.

62 **INTRODUCTION**

63 Eighteen of the 20 amino acids are each encoded by more than one codon, but the
64 synonymous codons are usually unequally used in a genome (*1, 2*). Among the synonymous
65 codons of an amino acid, those used more often than the average are referred to as preferred
66 codons while the rest unpreferred. This phenomenon of codon usage bias (CUB), initially
67 discovered over four decades ago from the first few determined gene sequences (*3-6*), is a result
68 of the joint forces of mutation, genetic drift, and natural selection, but the specific selective
69 agents have not been fully deciphered (*1, 2*). One long-standing hypothesis known as the
70 translational accuracy hypothesis (TAH) asserts that different synonymous codons are translated
71 with different accuracies and that CUB results at least in part from natural selection for
72 translational accuracy (*7*). Indeed, the importance of accurate protein translation cannot be
73 overstated, because mistranslation may lead to the loss of normal protein functions and gain of
74 cellular toxicity (*8*) and cause severe diseases including cancer and neurodegenerative diseases
75 (*9*). In fact, several cellular mechanisms are known to ensure the overall fidelity of protein
76 synthesis. For example, conformational changes of the ribosome decoding center can be more
77 efficiently induced by cognate codon-anticodon interactions than near-cognate codon-anticodon
78 interaction (*10*), allowing discrimination against incorrect decoding. Additionally, the accuracy
79 of many steps in translation, such as tRNA aminoacylation (*10*) and codon-anticodon matching,
80 is enhanced by the energy-consuming kinetic proofreading (*11*). Notwithstanding, even if
81 synonymous codons differ in translational accuracy, relatively accurate synonymous codons may
82 not be preferentially used. This is because synonymous codons differ in other properties such as
83 the translational elongation speed (*12, 13*); selection related to these other features(*14*) could
84 triumph selection for translational accuracy.

85 Several groups have attempted to test the TAH of CUB. In particular, Akashi (*7*)
86 developed an indirect test based on the idea that the benefit of using relatively accurate codons
87 should be greater at evolutionarily conserved amino acid sites than unconserved sites of the same
88 protein; consequently, a higher usage of preferred codons at conserved sites than at unconserved
89 sites supports the TAH. While Akashi's test is positive for several model organisms investigated
90 (*7, 15, 16*), this test does not directly compare translational accuracies of synonymous codons so
91 cannot completely exclude alternative explanations (*7, 17*). In an early study, Precup and Parker
92 used site-directed mutagenesis followed by peptide sequencing to show that AAU, an

93 unpreferred codon of Asn, is misread as Lys 4-9 times more often than is AAC, a preferred codon
94 of Asn, at a particular position of the coat protein gene of the bacteriophage MS2 under Asn
95 starvation (*18*). Similarly, Kramer and Farabaugh observed that AAU has a significantly higher
96 rate of mistranslation to Lys than AAC at a particular position of a reporter gene in *Escherichia*
97 *coli* (*19*). Nonetheless, Kramer and Farabaugh also observed that the unpreferred Arg codons of
98 CGA and CGG and the preferred Arg codons of CGU and CGC exhibited similar rates of
99 mistranslation to Lys (*19*). While the above experiments directly tested the TAH, they were each
100 based on the investigation of one amino acid site of one protein, so its genome-wide generality is
101 unknown. As such, a direct and global test of the TAH is needed.
102 Capitalizing on a proteome-wide probe of mistranslation in *E. coli* (*20*), we here provide
103 direct evidence that preferred codons are generally translated more accurately than unpreferred
104 codons. We then use the *E. coli* data to validate a sequence-based proxy for relative translational
105 accuracies of synonymous codons. Using this proxy, we show that the TAH of CUB is supported
106 in the vast majority of over 1000 diverse taxa surveyed, but that the relative translational
107 accuracies of synonymous codons vary substantially among taxa. We find that the relative
108 translational accuracy of a synonymous codon is strongly correlated with its cognate tRNA
109 abundance relative to near-cognate tRNA abundance. These and other results suggest a model in
110 which selections for translational efficiency and accuracy drive the CUB and its coevolution with
111 the tRNA pool.
112
113 **RESULTS**
114 **Preferred codons are more accurately decoded**
115 A direct test of the TAH of CUB requires comparing the mistranslation rate among
116 synonymous codons. Using mass spectrometry, Mordret *et al.* quantified mistranslations at
117 individual sites of the *E. coli* proteome (*20*). After removing sites and codons where
118 mistranslation rates cannot be quantified due to technical reasons (see Materials and Methods),
119 we grouped mistranslation events according to the identities of their original codons. We then
120 computed the absolute mistranslation rate of a codon as the ratio of the total intensity of
121 mistranslated peptides to that of all peptides mapped to the codon. Finally, we computed the
122 relative mistranslation rate (*RMR*) of a codon by dividing its absolute mistranslation rate by the
123 mean absolute mistranslation rate of all codons coding for the same amino acid. *RMR* >1 means

124 that the codon has a higher mistranslation rate than the average among all codons for the same
125 amino acid, whereas *RMR* <1 means the opposite. Codon usage was assessed by the relative
126 synonymous codon usage (*RSCU*). The *RSCU* of a codon equals its frequency in the genome
127 relative to the average frequency of all codons for the same amino acid (*21*). A codon with
128 *RSCU* >1 is preferred while a codon with *RSCU* <1 is unprefered.
129 We were able to estimate the *RMR* for 27 codons of nine amino acids (**Fig. 1a**). Except
130 for Gly, the most preferred synonymous codon of an amino acid shows *RMR* <1, providing a
131 significant support for the TAH ($P = 0.020$, one-tailed binomial test). Similarly, except for Gly
132 and Val, the least prevalent synonymous codon of an amino acid shows *RMR* >1 ($P = 0.090$, one-
133 tailed binomial test). Because both *RSCU* and *RMR* of a codon are relative to the mean of all
134 codons for the same amino acid, they can be compared among codons of different amino acids.
135 Indeed, a strong negative correlation was observed between *RSCU* and *RMR* among the 27
136 codons (Pearson's $r = -0.56$, $P < 0.001$, permutation test; Spearman's $\rho = -0.49$, $P = 0.005$,
137 permutation test; **Fig. 1b**). Together, these findings from the proteomic data of *E. coli*
138 demonstrate that preferred codons tend to have lower mistranslation rates, supporting the TAH of
139 CUB.
140
141 **Relative translational accuracies of synonymous codons vary across taxa**
142 How do certain synonymous codons achieve higher translational accuracies than others?
143 There are two general scenarios. In the first scenario, referred hereinafter as the constant
144 accuracy hypothesis, the translational accuracy is intrinsically higher for a synonymous codon
145 than another because of their different chemical nature (*22*). Consequently, the relative
146 translational accuracies of synonymous codons should be more or less the same in different
147 species. For instance, given that AAA (Lys) is more accurate than AAG (Lys) in *E. coli* (**Fig.
148 1a**), we expect the same trend in the vast majority if not all species. Alternatively, relative
149 translational accuracies of synonymous codons may be greatly influenced by species-specific
150 factors such as the tRNA pool. Under this scenario, referred to as the variable accuracy
151 hypothesis hereinafter, the relative accuracies of synonymous codons vary among species. That
152 is, AAA is more accurate than AAG in many species but the opposite is true in many other
153 species.

154 Measuring the relative translational accuracies of synonymous codons in a large number
155 of species will allow differentiating between the above two hypotheses, which will in turn help
156 understand the mechanism underlying the translational accuracy differences among synonymous
157 codons. Because codon-specific, proteome-based translational accuracies have not been
158 measured beyond *E. coli*, we resort to a sequence-based proxy referred to as the odds ratio (*OR*)
159 that originated from Akashi's test (*7*). Specifically, the *OR* of synonymous codon X that encodes
160 amino acid Y in a gene is the number of times that X is used at invariant Y sites relative to the
161 number of times that X is not used at invariant Y sites, divided by the number of times that X is
162 used at variant Y sites relative to the number of times that X is not used at variant Y sites (**Fig.**
163 **2a**). Here, invariant and variant Y sites refer to Y sites in the focal species whose counterparts in
164 the ortholog from a related species have Y and non-Y, respectively. The *OR* values computed
165 from individual genes can be combined to yield a single *OR* using the Mantel-Haenszel
166 procedure (see Materials and Methods). While *OR* was originally developed for preferred
167 codons, it can be computed for any codon of the 18 amino acids that have multiple synonymous
168 codons (*14*). Based on Akashi's test, *OR* has been used as a proxy for the relative translational
169 accuracy of a codon (*14*). To verify the relationship between *OR* and relative translational
170 accuracy, we computed *OR* values by aligning *E. coli* genes with their *Salmonella enterica*
171 orthologs. Indeed, for the 27 codons with *RMR* estimates, *OR* and *RMR* are strongly negatively
172 correlated (r = -0.63, P = 0.001; ρ = -0.43, P = 0.01; **Fig. 2b**), confirming that the *OR* of a codon
173 is a valid proxy for its relative translational accuracy.
174 To examine whether the relative translational accuracies of synonymous codons vary
175 among species, we took advantage of a recently built phylogenetic tree of 10,575 microbial taxa
176 (*23*). Because most taxa (9,906) in the tree are from the domain Bacteria, we first focused our
177 analysis on Bacteria. We picked all 1,197 pairs of sister bacterial taxa from the tree and aligned
178 their orthologous genes (see Materials and Methods). We randomly assigned one taxon in each
179 pair as the focal taxon and computed *OR* for each codon as described above. We found a
180 positive correlation between *RSCU* and *OR* across codons in 95% of the taxa examined (**Fig. 2c**),
181 demonstrating an overwhelming support for the TAH of CUB in Bacteria.
182 We computed ln(*OR*) to make its distribution relatively symmetric to aid visualization,
183 and examined as an example ln(*OR*) for codon CAT (His) in each of the focal taxa arranged
184 according to the bacterial tree (one taxon per order is presented in **Fig. 2d**). We found ln(*OR*) to

185 vary greatly from negative values to positive values, with a high density near 0 (**Fig. 2e**).
186 Furthermore, the extreme values of ln(*OR*) (bright red and bright green in **Fig. 2d**) are scattered
187 across the tree rather than concentrated in a few clades, suggesting that the relative translational
188 accuracy of CAT has changed substantially and frequently in evolution. The across-taxon
189 variation of *OR* indicates that CAT is the relatively inaccurate one of the two synonymous
190 codons of His in many taxa (red in **Fig. 2d**) but the relatively accurate one in many other taxa
191 (green), supporting the variable accuracy hypothesis. From **Fig. 2e**, which shows the 18 amino
192 acids each with multiple codons, it is clear that the pattern observed for CAT applies to all
193 codons. Furthermore, every codon has *OR* >1 in at least 8.9% of the taxa examined (**Fig. S1a**).
194 These results thus support the variable accuracy hypothesis for all synonymous codons. The
195 above observations of *OR* variation among taxa are not primarily caused by sampling error,
196 because a similar pattern was detected when we analyzed a subset of taxa for each amino acid
197 where the number of occurrences of each synonymous codon considered in *OR* estimation is at
198 least 1000 per taxon (**Fig. S1b**). They are not mainly caused by genetic drift either, because a
199 similar pattern was found when we analyzed a subset of taxa with strong signals of selection for
200 translational accuracy (correlation between *RSCU* and *OR* exceeding 0.5) (**Fig. S1c**). It is worth
201 pointing out that, despite the general support for the variable accuracy hypothesis, for a minority
202 of codons such as ATA (Ile), AGA (Arg), and AGG (Arg), the distribution of ln(*OR*) is strongly
203 skewed toward negative values (**Fig. 2e**), suggesting that their relative translational accuracies
204 are somewhat constrained although not invariable in evolution.
205 To investigate if the above observations from Bacteria are generalizable to the other two
206 domains of life, we first expanded our analysis to Archaea represented in the large phylogeny
207 mentioned (*23*). We found that the correlation between *RSCU* and *OR* is positive in 90% of taxa
208 examined and that ln(*OR*) varies greatly across taxa for each codon (**Fig. S2**), further supporting
209 the TAH and the variable accuracy hypothesis. For Eukaryota, we analyzed five commonly used
210 model organisms: human, mouse, worm, fly, and budding yeast (see Materials and Methods). In
211 each of these species, the correlation between *RSCU* and *OR* is significantly positive (**Table S1**),
212 supporting the TAH. Except for the two mammals, which are closely related, the *OR*s estimated
213 from one species are not well correlated with those estimated from another species (**Fig. S3**).
214 Furthermore, the correlation in *OR* generally declines with the divergence time between the two
215 species (**Fig. S3**), consistent with the variable accuracy hypothesis. Taken together, our results

216 show that the TAH is generally supported in all domains of life but the relative translational
217 accuracies of synonymous codons vary across taxa.
218
219 **Mechanistic basis of among-codon and across-taxon variations of translational accuracies**
220 The empirical support for the variable accuracy hypothesis strongly suggests that the
221 determinants of the *RMR*s of synonymous codons vary among species. In the aforementioned
222 Kramer-Farabaugh study (*19*), the authors found that artificially increasing the expression level
223 of the cognate tRNA for Arg codons AGA and AGG reduces their mistranslations to Lys, so
224 proposed that the competition between cognate and near-cognate tRNAs determines the
225 mistranslation rate of a codon. Here, the cognate tRNA is the tRNA whose anticodon pairs with
226 the codon correctly (allowing wobble pairing), whereas the near-cognate tRNA corresponds to a
227 different amino acid and has an anticodon that mismatches the codon at one position. Consistent
228 with the above proposal, Mordret *et al.* (*20*) inferred that most of the mistranslation events in *E.*
229 *coli* arose from mispairing between codons and near-cognate tRNAs. They further noted that,
230 for certain types of mistranslation, there is a negative correlation across growth phases between
231 the mistranslation rate and the ratio ($R_{c/nc}$) in abundance between cognate and near-cognate
232 tRNAs, although the correlation was rarely statistically significant (*20*). Based on these past
233 observations, we hypothesize that the relative translational accuracy of a synonymous codon
234 increases with its relative $R_{c/nc}$, or $RR_{c/nc}$, which is $R_{c/nc}$ divided by the mean $R_{c/nc}$ of all codons
235 coding for the same amino acid (see Materials and Methods). We further hypothesize that,
236 because the tRNA pool varies substantially among species (*24*), the among-species variation of
237 relative translational accuracies arises from the among-species variation in $RR_{c/nc}$.
238 To test the above hypotheses, we computed $RR_{c/nc}$ for each codon using published tRNA
239 expression levels in *E. coli* (*20*). Indeed, we observed a significant negative correlation between
240 $RR_{c/nc}$ and *RMR* (r = -0.47, P = 0.009; ρ = -0.53, P = 0.005; **Fig. 3a**) and a significant positive
241 correlation between $RR_{c/nc}$ and *OR* (r = 0.49, P = 0.07; ρ = 0.74, P = 0.00001; **Fig. 3b**) across
242 codons, supporting the hypothesis that the relative ratio of cognate to near-cognate tRNA
243 abundances is a major determinant of a codon's relative translational accuracy in *E. coli*. Note
244 that the relative cognate tRNA abundance alone is not significantly correlated with *RMR* (r = -
245 0.22, P = 0.2; ρ = -0.04, P = 0.4; **Fig. S4a**), supporting the role of competition between cognate
246 and near-cognate tRNAs in determining *RMR*. As previously reported (*14*), the relative cognate

247 tRNA level is highly correlated with *RSCU* ($r = 0.61$, $P = 0.03$; $\rho = 0.48$, $P = 0.02$; **Fig. S4b**),
248 which is likely a result of selection for high translational efficiency (i.e., more codons translated
249 per unit time per cell) because balanced codon usage relative to cognate tRNA concentrations
250 maximizes translational efficiency (*14*).

251 We next investigated whether the above finding in *E. coli* applies to other bacterial taxa
252 surveyed in Fig. 2. Because tRNA expression levels are unknown for the vast majority of these
253 taxa, we used the gene copy number of each tRNA species as a proxy for the total expression
254 level of the tRNA species (*25*). Indeed, *E. coli* $RR_{c/nc}$ computed from tRNA gene copy numbers
255 is highly correlated with that computed from tRNA expression levels ($r = 0.77$, $P = 1.64\times10^{-6}$; ρ
256 $= 0.90$, $P = 1.36\times10^{-10}$). Furthermore, *E. coli* $RR_{c/nc}$ computed from tRNA gene copy numbers is
257 significantly correlated with *RMR* (**Fig. 3c**), confirming the validity of using this proxy. We
258 obtained the tRNA gene annotations for 1094 of the 1197 focal bacterial taxa examined in Fig. 2.
259 However, in many of these taxa, there is little tRNA gene redundancy or variation in cognate
260 tRNA gene copy number among synonymous codons despite considerable CUB; in these taxa,
261 the tRNA gene copy number is unlikely a good proxy for tRNA abunadnce (*26*). Because the
262 tRNA gene copy number is a good proxy for tRNA abundance in *E. coli*, which has 85 tRNA
263 genes, we decided to filter out taxa with fewer than 81 tRNA genes to strike a balance between
264 the noise level and number of taxa in our analysis. This filtering left us with 59 taxa, in each of
265 which we correlated the *OR* of a codon with its $RR_{c/nc}$ computed from tRNA gene copy numbers.
266 The vast majority (92%) of the taxa show a positive correlation (**Fig. 3d**), supporting the
267 generality of our hypothesis on the role of $RR_{c/nc}$ in determining the relative translational
268 accuracy of a codon in Bacteria.

269 To investigate whether the above finding is generalizable to other domains of life, we
270 analyzed tRNA genes in Archaea taxa and Eukaryotic model organisms. Unfortunately, no
271 Archaea taxa examined have more than 80 tRNA genes. For each of the five eukaryotes (human,
272 mouse, fly, worm, and yeast), the correlation between *OR* and $RR_{c/nc}$ computed from tRNA gene
273 copy numbers is significantly positive for linear or rank correlation (**Table S2**). Together, our
274 findings strongly support that, in the diverse taxa surveyed, the ratio of cognate tRNA abundance
275 to near-cognate tRNA abundance is generally a major determinant of the relative translational
276 accuracy of a codon. Hence, the variation of the tRNA pool among species can explain the
277 across-species variation of the relative translational accuracies of synonymous codons.

DISCUSSION

Analyzing published proteomic data from *E. coli*, we provided direct, global evidence that preferred synonymous codons are generally decoded more accurately than unpreferred codons. We found that relative translational accuracies of synonymous codons vary substantially among species, supporting the variable accuracy hypothesis. We obtained strong evidence that the ratio of cognate tRNA abundance to near-cognate tRNA abundance is a major determinant of a codon's relative translational accuracy. Hence, the variable accuracies observed are mechanistically explained by the variation of the tRNA pool across species. These findings, together with the previous report on the selection for translational efficiency (*14*), suggest a model in which the tRNA pool and codon usage coevolve to improve both translational efficiency and accuracy (**Fig. 4a**). Specifically, mutation and drift can alter both codon frequencies and tRNA concentrations. The cellular translational efficiency is maximized when (transcriptomic) codon frequencies equal relative cognate tRNA concentrations (*14*), whereas the translational accuracy of a codon is maximized when the ratio of its cognate tRNA concentration to near-cognate tRNA concentration is maximized. Under this model, selections for translational efficiency and translational accuracy are related but not perfectly aligned, which could introduce tradeoffs between translational efficiency and accuracy (*27*). Indeed, our simulation of a simple genetic system with two amino acids, each encoded by two synonymous codons (**Fig. 4b**), found that imposing a selection for translational accuracy can lower translational efficiency (**Fig. 4c**).

Interestingly, our results imply that, even in the absence of selection for translational accuracy, the positive correlation between synonymous codon frequency and cognate tRNA concentration resulting from selection for translational efficiency (*14*) will likely render the cognate tRNA concentration relative to near-cognate tRNA concentration higher for more frequently used synonymous codons. Consequently, the positive correlation between the relative codon frequency and relative translational accuracy may arise in the absence of selection for translational accuracy. In fact, in *E. coli*, for 16 of the 18 amino acids with multiple synonymous codons, the codon with the highest cognate tRNA concentration has the highest $RR_{c/nc}$. Upon shuffling the expression levels among tRNA species, we found that, for over one half of the 18 amino acids, the codon with the highest cognate tRNA concentration has the highest $RR_{c/nc}$. This was true in each of 1000 shufflings. Nevertheless, in only 6 of these 1000 shufflings did all 18

309 amino acids exhibit the above feature. Thus, a high but non-perfect concordance between
310 translational efficiency and accuracy is expected. Therefore, strictly speaking, the correlation in
311 Fig. 1b by itself does not prove selection for translational accuracy. However, this correlation, in
312 conjunction with the correlation between *RSCU* and *OR*, demonstrates that evolutionarily
313 conserved sites tend to use preferred synonymous codons, which tend to be relatively accurately
314 translated, hence proving the role of selection for translational accuracy in causing CUB, or the
315 TAH. How codon usage and the tRNA pool evolve under the joint forces of selections for
316 translational efficiency and accuracy in addition to mutation and drift is quite complex. For
317 instance, because any tRNA is simultaneously a cognate tRNA for one or more codons and a
318 near-cognate tRNA for some other codons, increasing the translational accuracy of a particular
319 codon might be at the expense of the translational accuracy of another codon. Indeed, a previous
320 study showed that artificially increasing the cognate tRNA expression levels for arginine codons
321 can result in proteotoxic stress (*28*). This subtle tradeoff could cause non-independent uses of
322 codons of different amino acids. This was indeed observed in the simulation aforementioned
323 (**Fig. 4d**). Future modeling work with realistic parameters might shed more light on this issue.
324 In addition to impacting translational efficiency and accuracy, synonymous mutations also affect
325 mRNA folding (*29*), mRNA stability (*30*), mRNA concentration (*30-32*), pre-mRNA splicing
326 (*33*), and co-translational protein folding (*34, 35*), so additional selections may shape CUB and
327 its evolution.
328 Our study has several caveats. First, in our calculation of a codon's mistranslation rate,
329 we lumped all mistranslations of the codon regardless of the erroneous amino acid it is translated
330 to. Because different mistranslations of the same codon likely have differential fitness costs and
331 because selection for translational accuracy likely minimizes the total fitness reduction caused by
332 mistranslation instead of the mistranslation rate *per se*, properly weighting different
333 mistranslations in *RMR* calculation will likely strengthen its correlation with *RSCU*. Second,
334 when calculating the ratio of cognate tRNA concentration to near-cognate tRNA concentration,
335 we did not consider the difference in interaction strength between different codons and
336 anticodons (*36*). Future research that takes into account this interaction under physiological
337 conditions may significantly improve the signal in the correlation analysis of Fig. 3. Third, our
338 analysis in Fig. 3d was limited to taxa with >80 tRNA genes. Future research using tRNA
339 expression levels (*26*) when they become available can confirm if the same pattern holds for taxa

340 with fewer tRNA genes. Finally, due to data limitation, we did not consider tRNA expression
341 variations across environments, cell cycle stages, or tissues (*37*). In the future, it would be
342 interesting to study how such variations simultaneously impact translational efficiency and
343 accuracy.
344 Our results might help design organisms with expanded code tables (*38*). Expanding the
345 code table is realized by introducing unnatural tRNAs that are charged with non-canonical amino
346 acids. The introduction of these tRNAs often leads to fitness defects due to mistranslation of
347 normal codons (*39*). Our research suggests that one way to alleviate the proteotoxic stress is to
348 identify potential near-cognate codons that could be mistranslated by the unnatural tRNA and
349 adjust the natural tRNA pool to minimize the impact.
350

351 **MATERIALS AND METHODS**

352 **Estimating relative mistranslation rates of synonymous codons from *E. coli* proteomic data**

353 The proteomic data analyzed came from Table S1 in Modret *et al.* (*20*). The authors
354 separately measured mistranslation events from high-solubility and low-solubility proteins using
355 mass spectrometry, and both groups of events were considered in our analysis. We focused on
356 the data from the wild-type strain BW25113 in the MOPS complete medium because (i) this
357 dataset is the largest among datasets from all strain-medium combinations and (ii) no artificial
358 perturbation such as mutation, drug, or amino acid depletion was applied (*20*). We first removed
359 sites that cannot be traced to a unique original codon. We also filtered out sites showing an
360 intensity of "NaN" for the unmodified (aka base) peptide or mistranslated (aka dependent)
361 peptide. Because different synonymous codons tend to generate different mistranslations by
362 mispairing with different near-cognate tRNAs, if these different mistranslations have different
363 detection probabilities, the comparison between synonymous codons would be unfair.
364 Unfortunately, some mistranslations produce mass shifts indistinguishable from post-
365 translational modifications so cannot be reliably identified through mass spectrometry (*26*),
366 which would produce exactly this situation in some cases. Therefore, we removed amino acids
367 with undetectable mistranslations except for Leu and Ile. We kept these two amino acids because
368 the only undetectable mistranslations for them are Leu to Ile and Ile to Leu, both can be
369 considered benign due to the high physicochemical similarity between Leu and Ile (*40*).
370 Considering the structure of the genetic code table, we found that the underestimation of the

371 mistranslation rate due to the negligence of mistranslations between Leu and Ile is severer for
372 unpreferred than preferred codons, suggesting that the actual strength of evidence for higher
373 mistranslation rates of unpreferred than preferred synonymous codons is stronger than what is
374 shown in Fig. 1. We then computed each codon's absolute mistranslation rate by dividing the
375 total intensity of mistranslated (i.e., dependent) peptides by that of all (i.e., dependent + base)
376 peptides mapped to the codon. We divided each codon's absolute mistranslation rate by the mean
377 absolute mistranslation rate of all codons coding for the same amino acid to obtain the codon's
378 relative mistranslation rate (*RMR*). We removed amino acids without data for all of its
379 synonymous codons because calculating *RMR* requires having data for all synonymous codons of
380 an amino acid. In total, we computed *RMR* for 27 codons of 9 amino acids.
381
382 **Relative synonymous codon usage (*RSCU*), odds ratio (*OR*), and relative ratio of cognate**
383 **tRNA concentration to near-cognate tRNA concentration ($RR_{c/nc}$) for *E. coli***
384 Peptide and cDNA sequences of *E. coli* (genome assembly ASM584v2) and *S. enterica*
385 (genome assembly ASM78381v1) were downloaded from Ensembl Bacteria (*41*). We computed
386 *RSCU* of codon *j* of amino acid *i* from all coding sequences of *E. coli* by $RSCU_{i,j} = \frac{n_i x_{i,j}}{\sum_{j=1}^{n_i} x_{i,j}}$,
387 where n_i is the number of synonymous codons of amino acid *i* and $x_{i,j}$ is the number of codon *j*
388 of amino acid *i* in all coding sequences (*21*). Conventionally, *RSCU* is computed from highly
389 expressed genes (*21*). However, due to the lack of gene expression information from most of the
390 species analyzed, we computed *RSCU* from all genes. This should not qualitatively affect our
391 analysis, because *RSCU* computed from highly expressed genes (e.g., the top 20% of genes) is
392 nearly perfectly correlated with that computed from all genes (e.g., in *E. coli*, $r = 0.96$, $P <$
393 2.2×10^{-16}).
394 To calculate the *OR* of each codon, we first identified one-to-one orthologous proteins
395 between *E. coli* and *S. enterica* using OrthoFinder (*42*). Next, we aligned these one-to-one
396 orthologs using MUSCLE (*43*), separating all amino acid sites into conserved and non-conserved
397 sites. For a focal codon in gene *i*, we tabulated a_i, number of times the focal codon is observed
398 at conserved amino acid sites; b_i, number of times the focal codon is observed at unconserved
399 sites; c_i, total number of times the focal codon's synonymous codons are observed at conserved
400 sites; and d_i, total number of times the focal codon's synonymous codons are observed at

401 unconserved sites. Here, the focal codon's synonymous codons do not include itself. *OR* for
402 gene *i* equals $(a_i d_i)/(b_i c_i)$. Using the Mantel-Haenszel procedure, we combined the odds ratios of
403 the focal codon from individual genes into one odds ratio (*7*) by $OR = \frac{\sum_i \frac{a_i d_i}{(a_i+b_i+c_i+d_i)}}{\sum_i \frac{b_i c_i}{(a_i+b_i+c_i+d_i)}}$.

404 To compute $RR_{c/nc}$ of a codon, we tabulated the cognate tRNAs and near-cognate tRNAs
405 of the codon. Cognate tRNAs are all tRNAs that can pair with the focal codon allowing wobble
406 pairing at the 3rd codon position, while near-cognate tRNAs are tRNAs coded for a different
407 amino acid but can pair with the focal codon with one base-pair mismatch (allowing wobble
408 pairing at the 3rd codon position). We then weighted each tRNA by their average relative
409 expression levels across three growth stages in the MOPS complete media (GEO number:
410 GSE128812). Finally, we normalized the ratio for each codon by the average ratio of all codons
411 coding for the same amino acid.
412
413 **_RSCU_, _OR_, and $RR_{c/nc}$ for other species**
414 *RSCU*, *OR*, and $RR_{c/nc}$ were calculated for non-*E. coli* taxa as for *E. coli*, with the
415 differences noted below. For the non-*E. coli* prokaryotic taxa, we downloaded the phylogenetic
416 tree of 10,575 taxa from the Web of Life (*23*) (https://biocore.github.io/wol/) and identified sister
417 taxa from the tree. Briefly, each pair of sister taxa are two taxa that are the single closest relative
418 to each other in the tree. For each pair of sister taxa, we downloaded from the same web site
419 their protein-coding DNA sequences, protein sequences, and tRNA gene copy number data. For
420 eukaryotic model organisms, we downloaded protein-coding DNA sequences and protein
421 sequences of human (*Homo sapiens*), mouse (*Mus musculus*), fly (*Drosophila melanogaster*),
422 worm (*Caenorhabditis elegans*), and budding yeast (*Saccharomyces cerevisiae*) from the NCBI
423 refseq database (*44*). We further downloaded the protein sequences of *Macaca mulatta* (as a
424 relative of *H. sapiens*), *Rattus norvegicus* (as a relative of *M. musculus*), *Drosophila erecta* (as a
425 relative of *D. melanogaster*), *Caenorhabditis briggsae* (as a relative of *C. elegans*), and
426 *Saccharomyces paradoxus* (as a relative of *S. cerevisiae*) from the NCBI refseq database. The
427 tRNA gene annotations in the five model organisms were downloaded from GtRNAdb (*24*).
428 $RR_{c/nc}$ was computed using tRNA gene copy numbers instead of tRNA expression levels.
429
430 **Statistical analysis**

431 Many of the quantities estimated in our work, such as *RMR*, $RR_{c/nc}$, *RSCU*, and *OR*, are
432 not independent among synonymous codons. To deal with this non-independence in statistical
433 tests, we applied permutation tests. Specifically, in **Fig. 1b**, we generated 1000 permuted
434 samples by shuffling the absolute mistranslation rates among all codons and then re-estimated
435 *RMR* values. We then computed the correlation between *RMR* and *RSCU* in each permuted
436 sample while holding the *RSCU* value of each codon unchanged. *P* equals the fraction of
437 permuted samples with the correlation coefficient more negative than that observed in the
438 original sample. Similarly, when testing the correlation between *RMR* and *OR* (**Fig. 2b**), we
439 shuffled the absolute mistranslation rate among all codons and recomputed *RMR* while holding
440 the *OR* for each codon unchanged. When testing the correlation between *RMR* (or *OR*) and
441 $RR_{c/nc}$ (**Fig. 3**), we shuffled the absolute mistranslation rates among codons and the expression
442 levels (or gene copy numbers) among tRNAs. Finally, when testing the correlation between
443 *RMR* (or *RSCU*) and relative cognate tRNA concentration (**Fig. S4**), we shuffled the absolute
444 mistranslation rate among codons and the expression level among tRNAs.
445 To estimate the standard error (SE) of the *RMR* of each codon, we constructed 1000
446 bootstrap samples by resampling the sites in the original data with replacement. Similarly, we
447 estimated the SE of the *OR* of each codon by constructing 1000 bootstrapped *E. coli* genomes via
448 resampling its genes that have one-to-one orthologs in *S. enterica*.
449
450 **Simulation**
451 To assess the impact of selections for translational accuracy and efficiency on codon
452 usage, we built a toy model with two amino acids: aa_0 and aa_1. Amino acid aa_0 is encoded by
453 synonymous codons 00 and 01 while aa_1 is encoded by synonymous codons 10 and 11 (**Fig. 4b**).
454 Codon-anticodon pairing follows the rule that 0 pairs with 1 and vice versa. The cognate tRNA
455 of a codon has an anticodon that pairs perfectly with the codon, while the near-cognate tRNA has
456 an anticodon that pairs with the codon with exactly one mismatch and carries the other amino
457 acid.
458 We considered a unicellular organism with one gene consisting of *n* codons. We assumed
459 that the mRNA level of the gene does not change in the evolution simulated and that ribosomes
460 are in shortage. We defined the organismal fitness as follows.
461 $Absolute\ fitness = Function - Cost$, where

$$Function = TE \times \sum_{i=1}^{n} f_i \text{ and } Cost = \sum_{i=1}^{n} c_i.$$

Here, f_i and c_i are the function and cost of codon i, respectively. We set $f_i = F_i$ if codon i encodes the pre-specified optimal amino acid at the codon; otherwise, $f_i = 0$. For each i, F_i is a random variable sampled from an exponential distribution with the mean equal to 1 (*45*). Following a previous study (*14*), we set the expected codon selection time per amino acid aa_0 as $t_0 = p_1^2/q_1+p_2^2/q_2$, where p_1 and $p_2 = 1-p_1$ are the fractions of amino acid aa_0 encoded by codon 00 and 01, respectively, and q_1 and $q_2 = 1-q_1$ are the fractions of corresponding cognate tRNAs among all tRNAs of aa_0, respectively. We similarly set the expected codon selection time per amino acid aa_1 and computed the total codon selection time of all codons. Translational efficiency *TE*, which is the number of codons translated per unit time, is the inverse of the total codon selection time. We set $c_i = TE \times C_i$ if codon i does not encode the pre-specified optimal amino acid at the codon; otherwise, $c_i = C_i \times TE \times \frac{1}{RR_{c/nc}}$. When there is no selection for translational accuracy, $C_i = 0$; otherwise, C_i for codon i is a random variable sampled from an exponential distribution with mean equal to 1. Note that C_i and F_i are independent from each other. $RR_{c/nc}$ is computed as described in Results; the inverse of $RR_{c/nc}$ measures the mistranslation rate.

We started the simulation with a coding sequence of 200 nucleotides, coding for 100 amino acids. Each site had a 50% chance to be 0 or 1. For simplicity, we assumed that the initial amino acid sequence is optimal such that the evolution in our simulation is largely about codon usage. For each of the four different tRNAs (with anticodons of 00, 01, 10, and 11, respectively), we sampled the initial copy number from 1, 2, and 3 with equal probabilities.

Next, we simulated the coevolution between the tRNA pool and codon usage following a strong selection, weak mutation regime. We first generate a mutation. With a probability of 0.02, it alters the copy number of a tRNA. In this case, we randomly pick a tRNA species to change its copy number by +1 or -1 with equal probabilities unless the copy number is 1, in which case it is +1. With a probability of 0.98, the mutation is a random point mutation at a randomly picked site of the coding sequence. The fitness of the mutant is then computed following the above fitness definition. The mutation is fixed with a probability of $\frac{1-r^{-1}}{1-r^{-N}}$, where r is the ratio of the absolute fitness of the mutant to that of the wild-type and N is the population size (*46*). The above mutation-selection process was repeated 100,000 rounds in each simulation

492 to reach an equilibrium. For each *N*, we simulated 200 times with and 200 times without
493 selection for translational accuracy.
494

495 **ACKNOWLEDGEMENTS**

496 We thank Dr. Yitzhak Pilpel for consultation on the *E. coli* proteomic data in Mordret *et*
497 *al.* (2019) and Wenfeng Qian, Jian-Rong Yang, and members of the Zhang laboratory for
498 valuable comments. This work was supported by the U.S. National Institutes of Health research
499 grant R35GM139484 to J.Z.
500

501 **REFERENCES**

502 1. R. Hershberg, D. A. Petrov, Selection on codon bias. *Annu Rev Genet* **42**, 287-299
503 (2008).
504 2. J. B. Plotkin, G. Kudla, Synonymous but not the same: the causes and consequences of
505 codon bias. *Nat Rev Genet* **12**, 32-42 (2011).
506 3. T. Ikemura, Correlation between the abundance of Escherichia coli transfer RNAs and the
507 occurrence of the respective codons in its protein genes: a proposal for a synonymous
508 codon choice that is optimal for the E. coli translational system. *J Mol Biol* **151**, 389-409
509 (1981).
510 4. W. Fiers *et al.*, A-protein gene of bacteriophage MS2. *Nature* **256**, 273-278 (1975).
511 5. G. M. Air *et al.*, Gene F of bacteriophage phiX174. Correlation of nucleotide sequences
512 from the DNA and amino acid sequences from the gene product. *J Mol Biol* **107**, 445-458
513 (1976).
514 6. A. Efstratiadis, F. C. Kafatos, T. Maniatis, The primary structure of rabbit beta-globin
515 mRNA as determined from cloned DNA. *Cell* **10**, 571-585 (1977).
516 7. H. Akashi, Synonymous codon usage in Drosophila melanogaster: natural selection and
517 translational accuracy. *Genetics* **136**, 927-935 (1994).
518 8. D. A. Drummond, C. O. Wilke, The evolutionary consequences of erroneous protein
519 synthesis. *Nat Rev Genet* **10**, 715-724 (2009).
520 9. B. Chen, M. Retzlaff, T. Roos, J. Frydman, Cellular strategies of protein quality control.
521 *Cold Spring Harb. Perspect. Biol.* **3**, a004374 (2011).
522 10. M. Ibba, D. Söll, Quality control mechanisms during translation. *Science* **286**, 1893-1897
523 (1999).
524 11. J. J. Hopfield, Kinetic proofreading: a new mechanism for reducing errors in biosynthetic
525 processes requiring high specificity. *Proceedings of the National Academy of Sciences* **71**,
526 4135-4139 (1974).
527 12. D. E. Weinberg *et al.*, Improved Ribosome-Footprint and mRNA Measurements Provide
528 Insights into Dynamics and Regulation of Yeast Translation. *Cell Rep* **14**, 1787-1799
529 (2016).
530 13. J. A. Hussmann, S. Patchett, A. Johnson, S. Sawyer, W. H. Press, Understanding Biases in
531 Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast.
532 *PLoS Genet* **11**, e1005732 (2015).

14. W. Qian, J.-R. Yang, N. M. Pearson, C. Maclean, J. Zhang, Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet* **8**, e1002603 (2012).
15. D. A. Drummond, C. O. Wilke, Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341-352 (2008).
16. N. Stoletzki, A. Eyre-Walker, Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Mol Biol Evol* **24**, 374-381 (2007).
17. P. Shah, M. A. Gilchrist, Effect of correlated tRNA abundances on translation errors and evolution of codon usage bias. *PLoS Genet* **6**, e1001128 (2010).
18. J. Precup, J. Parker, Missense misreading of asparagine codons as a function of codon identity and context. *J Biol Chem* **262**, 11351-11355 (1987).
19. E. B. Kramer, P. J. Farabaugh, The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *RNA* **13**, 87-96 (2007).
20. E. Mordret *et al.*, Systematic detection of amino acid substitutions in proteomes reveals mechanistic basis of ribosome errors and selection for translation fidelity. *Mol. Cell* **75**, 427-441. e425 (2019).
21. P. M. Sharp, T. M. Tuohy, K. R. Mosurski, Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* **14**, 5125-5143 (1986).
22. R. Hershberg, D. A. Petrov, General rules for optimal codon choice. *PLoS Genet* **5**, e1000556 (2009).
23. Q. Zhu *et al.*, Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nature communications* **10**, 1-14 (2019).
24. P. P. Chan, T. M. Lowe, GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* **37**, D93-D97 (2009).
25. T. Tuller *et al.*, An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**, 344-354 (2010).
26. Y. Wei, J. R. Silke, X. Xia, An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria. *Sci. Rep.* **9**, 1-11 (2019).
27. J. R. Yang, X. Chen, J. Zhang, Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS Biol* **12**, e1001910 (2014).
28. A. H. Yona *et al.*, tRNA genes rapidly change in evolution to meet novel translational demands. *Elife* **2**, e01339 (2013).
29. C. Park, X. Chen, J. R. Yang, J. Zhang, Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A* **110**, E678-686 (2013).
30. V. Presnyak *et al.*, Codon optimality is a major determinant of mRNA stability. *Cell* **160**, 1111-1124 (2015).
31. S. Chen *et al.*, Codon-Resolution Analysis Reveals a Direct and Context-Dependent Impact of Individual Synonymous Mutations on mRNA Level. *Mol Biol Evol* **34**, 2944-2958 (2017).
32. Z. Zhou *et al.*, Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc Natl Acad Sci U S A* **113**, E6117-E6125 (2016).
33. J. V. Chamary, J. L. Parmley, L. D. Hurst, Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* **7**, 98-108 (2006).
34. F. Buhr *et al.*, Synonymous Codons Direct Cotranslational Folding toward Different

Protein Conformations. *Mol Cell* **61**, 341-351 (2016).
35. I. M. Walsh, M. A. Bowman, I. F. Soto Santarriaga, A. Rodriguez, P. L. Clark, Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proc Natl Acad Sci U S A* **117**, 3528-3534 (2020).
36. M. d. Reis, R. Savva, L. Wernisch, Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32**, 5036-5044 (2004).
37. H. Gingold *et al.*, A dual program for translation regulation in cellular proliferation and differentiation. *Cell* **158**, 1281-1292 (2014).
38. E. Ros, A. G. Torres, L. R. de Pouplana, Learning from Nature to Expand the Genetic Code. *Trends Biotechnol.*, (2020).
39. J. W. Chin, Expanding and reprogramming the genetic code. *Nature* **550**, 53-60 (2017).
40. S. Henikoff, J. G. Henikoff, Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915-10919 (1992).
41. K. L. Howe *et al.*, Ensembl 2021. *Nucleic Acids Res.* **49**, D884-D891 (2021).
42. D. M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1-14 (2019).
43. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797 (2004).
44. N. A. O'Leary *et al.*, Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-D745 (2016).
45. A. Eyre-Walker, P. D. Keightley, The distribution of fitness effects of new mutations. *Nat. Rev. Genet* **8**, 610-618 (2007).
46. P. A. P. Moran, in *Mathematical proceedings of the cambridge philosophical society*. (Cambridge University Press, 1958), vol. 54, pp. 60-71.

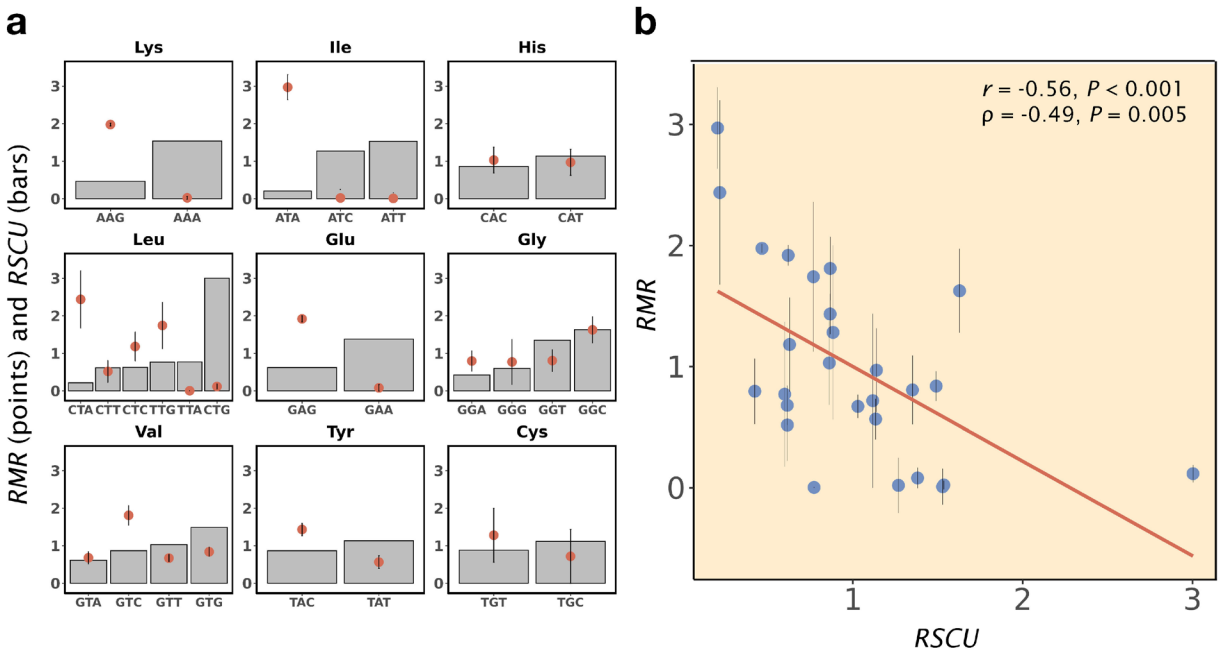

Figure 1. More frequently used synonymous codons tend to be decoded more accurately in *E. coli*. a, Comparison of relative synonymous codon usage (*RSCU*, bars) and relative mistranslation rate (*RMR*, dots) among synonymous codons for nine amino acids with proteome-based *RMR* estimates. **b**, A significant negative correlation between *RSCU* and *RMR* across the 27 codons in panel a. The red line is the linear regression. In both panels, error bars represent one standard error estimated by the bootstrap method. The standard error of *RSCU* estimated by the bootstrap method is negligible due to the large number of each codon in the genome, so is not shown. *P*-values are based on permutation tests.

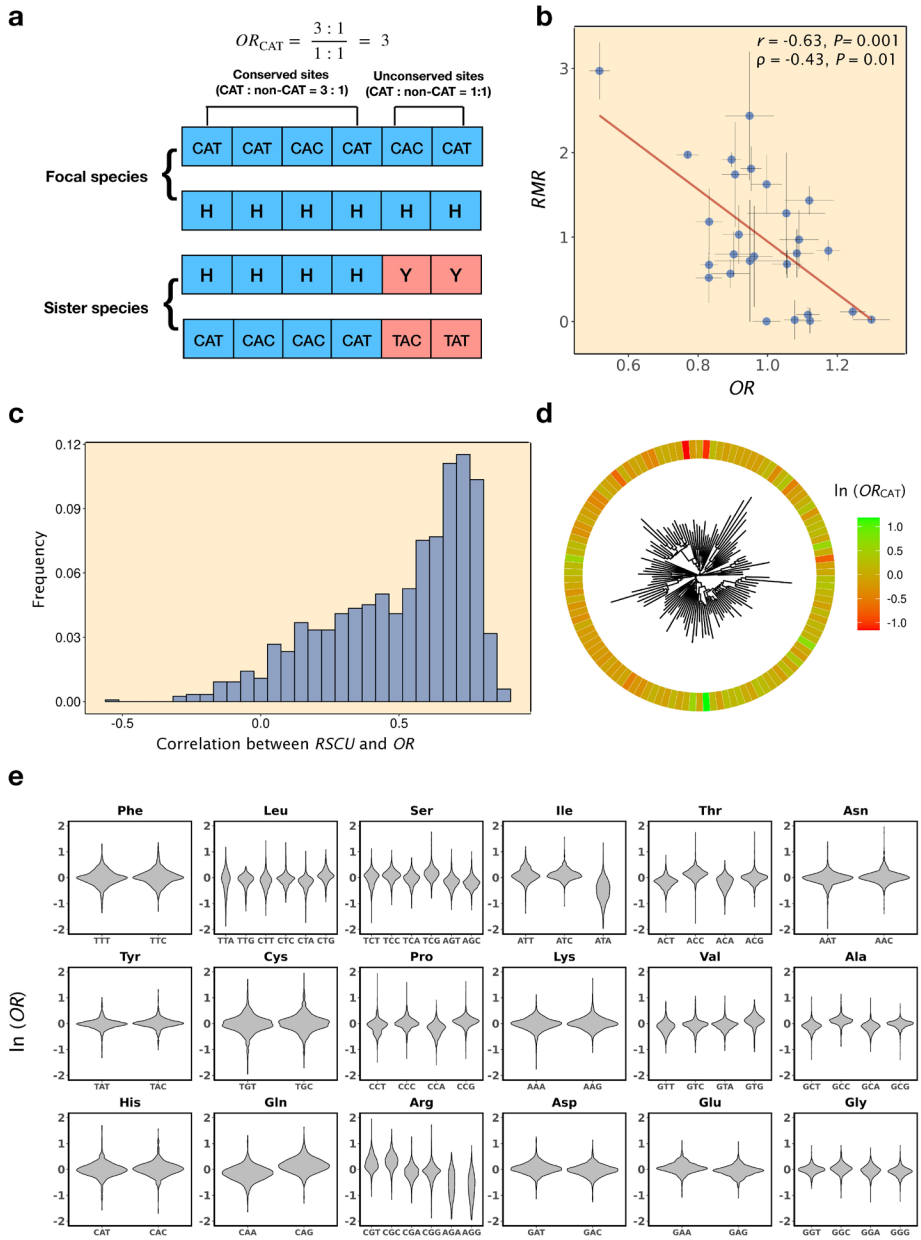

Figure 2. Variation of relative translational accuracies of synonymous codons across taxa. **a**, Diagram explaining the calculation of odds ratio (*OR*) of the codon CAT that serves as a proxy for its relative translational accuracy. Showing here is a hypothetical alignment of orthologous proteins (and the underlying coding sequences) between the focal species and a related species. **b**, *OR* is negatively correlated with *RMR* across codons in *E. coli*. *P*-values are based on permutation tests. The red line is the linear regression. **c**, Frequency distribution of Pearson's correlation between *RSCU* and *OR* in 1197 bacterial taxa. Ninety-five percent of these taxa show positive correlations. **d**, ln(*OR*) of codon CAT for each of 118 bacterial taxa, one per order, arranged according to their phylogeny shown in the middle. **e**, Violin plots showing frequency distributions of ln(*OR*) of individual codons of the 18 amino acids that have multiple synonymous codons across 1197 bacterial taxa.

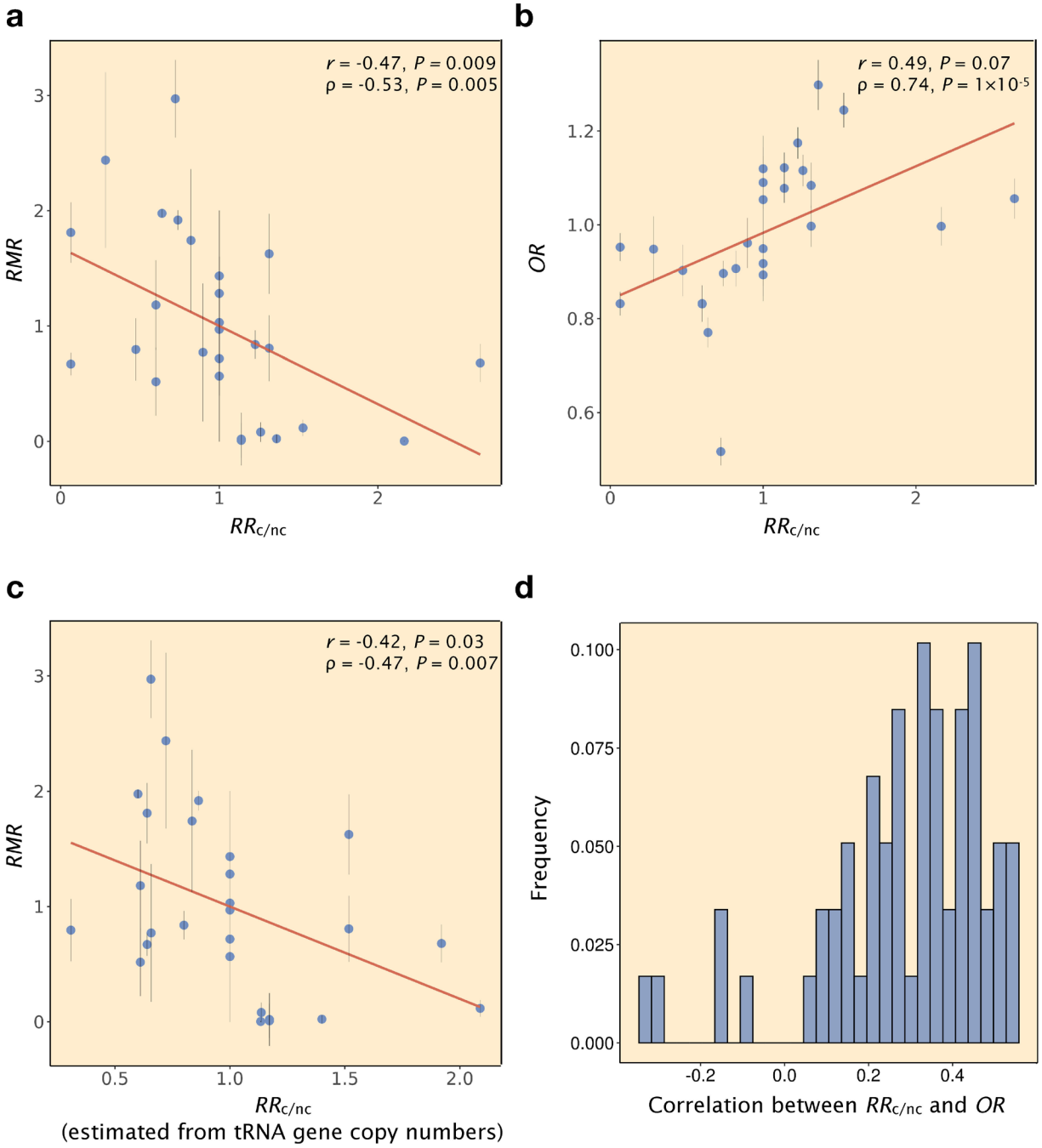

Figure 3. The relative ratio of cognate tRNA concentration to near-cognate tRNA concentration ($RR_{c/nc}$) is a major determinant of a codon's relative translational accuracy. **a**, *RMR* is negatively correlated with $RR_{c/nc}$ across codons in *E. coli*. **b**, *OR* is positively correlated with $RR_{c/nc}$ across codons in *E. coli*. **c**, *RMR* is negatively correlated with $RR_{c/nc}$ computed using tRNA gene copy numbers instead of tRNA concentrations in *E. coli*. **d**, Frequency distribution of Pearson's correlation between *OR* and $RR_{c/nc}$ computed using tRNA gene copy numbers in bacterial taxa with >80 tRNA genes. All *P*-values are based on permutation tests. In a-c, the red line is the linear regression.

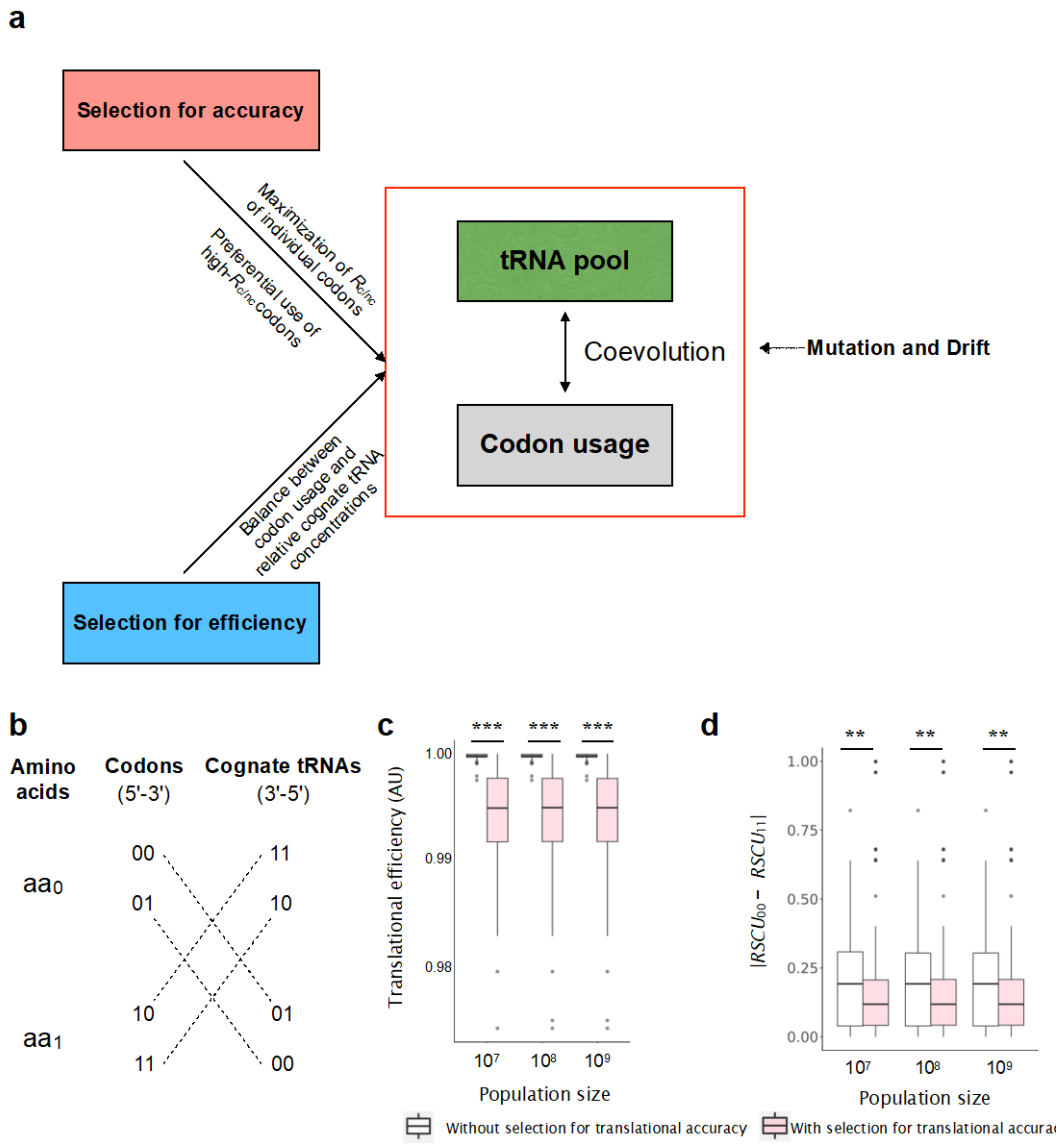

Figure 4. Selections for translational efficiency and accuracy shape the tRNA pool and codon usage. a, A model for the coevolution of the tRNA pool and codon usage driven by selections for translational efficiency and accuracy. **b.** A toy model with two amino acids, each encoded by two synonymous codons. A dotted line connects a codon with its near-cognate tRNA. **c**, Translational efficiency is significantly lower in the presence of selection for translational accuracy than in the absence of this selection. **d,** The absolute difference between the *RSCU* of 00 ($RSCU_{00}$) and that of 11 ($RSCU_{11}$) is smaller under the selection for translational accuracy than without this selection. With the selection, codon usage for aa_0 and that for aa_1 become coupled, because selection disfavors the cognate tRNA of the common codon of aa_0 to become the near-cognate tRNA of the common codon of aa_1, and vice versa. In c and d, each box plot shows the distribution from 200 replicates. The lower and higher edges of a box represent the first (qu_1) and third (qu_3) quartiles, respectively; the horizontal line inside the box indicates the median (md); the whiskers extend to the most extreme values inside inner fences, md ± $1.5(qu_3-qu_1)$; and the dots are outliers. *, $0.01 \le P < 0.05$, Wilcoxon rank-sum tests; **, $0.001 \le P < 0.01$; ***, $P < 0.001$.

Supplementary materials of
"Preferred synonymous codons are translated more accurately: Proteomic evidence, among-species variation, and mechanistic basis" by Sun & Zhang

The supplementary materials include:
Tables S1-S2
Figures S1-S4

Table S1. Pearson (*r*) and Spearman (*ρ*) correlations Between *OR* and *RSCU* in eukaryotic model organisms.

Species	*r*	*P*	*ρ*	*P*
Fly	0.84	<0.001	0.83	<0.001
Human	0.80	<0.001	0.74	<0.001
Mouse	0.74	<0.001	0.68	<0.001
Worm	0.42	<0.001	0.39	0.002
Yeast	0.35	0.01	0.39	0.003

Table S2. Pearson (*r*) and Spearman (*ρ*) correlations between *OR* and $RR_{c/nc}$ in eukaryotic model organisms.

Species	*r*	*P*	*ρ*	*P*
Fly	0.21	0.109	0.25	0.046
Human	0.39	0.007	0.29	0.055
Mouse	0.29	0.081	0.41	0.018
Worm	0.54	<0.001	0.57	<0.001
Yeast	0.42	0.045	0.37	0.053

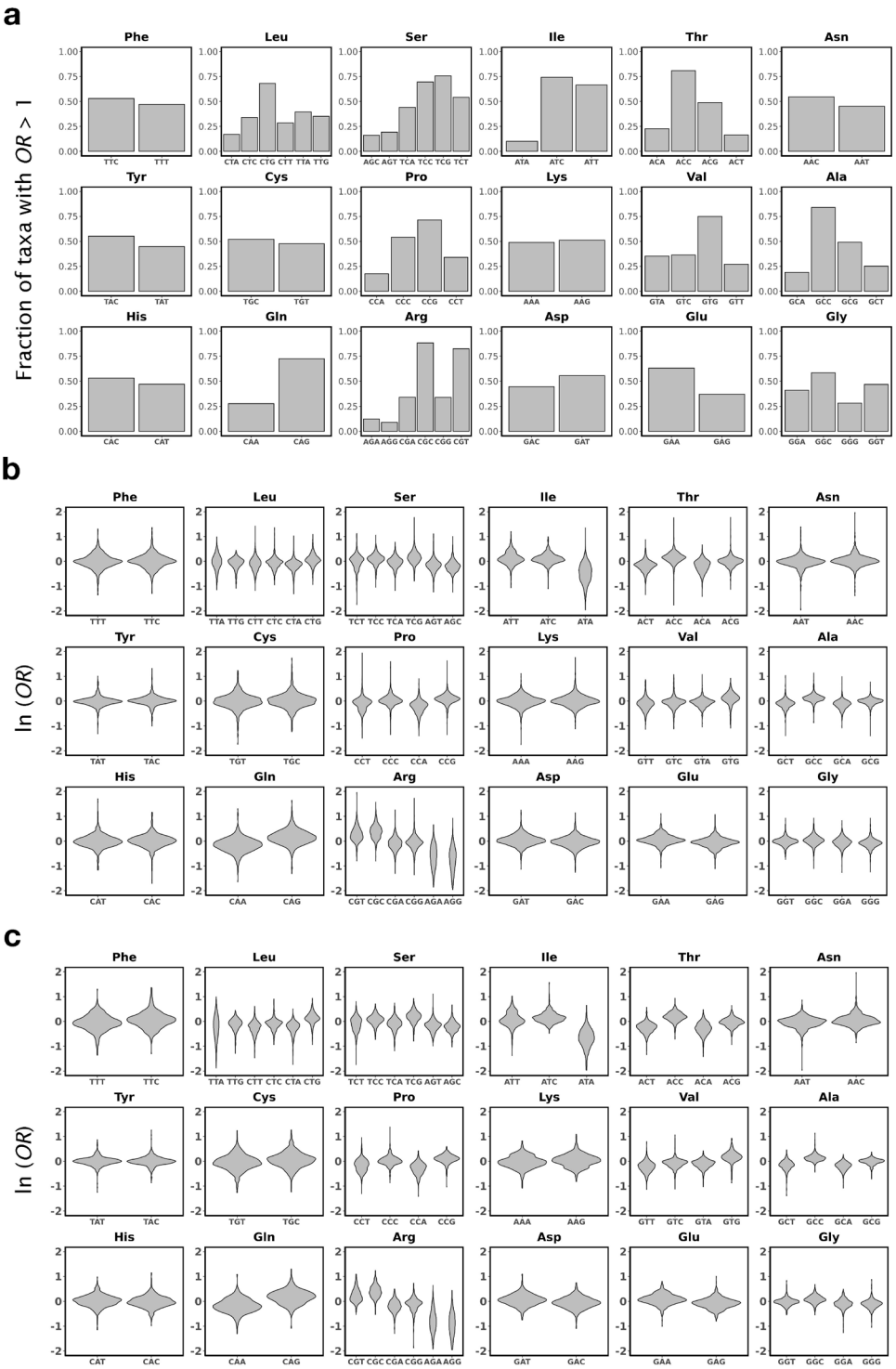

Fig. S1. Observed patterns of variation of relative translational accuracies of synonymous codons across bacterial taxa are robust. a, Fraction of 1197 taxa with *OR* > 1 for each codon. **b**, Violin plots showing frequency distributions of ln(*OR*) of individual codons across a subset of taxa in which the number of occurrences of each synonymous codon considered in *OR* estimation is at least 1000. **c**, Violin plots showing frequency distributions of ln(*OR*) of individual codons across a subset of taxa with strong signals of selection for translational accuracy (i.e., Pearson's correlation between *RSCU* and *OR* exceeds 0.5).

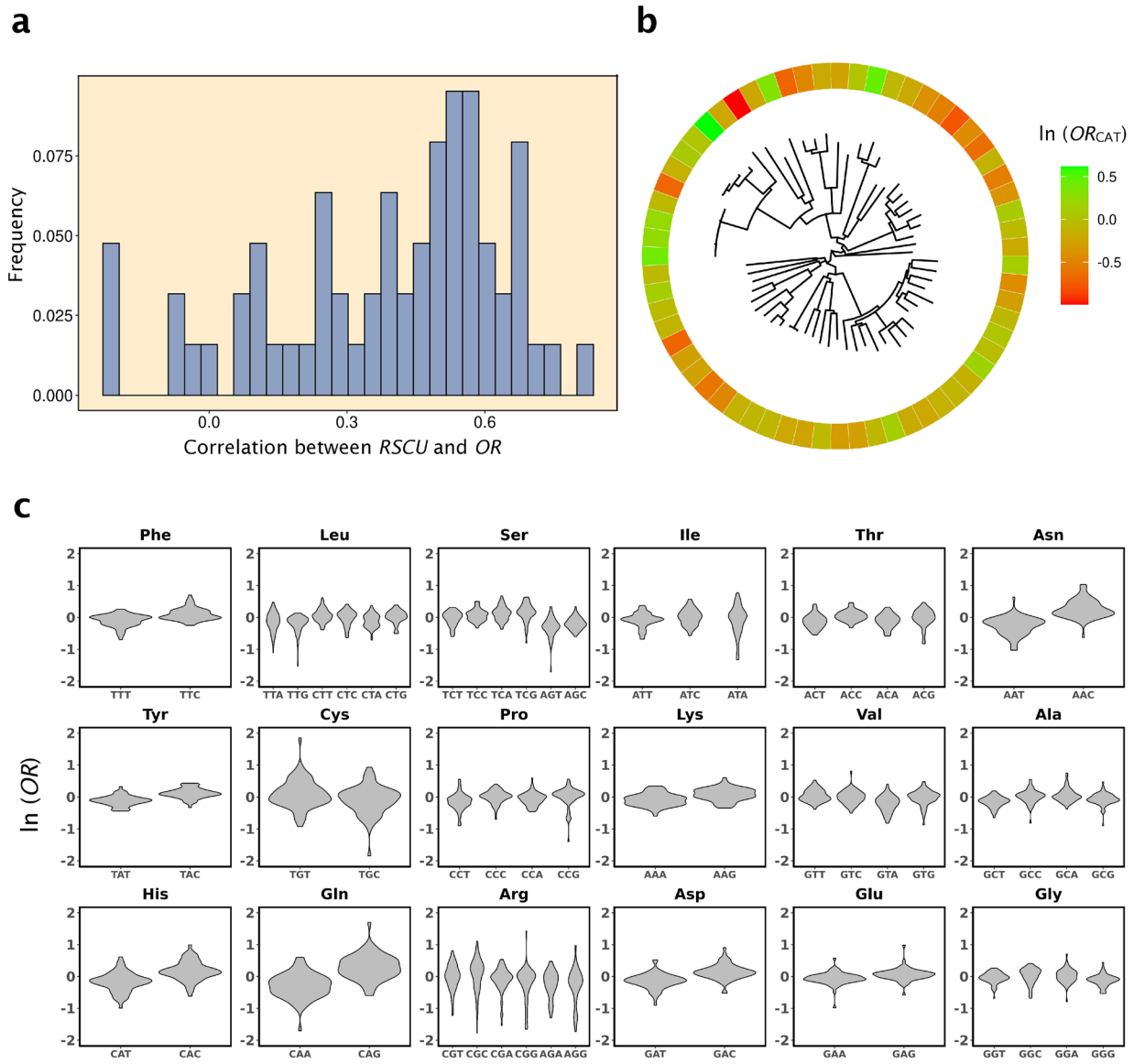

Fig. S2. Variation of relative translational accuracies of synonymous codons among Archaea taxa. a, Frequency distribution of Pearson's correlation between *RSCU* and *OR* in 63 taxa. Ninety percent of taxa show positive correlations. **b**, ln(*OR*) of codon CAT for each of the taxa arranged according to their phylogeny shown in the middle. **c**, Violin plots showing frequency distributions of ln(*OR*) of individual codons across taxa. ln(*OR*) appears less variable here than in Bacteria because of the much fewer Archaea taxa examined.

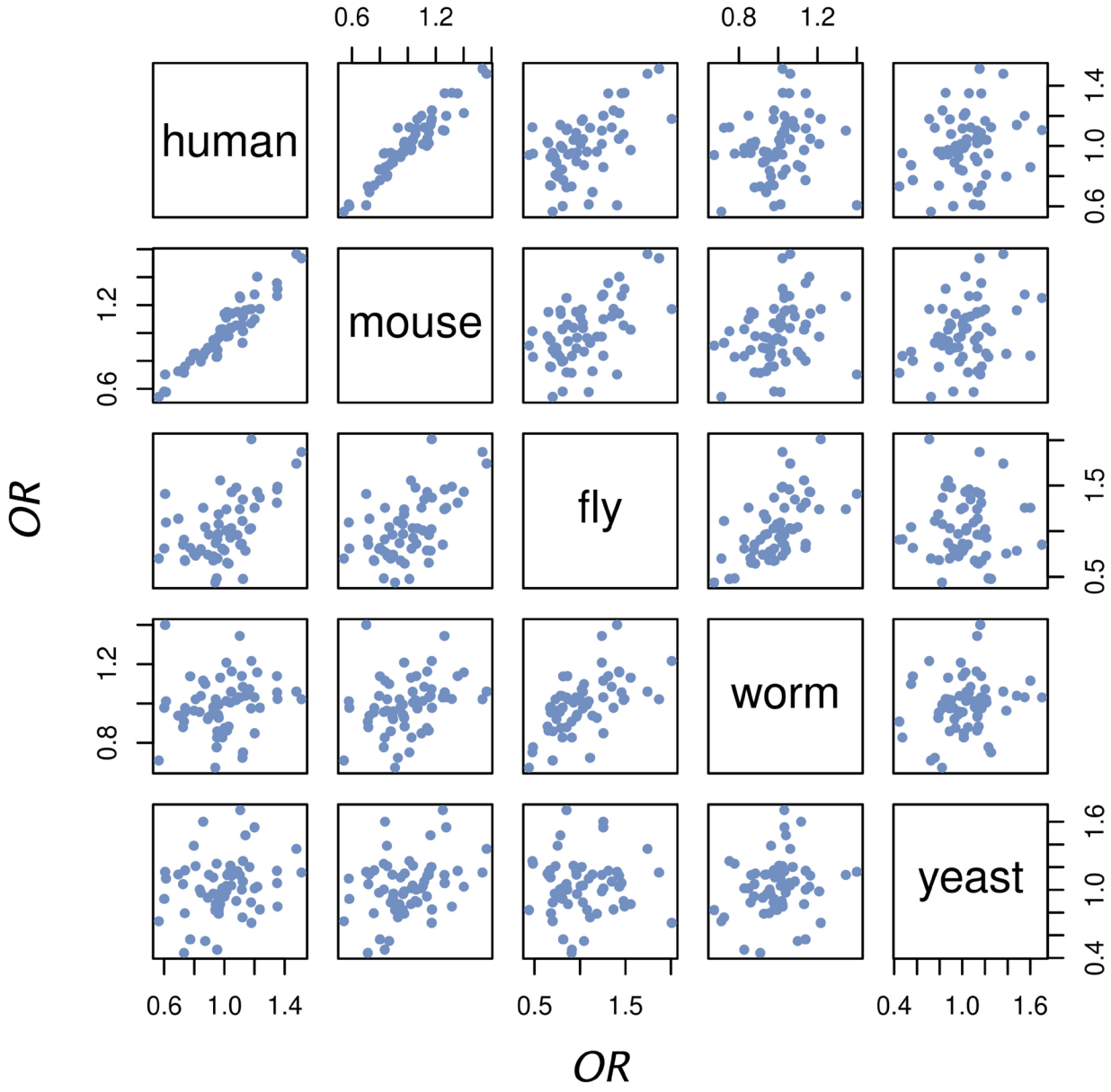

Fig. S3. Correlation in odds ratio (*OR*) across codons between eukaryotic model organisms. Each dot is a codon.

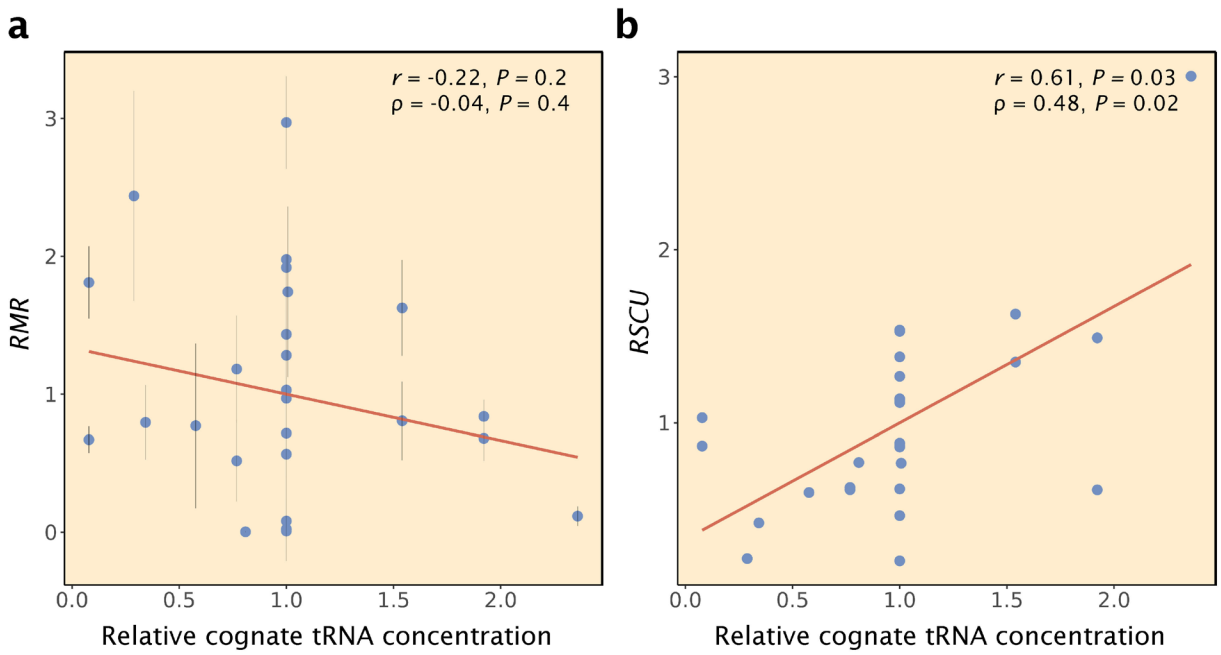

Fig. S4. Relationship between the cognate tRNA concentration and *RMR* or *RSCU* in *E. coli*. **a**, *RMR* of a codon is not significantly correlated with its relative cognate tRNA concentration, which is its cognate tRNA concentration divided by the mean cognate tRNA concentration of all codons coding for the same amino acid. **b**, *RSCU* of a codon is significantly positively correlated with its relative cognate tRNA concentration. *P*-values are based on permutation tests. Only the 27 codons with *RMR* estimates are analyzed in each panel to allow a direct comparison.