High transcriptional error rates vary as a function of gene expression level

Authors: K.M. Meer[1,2], P.G. Nelson[1], K. Xiong[3,4], J. Masel[1]*

Affiliations:

[1]Department of Ecology & Evolutionary Biology, University of Arizona, 1041 E Lowell St Tucson AZ 85721 USA.

[2]Present address: Computational Bioscience Program, University of Colorado Anschutz Medical Campus, 12800 E 19th Ave, Mail Stop 8303, Aurora, CO 80045 USA.

[3] Department of Molecular & Cellular Biology, University of Arizona, 1007 E Lowell St Tucson AZ 85721 USA.

[4]Present address: Computational Biology & Bioinformatics Program, Yale University, 266 Whitney Avenue, New Haven, CT 06520 USA.

*Correspondence to: masel@email.arizona.edu

Abstract

Errors in gene transcription can be costly, and organisms have evolved to prevent their occurrence or mitigate their costs. The simplest interpretation of the drift barrier hypothesis suggests that species with larger population sizes would have lower transcriptional error rates. However, *Escherichia coli* seems to have a higher transcriptional error rate than species with lower effective population sizes, e.g. *Saccharomyces cerevisiae*. This could be explained if selection in *E. coli* were strong enough to maintain adaptations that mitigate the consequences of transcriptional errors through robustness, on a gene by gene basis, obviating the need for low transcriptional error rates and associated costs of global proofreading. Here we note that if selection is powerful enough to evolve local robustness, selection should also be powerful enough to locally reduce error rates. We therefore predict that transcriptional error rates will be lower in highly abundant proteins on which selection is strongest. However, we only expect this result when error rates are high enough to significantly impact fitness. As expected, we find such a relationship between expression and transcriptional error rate for non C→U errors in *E. coli* but not in *S. cerevisiae*. We do not find this pattern for C→U changes in *E. coli*, presumably because most deamination events occurred during sample preparation, but do for C→U changes in *S. cerevisiae*, supporting the interpretation that C→U error rates estimated with an improved protocol, and which occur at rates comparable to *E. coli* non C→U errors, are biological.

Keywords: nearly neutral theory; weak selection; Cir-Seq; RNA editing; phenotypic mutation

Main Text

Errors are costly, and we therefore expect natural selection to reduce their rate. However, selection cannot achieve everything. In particular, it is only able to purge deleterious mutations when their selection coefficient *s* is significantly greater than one divided by the "effective population size". This numerical limit to selection may reflect not just the number of individuals in a population, but also competing selection at linked sites (Good and Desai 2014; Lynch 2007). The "nearly neutral theory" holds that deleterious mutations close to this limit are abundant (Ohta 1973), and the "drift barrier hypothesis" holds that differences in the precise location of this limit explain important differences among species (Lynch 2007). For example, codon usage bias is stronger in species believed to have higher effective population sizes (Vicario et al. 2007), indicating stronger selection to purge slightly deleterious synonymous mutations.

Rajon and Masel (2011) highlighted the distinction between a "global" solution that ameliorates a problem at many loci at once, and a set of "local" solutions that solve them one at a time. Because mutations affecting single loci are likely to have smaller fitness consequences than mutations with genome-wide effects, the drift barrier forms a more formidable barrier to local solutions than it does to global solutions. When local solutions evolve (in populations with large effective population sizes), they can obviate the need for global solutions. This yields the counterintuitive prediction that when global solutions are examined, it may be species with low effective population sizes that show the most extreme adaptations. Specifically, rates of error in transcription and translation could be higher in species with high effective population sizes,

since reducing error rates by kinetic proofreading is a costly global solution (Rajon and Masel 2011).

Here we focus on mistranscription errors, where during transcription, the wrong nucleic acid is incorporated at a single site. This can lead to non-functional proteins, incurring three types of costs. First is the energetic cost of futile transcription and translation (Wagner 2007); which can be significant in bacteria with large population sizes (Lynch and Marinov 2015; Petrov and Hartl 2000). Second, there is the opportunity cost of not using ribosomes to make other gene products (Dekel and Alon 2005; Kafri et al. 2016; Scott et al. 2014). Third, there is the cost of disposing of a misfolded and potentially toxic protein (Drummond and Wilke 2009; Geiler-Samerotte et al. 2011; Tomala and Korona 2013). Rajon and Masel (2011) predicted that in populations with smaller effective population sizes and more loci, costly proofreading might evolve to reduce the rate of mistranscription and hence the frequency with which these three costs are born, while in populations with very large effective population sizes and fewer loci, local solutions might evolve to reduce the cost of each mistranscription event, allowing their rate to stay high.

This prediction seems to have been confirmed for mistranscription (Xiong et al. 2017), whose rate of 8.2×10^{-5} in *Escherichia coli* (Traverse and Ochman 2016b) is far higher than that in *Saccharomyces cerevisiae* (3.9×10^{-6}) (Gout et al. 2017) or *Caenorhabditis elegans* (4.1×10^{-6}) (Gout et al. 2013), which have lower effective population sizes. Indeed, the rate is higher even

81 than that of *Buchnera aphidicola* ($4.7x10^{-5}$) (Traverse and Ochman 2016b). *Buchnera* is a highly
82 mutationally degraded species in which the drift barrier is an obstacle to the maintenance of
83 fidelity in many other important cellular functions (McCutcheon and Moran 2012); this high
84 rate in *Buchnera* may thus indicate that the drift barrier forms an obstacle even to global
85 solutions (Xiong et al. 2017). All these error rates except for that of *C. elegans* (Gout et al.
86 2013), were estimated using Cir-Seq (Acevedo and Andino 2014), and should therefore be
87 comparable, although sample preparation techniques differ in vulnerability to deamination.

88

89 If the drift barrier theory of Rajon and Masel (2011) explains the high rate of mistranscription in
90 *E. coli*, this implies that selection in *E. coli* must be potent enough to be sensitive to the
91 consequences of transcription errors in a local (i.e. site-specific) way, not just to its global rate.
92 Local solutions to mistranscription fall into two categories: local robustness to the
93 consequences of mistranscription when it occurs (this evolved robustness is hypothesized to be
94 responsible for permitting globally high mistranscription rates), and locally reduced
95 mistranscription rates at the sites most sensitive to it.

96

97 Here we test whether selection is able to maintain locally lower transcriptional error rates in
98 highly expressed genes. Selection to purge deleterious mutations is generally more effective in
99 highly expressed genes, as evidenced, for example, by stronger codon bias (Cutter and
100 Charlesworth 2006; Duret and Mouchiroud 1999; Ran et al. 2014; Sharp et al. 2010), which
101 lowers translational error rates (Zhang et al. 2016). Somatic mutations (Frigola et al. 2017),

102 alternative transcriptional start sites (Xu et al. 2019), post-transcriptional modifications (Liu and
103 Zhang 2018a, b), alternative mRNA polyadenylation (Xu and Zhang 2018), and translation errors
104 (Mordret et al. 2018) also occur at lower rates at sites where they are likely to have larger
105 effects. We similarly predict that because high mistranscription rates matter more for highly
106 expressed genes, highly expressed genes should evolve a lower rate of mistranscription. We
107 make this prediction for *E. coli*, where mistranscription rates are globally high and thus so is
108 local selection pressure. In contrast, we do not expect a relationship between expression level
109 and mistranscription rate in *S. cerevisiae*, where mistranscription rates are globally much lower.

110

111 Mistranscription rate data in *E. coli* were taken from Traverse and Ochman (2016a), who used
112 Cir-Seq (Acevedo and Andino 2014) to distinguish mistranscription events from sequencing
113 errors. Within the largest and highest-quality batch of their data (see Methods), data from four
114 experimental conditions (minimal vs. rich media, and midlog vs. stationary phase) were
115 sometimes analyzed separately and sometimes pooled. Mistranscription rates are much higher
116 for C→U substitutions: ~10^{-4} rather than ~10^{-5} for other mistranscription types. Since C→U
117 changes likely represent post-transcriptional deamination events, they are probably not
118 mistranscription errors per se, and they are also potentially more sensitive to preparation
119 artifacts (Chen et al. 2014), so we excluded them from most of our analysis.

120

121 To further ensure the quality of the data, we exclude "hotspot" nucleotide sites experiencing
122 significantly ($p<10^{-9}$) more errors of one type than expected from our model fitted as described

123 below. This eliminates recent mutations, inaccurate mapping of reads to the genome, or other
124 artifacts of the experiment or pipeline, as well as eliminating any sites subject to programmed
125 post-transcriptional RNA editing. We identified 5 protein-coding and 2,390 non-coding sites
126 that met this "hotspot" criteria for at least one experimental condition, and excluded these
127 sites from our analysis. The high rate of apparent mistranscription hotspots in non-coding genes
128 has been interpreted (Traverse and Ochman 2016a) as a consequence of the fact that *E. coli* has
129 multiple polymorphic rRNA operons, making mapping of reads inaccurate. We therefore restrict
130 our analysis to protein-coding genes only.

131

132 We modeled the number of errors observed per nucleotide site as count data, using a
133 generalized linear model. The number of errors expected is the product of the number of
134 observations of that nucleotide site, and the modeled mistranscription rate, the latter a linear
135 function of log protein abundance, experimental condition, and substitution type (see Methods,
136 model fitting culminated in the "final model" given in Equation 1, which includes protein
137 abundance, and experimental condition and substitution type each having effects on intercept
138 but not slope). The dependence on protein abundance (Figure 1; slope of 0 rejected with p
139 $=2\times10^{-12}$ in final model) supports our prediction from drift barrier theory, a result that gets
140 slightly stronger if we omit our hotspot removal procedure. The 11 non-C→U substitution types
141 have substantially different mistranscription rates; fitting different intercepts for each type
142 (while leaving their slopes the same) is strongly supported for inclusion in our final model (p =
143 2×10^{-16}). Similarly, different intercepts for different experimental conditions are supported (p =
144 1.5×10^{-3}).

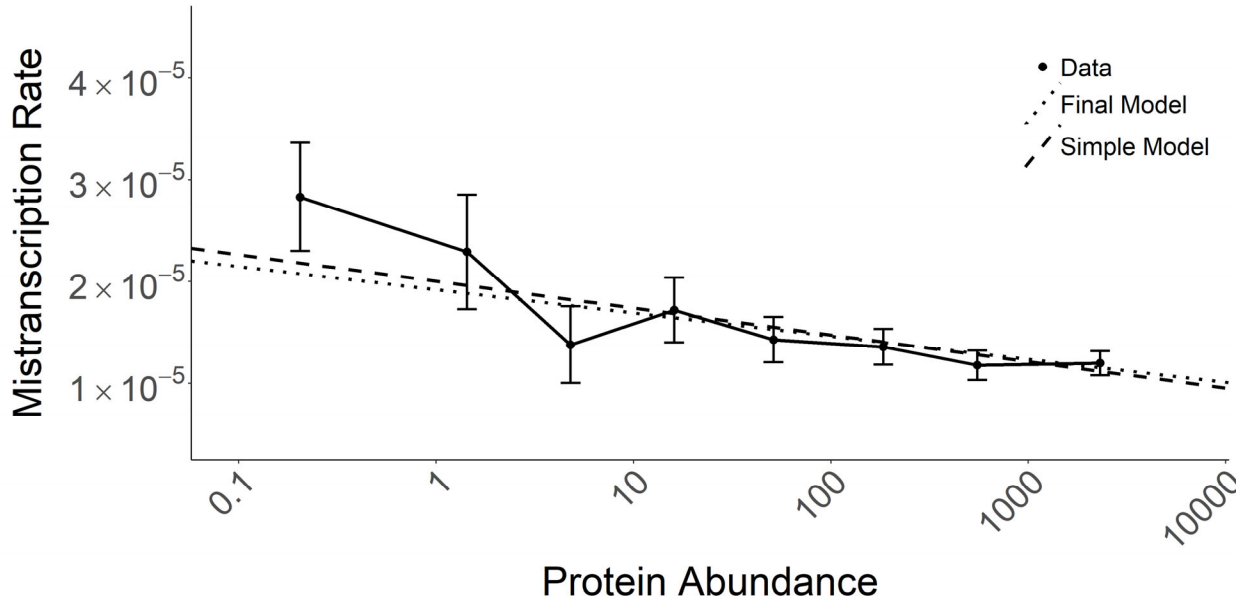

Figure 1. Highly expressed *E. coli* genes are subject to lower mistranscription rates. Dashed "simple model" shows a linear model fitted to a pooled dataset of 11 substitution types (i.e. excluding C→U) for all four conditions in batch #1. Dotted line shows the linear model with both condition and substitution type affecting the intercept but not the slope, plotted as a weighted average over conditions and substitutions, with weights proportional to the frequencies of opportunity to occur (i.e. by the numbers of reads of sites with A/C/G/U). Solid line shows the pooled data, binned by protein abundance as described in the Methods, and plotted according to mean protein abundance and the mean and 95% CI of the mistranscription rate within each bin. Data were divided into 10 bins; because of the limited availability of reads for low-expression genes, data within the first three bins were pooled. Note that mistranscription rate is per possible error, so the total mistranscription rate per nucleotide is around three times larger.

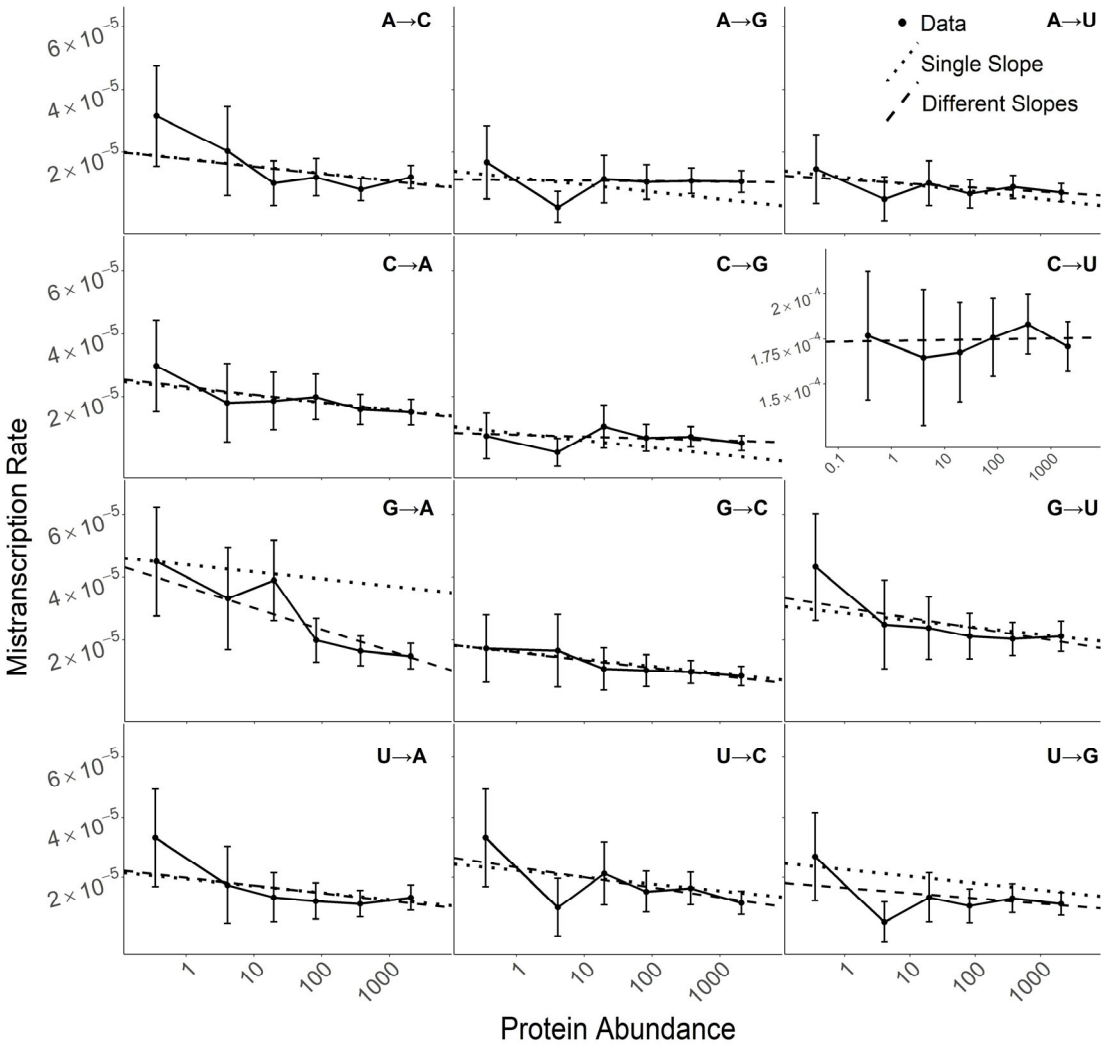

Figure 2. While different substitution types occur at different rates, all but C→U occur less often for highly abundant *E. coli* proteins. Dotted lines show our final linear model, in which the slope is the same across all error types. To display this model, we averaged the intercept over the four conditions, weighted according to the numbers of reads in each condition. Dashed lines show linear models with both the slope and intercept fitted separately for each error type using data pooled across all four conditions. Solid lines show the mean mistranscription rates, binned by protein abundance as described in the Methods, plotted according to mean protein abundance within each bin; error bars show 95% CI. Data were divided into 8 bins; because of the limited availability of reads for low-expression genes, data within the first three bins were pooled. Note that mistranscription rate is per possible error, so total mistranscription rate per nucleotide is around three times larger. The C→U error plot is shown inset due to very different y-axis values.

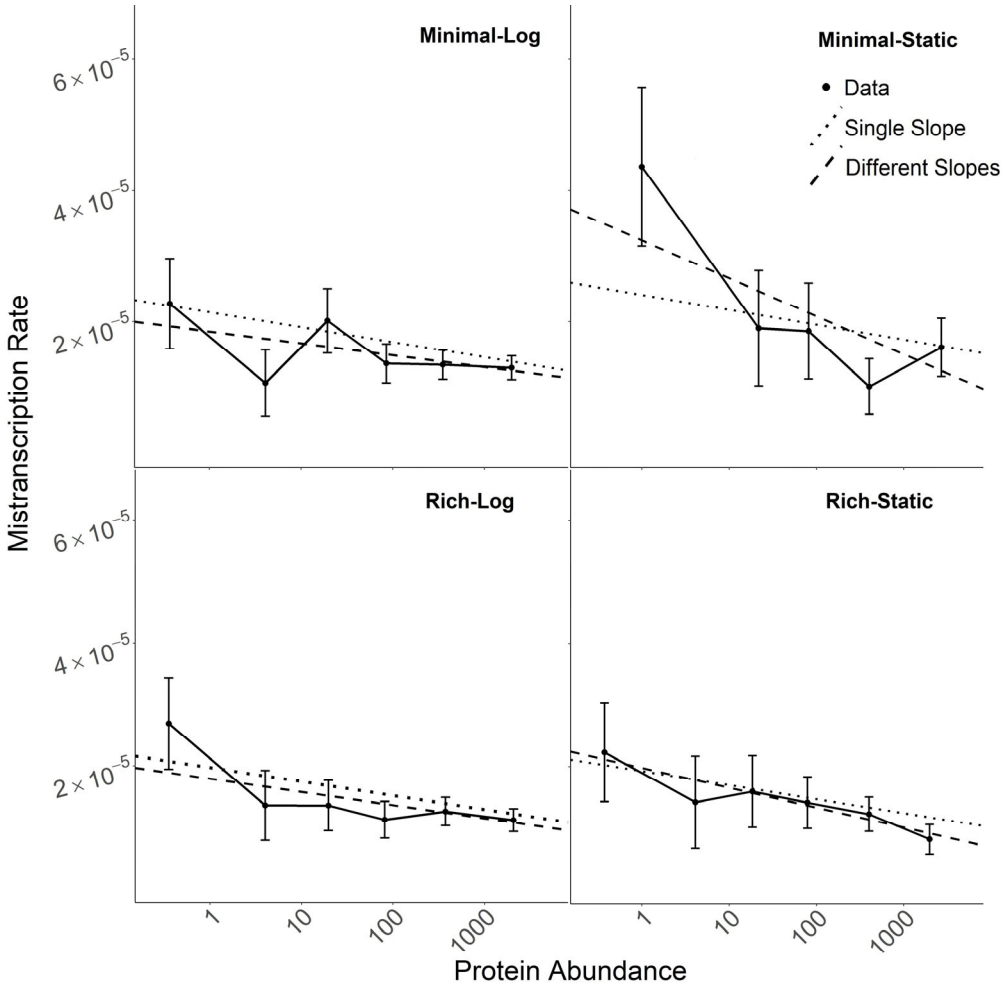

Figure 3. Different experimental conditions for *E. coli* yield similar mistranscription rates. Dotted lines show our final model, in which the slope is constant across all conditions. To display this model, we averaged the intercept over all error types, weighted by the frequencies of opportunity for the error type to occur (i.e. by the numbers of reads of sites with A/C/G/U). Dashed lines show linear models with both the slope and intercept fitted separately for each condition using data pooled across all error types. Solid lines show the mean mistranscription rates, binned by protein abundance as described in the Methods, plotted according to mean protein abundance within each bin; error bars show 95% CI. Data were divided into 8 bins; because of the limited availability of reads for low-expression genes, data within the first three bins were pooled. Note that mistranscription rate is per possible error, so total mistranscription rate per nucleotide is around three times larger.

185 Also fitting different slopes for different substitution types yields only a marginally significant
186 additional improvement over our final model, mostly due to the steeper slope for G→A
187 substitutions (Figure 2; p=0.051). Similarly, also fitting different slopes for each experimental
188 condition only marginally improves the fit relative to our final model (p = 0.026) with statistical
189 significance driven by a steeper slope in the minimal-static condition, which had far fewer data
190 points than the other conditions (Figure 3). The negligible differences in slope obtained from
191 pooling vs. stratifying (Figure 1) indicates that the slope is not spuriously generated by
192 systematic differences among the combined substitution types and experimental conditions.

193

194 Traverse and Ochman (2016a) reported that mistranscription errors were more commonly
195 synonymous (32%) than would be predicted if errors occurred at random across the genome
196 (24%). When controlling for the effects of substitution type, condition, and protein abundance
197 in our final model of mistranscription rates, the synonymous vs. non-synonymous status of the
198 potential mistranscription error did not predict the error rate (p = 0.08). Indeed, following our
199 data processing and quality filters, the overall frequency with which a mistranscription error
200 was synonymous was 23.4%, suggesting that the previously reported excess of synonymous
201 mistranscription events was due to data quality issues. In any case, whatever molecular
202 mechanism is responsible for variation in mistranscription rates, it seems to act at the level of
203 the gene rather than at the level of the nucleotide site.

204

205 Molecular chaperones play a critical role in mitigating the harm from mistranscription by
206 reducing misfolding. Genes that are chaperone clients might tolerate higher mistranscription
207 rates. Alternatively, sensitivity to mistranscription might select both for a lower
208 mistranscription rate and chaperone use. We found no support for either hypothesis; adding an
209 intercept term for *GroEL* chaperonin use was not a significant improvement on top of our final
210 model (p = 0.1).

211

212 As discussed in the Introduction, Cir-Seq data on the yeast *S. cerevisiae* indicates a much lower
213 mistranscription rate than *E. coli* (Gout et al. 2017), suggesting that it uses a global solution,
214 reducing site-specific selection pressures on mistranscription rates. We therefore do not predict
215 a relationship between gene expression and local mistranscription rate in this species, and do
216 not find one for the 11 non C→U substitution types (Figures 4 and 5; *p*=0.2 in our linear model
217 controlling for substitution type as a fixed effect).

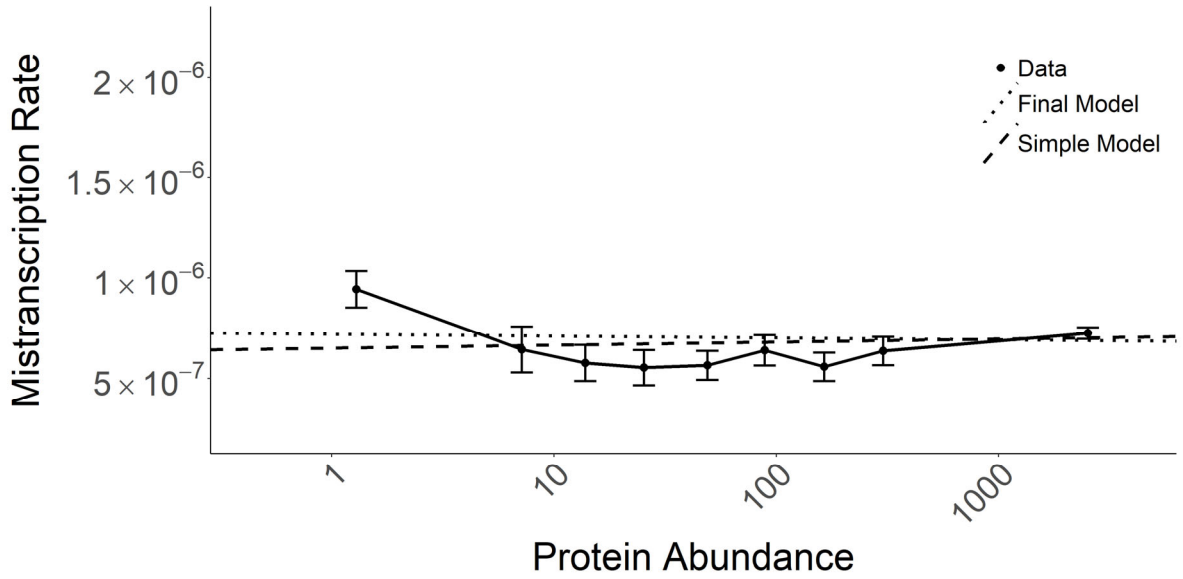

Figure 4. No relationship between protein abundance and mistranscription rate in *S. cerevisiae* non-C→U errors. Dashed line shows a simple linear model fitted to a pooled dataset of the 11 non-C→U substitution types. Dotted line shows the linear model that corresponds to the *E. coli* "Final Model", with substitution type affecting the intercept but not the slope. To display this model, we took a weighted average of the intercept over substitution types as a function of the frequencies of opportunity to occur. Solid line shows the pooled data, binned by protein abundance as described in the Methods, and plotted according to mean protein abundance and the mean and 95% CI of the mistranscription rate within each bin. Data were divided into 10 bins; because of the limited availability of reads for low-expression genes, data within the first two bins were pooled. The y-axis spans one order of magnitude, as in Figure 1, making the two figures visually comparable in terms of the magnitude of a slope or lack thereof. Note that mistranscription rate is per possible error, so total mistranscription rate per nucleotide is around three times larger.

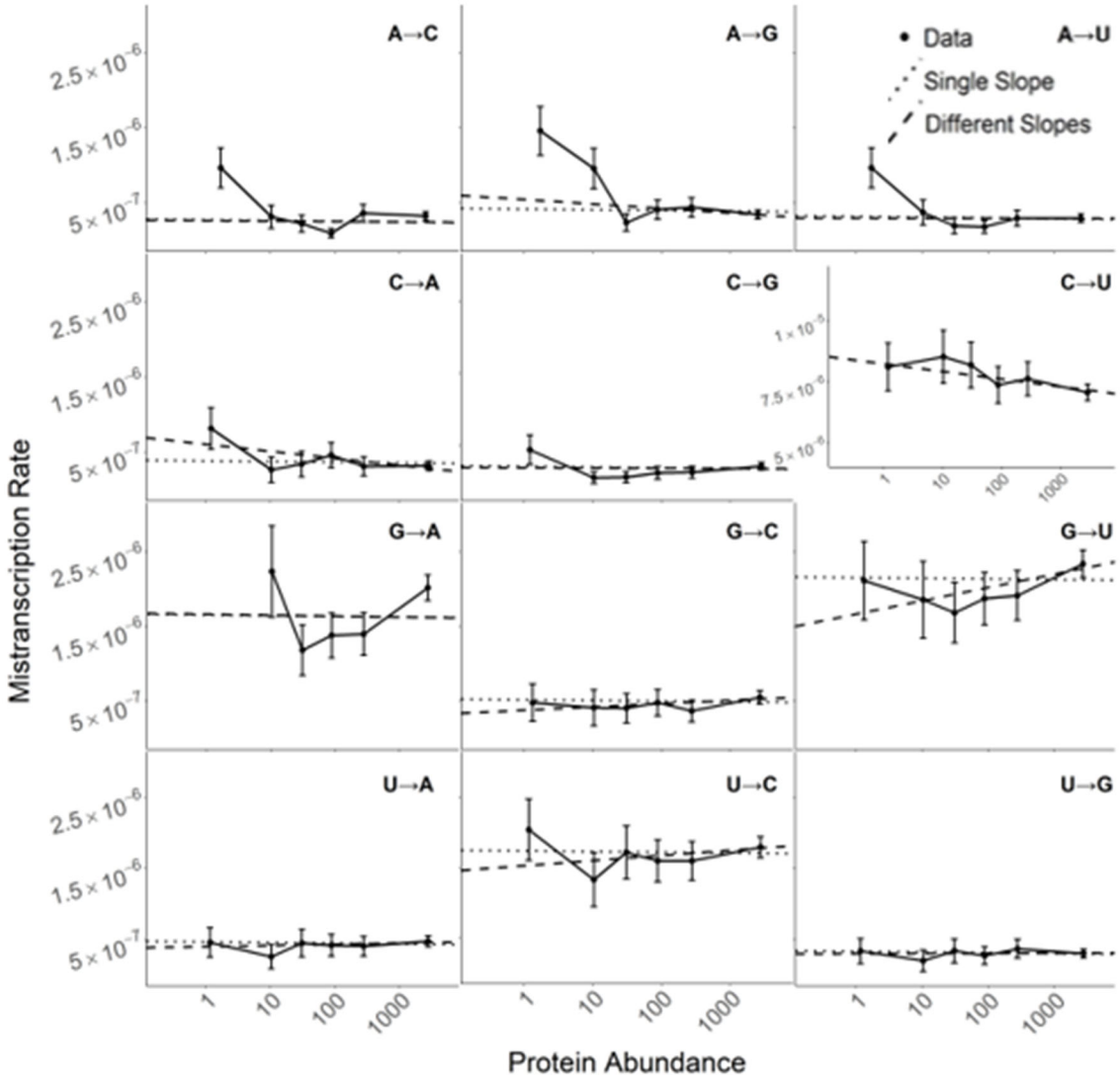

Figure 5. *S. cerevisiae* mistranscription rates broken down by base-substitution error types. Dotted lines show our final linear model, in which the intercept varies by error type, but the slope is the same across all error types. Dashed lines show linear models with both the slope and intercept fitted separately for each error type. Solid lines show the mean mistranscription rates, binned by protein abundance as described in the Methods, plotted according to mean protein abundance within each bin; error bars show 95% CI. Data were divided into 8 bins; because of the limited availability of reads for low-expression genes, data within the first three bins were pooled. Note that mistranscription rate is per possible error, so total mistranscription rate per nucleotide is around three times larger. The C→U error plot is shown inset due to radically different y-axis values than for the other error types.

243 However, C→U substitutions, which occur at much higher rates than other substitution types
244 and hence are subject to more selection even in *S. cerevisiae*, are less frequent for highly
245 abundant proteins (Figure 5; p= 0.006 for model restricted to C→U data). This confirms that the
246 protocol of Gout et al. (2017) succeeded in avoiding deamination events during sample
247 preparation (which should not depend on protein abundance), where that of Traverse &
248 Ochman (2016a) did not.

249

250 The high rate of mistranscription errors in *E. coli* came as a surprise to many (Traverse and
251 Ochman 2016a, b). This naturally raises the hypothesis that it is the data that are in error. While
252 the Cir-Seq technique is effective in preventing sequencing errors from inflating estimated
253 mistranscription rates (Acevedo and Andino 2014), it does not eliminate artifacts of the sample
254 preparation and analysis such as mutations occurring during the Cir-Seq experiment, nor
255 inaccurate mapping of reads to the genome. While these could artificially inflate estimated
256 mistranscription rates, we are not aware of any plausible mechanism by which the degree of
257 such inflation would be a function of protein abundance. Our results thus confirm the credibility
258 of the data, and hence of the statement that *E. coli* has a strikingly high non-C→U
259 mistranscription rate. After applying our quality filters, we calculate the total rate of all non-
260 C→U errors as 4.1×10^{-5} per site, or 8.6×10^{-5} if C→U errors are also included. In contrast, in *S.*
261 *cerevisiae*, we calculate from the data of Gout et al. (2017) a non-C→U mistranscription rate of
262 2.3×10^{-6}, or 3.5×10^{-6} with the C→U error type included.

263

264 The dependence of the *E. coli* mistranscription rate on the strength of selection (as reflected by
265 protein abundance), but not the *S. cerevisiae* mistranscription rate, is consistent with proposed
266 drift barrier explanations (McCandlish and Plotkin 2016; Rajon and Masel 2011; Xiong et al.
267 2017). In particular, *E. coli* is smaller and is generally accepted to have a larger effective
268 population size than *S. cerevisiae*. *E. coli* also has fewer loci, occurring within 4453 genes in K-12
269 (Riley et al. 2006) compared to 5178 genes in *S. cerevisiae* (Engel et al. 2014), which makes it
270 easier to evolve robustness at each one. A typical *E. coli* mRNA encodes 0.02% of the total
271 protein molecules in a cell, while a yeast mRNA encodes only 0.01% (Lu et al. 2007), meaning
272 that each transcription error in *E. coli* has a proportionately greater impact at one moment in
273 time. While a typical yeast mRNA has a longer half-life and so makes proteins over a longer time
274 (Siwiak and Zielenkiewicz 2013), this is not enough to counteract all other factors making local
275 solutions easier to evolve in *E. coli*.

276

277 We have shown that local mistranscription rates vary in a systematic way on a per-gene basis,
278 but have not determined the mechanisms by which expression error rates vary.
279 Mistranscription rates are affected at the gene level by the presence or absence of specific RNA
280 polymerase subunits (Thomas et al. 1998; Walmacq et al. 2009) or transcription factors
281 (Bubunenko et al. 2017; Irvin et al. 2014; Roghanian et al. 2015), and at a level higher than
282 nucleotide site but lower than gene by sequence characteristics such as long mononucleotide
283 repeats (Ackermann and Chao 2006; Gu et al. 2010). The degree to which such factors vary
284 systematically is unknown.

285

286 We have also shown that the local mistranscription rates even of highly expressed *E. coli* genes
287 are higher than the global mistranscription rate in *S. cerevisiae*, suggesting that *E. coli* genes are
288 somehow more robust to the consequences of mistranscription than are *S. cerevisiae* genes.
289 One sequence-based measure of robustness to transcription errors is the use or avoidance of
290 "fragile" codons that are one substitution away from a stop codon (Cusack et al. 2011). Serine,
291 arginine, glycine, and leucine can be encoded by either fragile or non-fragile codons. Of these
292 four amino acids, preference for fragile vs. robust codons is independent of %GC content only
293 in serine (TCT, TCC, AGT, and AGC are robust, while TCA and TCG are fragile), making serine
294 ideal for comparing codon fragility between species. We find that 25.7% of serine residues are
295 encoded by fragile codons in the *E. coli* proteome and 31.6% in *S. cerevisiae*. Dividing by the
296 expected frequency of robust serine residues (1/3) shows that *E. coli* uses fragile serine codons
297 77% as often as expected, compared to 95% in *S. cerevisiae*. These numbers suggest that
298 avoidance of fragile codons might explain at least some of the robustness of the *E. coli* genome.
299 However, the robustness associated with *E. coli*'s global solution is not so complete as to
300 eliminate selection for locally lower mistranscription rates in the genes subject to the strongest
301 selection, leading to the trend detected here.

302

303 **Methods**

304 Scripts used in these analyses are available at https://github.com/MaselLab/Meer-et-al-
305 Transcriptional-Error-Rates.

306

307 ***E. coli mistranscription data***

308 Pre-processed data were obtained from Traverse and Ochman (2016a), that included how many
309 times each of the 4,641,652 nucleotide loci in the K-12 MG155 reference genome (GenBank
310 accession: NC.000913.3) was observed, and how often each nucleotide was seen there. We
311 assigned these loci to 4,140 protein coding genes and 178 non-coding genes using the
312 annotation of GenBank accession NC.000913.3. We analyzed the 3,935,551 nucleotide loci
313 within annotated non-overlapping protein-coding ORFs, and 47,344 nucleotide loci from non-
314 coding genes based on annotated 'start' and 'stop' positions. We excluded any sites that were
315 present in overlapping genes, as we could not assign a single error rate or protein abundance in
316 such cases.

317

318 Traverse and Ochman (2016a) data were obtained in multiple batches (referred to as
319 "replicates" in their data tables), with results reported only on two of the batches. Batch #2 had
320 approximately half as much data and twice the error rate of batch #1, so we restrict our
321 analysis to batch #1 only. Combining the data from each of the four experimental conditions
322 (minimal vs. rich media, and midlog vs. stationary phase) within batch #1 effectively yielded
323 15,742,204 protein-coding sites and 189,376 non-coding sites, where "site" is used here as
324 shorthand for condition×nucleotide locus, i.e. to describe the set of reads of a nucleotide locus
325 within just one experimental condition.

326

327 We excluded any site that had no reads and any protein-coding transcript site with no protein
328 abundance measure, leaving 5,994,463 coding and 182,233 non-coding sites. Each site can
329 experience three different substitution error types (e.g. C→U, C→A, and C→G), which we
330 treated separately, yielding 17,983,389 coding and 546,699 non-coding "possible errors" for
331 analysis. Note that data for the three alternative errors at the same site are not, strictly
332 speaking, independent, because the occurrence of one error reduces the denominator for the
333 other two. However, at low error rates, this effect is negligible.

334

335 Mutations occurring during the Cir-Seq experiment, inaccurate mapping of reads to the
336 genome, or other artifacts of the experiment or pipeline can result in the appearance of
337 mistranscription "hot spots" that are best removed. We calculated the likelihoods of seeing that
338 many or more errors for each of the 18,530,088 possible errors being analyzed, using a
339 significance cutoff of 10^{-9} to ensure that only 10^{-9}×18,530,088=0.02 possible errors are falsely
340 excluded, or slightly more if there is genuine biological variation in mistranscription rates
341 beyond that captured by our linear model. We calculated likelihoods from a cumulative
342 binomial distribution based on the number of reads at that site and the rate of error expected
343 at that site from our model. When a possible error was excluded with likelihood $< 10^{-9}$, we
344 excluded the entire nucleotide locus (i.e. all three possible substitutions in all four conditions).
345 We performed an iterative procedure, first fitting a model of constant error rate for all non
346 C→U errors and a separate error rate for C→U errors, using expectations from this model to
347 exclude outliers, then using the cleaned-up data to develop a more sophisticated error rate
348 model of all conditions/substitution types, and using the revised expectations from this model

349 to update which loci should be excluded etc. until convergence. In the final iteration, one or
350 more possible errors was determined to be an outlier at 5 protein-coding and 2,390 non-coding
351 loci. For protein-coding outliers, we excluded all possible errors at each of the 5 outlier loci, i.e.
352 up to 60 possible errors (3 possible errors at 5 loci in 4 conditions). Some sites had no transcript
353 reads in some conditions, resulting in only 48 rather than 60 possible errors being excluded by
354 this procedure, leaving 17,983,341 possible errors in protein-coding transcript regions for
355 analysis. Excluding C→U substitutions, due to their significantly higher error rate and likelihood
356 of occurring post-transcriptionally, further reduced this to 16,466,559 non-C→U possible errors
357 for analysis.

358

359 ***S. cerevisiae mistranscription data***

360 Similarly pre-processed transcript data were obtained from Gout et al. (2017), who recorded
361 how many times each nucleotide locus was observed in the S288C reference genome (GenBank
362 accession: GCA_000146045.2), to which the wild-type BY4741 strain used in their experiment is
363 very closely related. Only one experimental condition was used in this study. Using the same
364 methodology as for the *E. coli* data, we used the accession to assign nucleotide sites to the
365 5,983 protein-coding nuclear gene regions based on the annotated 'start' and 'stop' positions.
366 This process identified 8,853,931 nucleotide loci within annotated protein-coding ORFs,
367 resulting in 26,561,793 possible errors for analysis.

368

369 Excluding any transcript site without reads or with unreported or zero protein abundance left
370 us with 18,649,818 possible errors. Using our outlier detection protocol, we identified 44 loci
371 containing possible errors as outliers and excluded all possible errors at the associated loci (132
372 possible errors in total), leaving 18,649,686 possible errors for analysis.

373

374 C→U errors were also identified as having a substantially higher error rate in the yeast data
375 (1.8×10^{-5} versus 2.3×10^{-6} for other mistranscription types), and were excluded from our analysis,
376 resulting in 17,394,875 non-C→U possible errors.

377

378 ***Protein abundance data***

379 Integrated protein abundance data were taken from PaxDB (Wang et al. 2015).

380

381 ***GroEL client status***

382 We labelled the 1,929,741 possible errors associated with 252 *E. coli* proteins as having GroEL
383 client status, based on the identification of those proteins by Kerner et al. (2005) as specific
384 interactors with the GroEL chaperonin.

385

386 ***Statistical model***

387 We modeled the error rate at site *i* within gene *j* as a linear function of the log-abundance of

388 protein *j*, i.e.

389 $$\frac{\mathrm{E}_i}{\mathrm{R}_i} = \rho + \beta \ln(Abundance_j)$$

390 where E_i is the number of reads containing a particular error and R_i is the total number of

391 reads at that nucleotide site.

392

393 To better model the error function in the linear model, we multiply both sides by R_i:

394 $$E_i \sim R_i + R_i \log_{10}(Abundance_j) + \varepsilon_{Poisson}$$

395 The observed number of errors E_i has the properties of count data, and so can be modeled as a

396 sample from a Poisson distribution. We fitted the statistical model above using a generalized

397 linear model function in R (glm, stats package), specifying the family of the model as

398 "poisson(link = identity)". For *E. coli*, experimental condition and type of error (excluding C→U)

399 were added as fixed effects to yield:

400 $$E_i \sim type : R_i + cond : R_i + R_i \log_{10}(Expression_j) + \varepsilon_{Poisson} \quad (1)$$

401

402 Slope as a function of expression level could also be made dependent on type and/or condition,

403 but adding these terms resulted in either insignificant or marginally significant improvements to

404 the *E. coli* model fit. For *S. cerevisiae*, the condition term does not apply, and expression was

405 not supported a predictive in the model. P-values associated with adding or removing terms to

406 a linear model were obtained using the anova command with the Chisq option to compare
407 nested models in R. P-values associated with comparisons to the final models for *E. coli* and *S.*
408 *cerevisiae* are given in the text.

409

410 ***Data Binning***

411 We binned data by protein abundance for visualization and comparison to the fitted models. All
412 possible errors were sorted by the abundance value of the corresponding protein. Bin
413 boundaries were evenly spaced along our log-abundance axis between the 5% quantile and the
414 95% quantile, with data beyond these quantiles included in the edge bins. For each bin, one
415 point was plotted with y-value equal to the mean and 95% confidence interval of the
416 mistranscription rate and an x-value equal to the geometric mean of protein abundance. The
417 number of mistranscription errors observed is expected to follow a binomial distribution with r
418 trials, each with probability p of an error. We thus estimated a standard error of $\sqrt{(1-\hat{p})\hat{p}/r}$,
419 where r is the total number of reads within the bin and $\hat{p}$ is the observed error frequency
420 within the bin. To generate the 95% confidence interval we multiplied this standard error by
421 1.96. To keep standard errors for low-abundance bins reasonably low, data from several low-
422 abundance bins were combined.

423

424 **Acknowledgements:** This work was supported by the Undergraduate Biology Research Program
425 at the University of Arizona, the John Templeton Foundation [39667, 60814], and the National
426 Institutes of Health (GM104040). We thank Christophe Herman for helpful discussions and

Charles Traverse and Jean-François Gout for making their preprocessed data readily available to us.

References

Acevedo A, Andino R. 2014. Library preparation for highly accurate population sequencing of RNA viruses. Nat Protoc. 9:1760–1769. doi: 10.1038/nprot.2014.118

Ackermann M, Chao L. 2006. DNA sequences shaped by selection for stability. PLoS Genetics. 2:e22.

Bubunenko MG, et al. 2017. A Cre Transcription Fidelity Reporter Identifies GreA as a Major RNA Proofreading Factor in *Escherichia coli*. Genetics. 206:179-187. doi: 10.1534/genetics.116.198960

Chen G, et al. 2014. Cytosine Deamination Is a Major Cause of Baseline Noise in Next-Generation Sequencing. Molecular Diagnosis & Therapy. 18:587-593. doi: 10.1007/s40291-014-0115-2

Cusack BP, Arndt PF, Duret L, Crollius HR. 2011. Preventing dangerous nonsense: selection for robustness to transcriptional error in human genes. PLoS Genetics. 7:e1002276.

Cutter AD, Charlesworth B. 2006. Selection intensity on preferred codons correlates with overall codon usage bias in *Caenorhabditis remanei*. Curr Biol. 16:2053-7. doi: 10.1016/j.cub.2006.08.067

Dekel E, Alon U. 2005. Optimality and evolutionary tuning of the expression level of a protein. Nature. 436:588-592.

Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. Nat Rev Genet. 10:715-24. doi: 10.1038/nrg2662

Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci U S A. 96:4482-4487. doi: 10.1073/pnas.96.8.4482

Engel SR, et al. 2014. The reference genome sequence of *Saccharomyces cerevisiae*: then and now. G3: Genes, Genomes, Genetics. 4:389-398.

Frigola J, et al. 2017. Reduced mutation rate in exons due to differential mismatch repair. Nat Genet. 49:1684-1692. doi: 10.1038/ng.3991

Geiler-Samerotte KA, et al. 2011. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. Proc Natl Acad Sci U S A. 108:680-685. doi: 10.1073/pnas.1017570108

460 Good BH, Desai MM. 2014. Deleterious Passengers in Adapting Populations. Genetics.
461 198:1183-1208. doi: 10.1534/genetics.114.170233

462 Gout J-F, et al. 2017. The landscape of transcription errors in eukaryotic cells. Sci Adv.
463 3:e1701484. doi: 10.1126/sciadv.1701484

464 Gout J-F, et al. 2013. Large-scale detection of in vivo transcription errors. Proc Natl Acad Sci U S
465 A. 110:18584-18589. doi: 10.1073/pnas.1309843110

466 Gu T, et al. 2010. Avoidance of long mononucleotide repeats in codon pair usage. Genetics.
467 186:1077-1084.

468 Irvin JD, et al. 2014. A genetic assay for transcription errors reveals multilayer control of RNA
469 polymerase II fidelity. PLoS Genet. 10:e1004532. doi: 10.1371/journal.pgen.1004532

470 Kafri M, Metzl-Raz E, Jona G, Barkai N. 2016. The Cost of Protein Production. Cell Rep. 14:22-31.
471 doi: 10.1016/j.celrep.2015.12.015

472 Kerner MJ, et al. 2005. Proteome-wide Analysis of Chaperonin-Dependent Protein Folding in
473 *Escherichia coli*. Cell. 122:209-220. doi: 10.1016/j.cell.2005.05.028

474 Liu Z, Zhang J. 2018a. Human C-to-U Coding RNA Editing Is Largely Nonadaptive. Mol Biol Evol.
475 35:963-969. doi: 10.1093/molbev/msy011

476 Liu Z, Zhang J. 2018b. Most m^6A RNA Modifications in Protein-Coding Regions Are Evolutionarily
477 Unconserved and Likely Nonfunctional. Mol Biol Evol. 35:666-675. doi:
478 10.1093/molbev/msx320

479 Lu P, et al. 2007. Absolute protein expression profiling estimates the relative contributions of
480 transcriptional and translational regulation. Nature Biotechnology. 25:117-124. doi:
481 10.1038/nbt1270

482 Lynch M. 2007. *The origins of genome architecture*. Sunderland: Sinauer Associates.

483 Lynch M, Marinov GK. 2015. The bioenergetic costs of a gene. Proc Natl Acad Sci U S A.
484 112:15690-5. doi: 10.1073/pnas.1514974112

485 McCandlish DM, Plotkin JB. 2016. Transcriptional errors and the drift barrier. Proc Natl Acad Sci
486 U S A. 113 3136-3138. doi: 10.1073/pnas.1601785113

487 McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. Nat Rev
488 Microbiol. 10:13-26. doi: 10.1038/nrmicro2670

489 Mordret E, et al. 2018. Systematic detection of amino acid substitutions in proteome reveals a
490 mechanistic basis of ribosome errors. bioRxiv.255943. doi: 10.1101/255943

491 Ohta T. 1973. Slightly Deleterious Mutant Substitutions in Evolution. Nature. 246:96-98. doi:
492 10.1038/246096a0

493 Petrov DA, Hartl DL. 2000. Pseudogene evolution and natural selection for a compact genome. J
494 Hered. 91:221-227.

495 Rajon E, Masel J. 2011. The evolution of molecular error rates and the consequences for
496 evolvability. Proc Natl Acad Sci U S A. 108:1082-1087. doi: 10.1073/pnas.1012918108

497 Ran W, Kristensen DM, Koonin EV. 2014. Coupling between protein level selection and codon
498 usage optimization in the evolution of bacteria and archaea. mBio. 5:e00956-14. doi:
499 10.1128/mBio.00956-14

500 Riley M, et al. 2006. Escherichia coli K-12: a cooperatively developed annotation snapshot—
501 2005. Nucleic Acids Res. 34:1-9.

502 Roghanian M, Zenkin N, Yuzenkova Y. 2015. Bacterial global regulators DksA/ppGpp increase
503 fidelity of transcription. Nucleic Acids Res. 43:1529-36. doi: 10.1093/nar/gkv003

504 Scott M, Klumpp S, Mateescu EM, Hwa T. 2014. Emergence of robust growth laws from optimal
505 regulation of ribosome synthesis. Mol Syst Biol. 10:747. doi: 10.15252/msb.20145379

506 Sharp PM, Emery LR, Zeng K. 2010. Forces that influence the evolution of codon bias. Philos
507 Trans R Soc Lond B Biol Sci. 365:1203-12. doi: 10.1098/rstb.2009.0305

508 Siwiak M, Zielenkiewicz P. 2013. Transimulation - Protein Biosynthesis Web Service. PLoS ONE.
509 8:e73943. doi: 10.1371/journal.pone.0073943

510 Thomas MJ, Platas AA, Hawley DK. 1998. Transcriptional fidelity and proofreading by RNA
511 polymerase II. Cell. 93:627-637.

512 Tomala K, Korona R. 2013. Evaluating the fitness cost of protein expression in *Saccharomyces*
513 *cerevisiae*. Genome Biol Evol. 5:2051-60. doi: 10.1093/gbe/evt154

514 Traverse CC, Ochman H. 2016a. Conserved rates and patterns of transcription errors across
515 bacterial growth states and lifestyles. Proc Natl Acad Sci U S A. 113:3311-3316. doi:
516 10.1073/pnas.1525329113

517 Traverse CC, Ochman H. 2016b. Correction for Traverse and Ochman, Conserved rates and
518 patterns of transcription errors across bacterial growth states and lifestyles. Proc Natl Acad Sci
519 U S A. 113:E4257-E4258. doi: 10.1073/pnas.1609677113

520 Vicario S, Moriyama EN, Powell JR. 2007. Codon usage in twelve species of Drosophila. BMC
521 Evol Biol. 7:226. doi: 10.1186/1471-2148-7-226

522 Wagner A. 2007. Energy costs constrain the evolution of gene expression. J Exp Zool Part B.
523 308B:322-324. doi: 10.1002/jez.b.21152

524 Walmacq C, et al. 2009. Rpb9 subunit controls transcription fidelity by delaying NTP
525 sequestration in RNA polymerase II. J Biol Chem. 284:19601-12. doi: 10.1074/jbc.M109.006908

526 Wang M, et al. 2015. Version 4.0 of PaxDb: Protein abundance data, integrated across model
527 organisms, tissues, and cell-lines. Proteomics. 15:3163-3168. doi: 10.1002/pmic.201400441

528 Xiong K, McEntee JP, Porfirio DJ, Masel J. 2017. Drift Barriers to Quality Control When Genes
529 Are Expressed at Different Levels. Genetics. 205:397-407. doi: 10.1534/genetics.116.192567

530 Xu C, Park J-K, Zhang J. 2019. Evidence that alternative transcriptional initiation is largely
531 nonadaptive. PLoS Biology. 17:e3000197. doi: 10.1371/journal.pbio.3000197

532 Xu C, Zhang J. 2018. Alternative Polyadenylation of Mammalian Transcripts Is Generally
533 Deleterious, Not Adaptive. Cell Syst. 6:734-742. doi: 10.1016/j.cels.2018.05.007

534 Zhang D, et al. 2016. The Effect of Codon Mismatch on the Protein Translation System. PLoS
535 ONE. 11:e0148302. doi: 10.1371/journal.pone.0148302

536