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.

*joint first author

*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* (especially G→A), 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. 2019) 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 are more sensitive to preparation artifacts (Chen et al. 2014), i.e. they may not be
118 mistranscription errors, we excluded them from most of our analysis.

119

120 To further ensure the data quality, we exclude "hotspot" nucleotide sites experiencing
121 significantly ($p<10^{-9}$) more errors of one type than expected from our model fitted as described
122 below. This eliminates recent mutations, inaccurate mapping of reads to the genome, or other

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

129

130 We modeled the number of errors observed per nucleotide site as count data, using a
131 generalized linear model. The number of errors expected is the product of the number of
132 observations of that nucleotide site, and the modeled mistranscription rate, the latter a linear
133 function of log protein abundance, experimental condition, and substitution type (see
134 Methods). The dependence on protein abundance (Figure 1; slope of 0 rejected from Eq. 1
135 model with $p=2\times10^{-14}$) supports our prediction from drift barrier theory, a result that gets
136 slightly stronger if we omit our hotspot removal procedure. The 11 non-C→U substitution types
137 have substantially different mistranscription rates (Supplementary Figure S1); fitting different
138 intercepts for each type (while leaving their slopes the same) is strongly supported for inclusion
139 in our Eq. 1 model ($p = 2\times10^{-16}$).

140

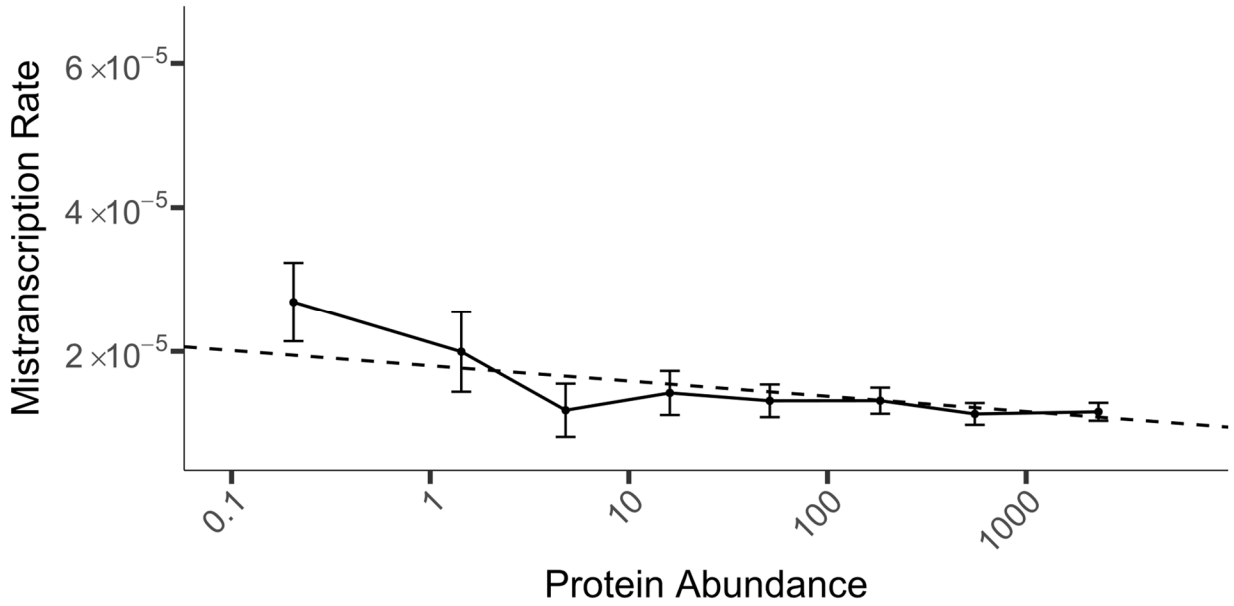

Figure 1. Highly expressed *E. coli* genes are subject to lower mistranscription rates. The dashed line shows the Eq. 1 model applied to the 11 non C→U substitution types, in which both condition and substitution type affect the intercept but not the slope, plotted as a weighted average over conditions 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 show 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.

Different intercepts for different experimental conditions are also supported, in addition (p = 1.5×10^{-3}). Fitting different slopes for each experimental condition only marginally improves the fit relative to our Eq. 1 model (p = 0.052), mostly attributable to a steeper slope in the minimal-static condition, which had far fewer data points than the other conditions (Figure S2).

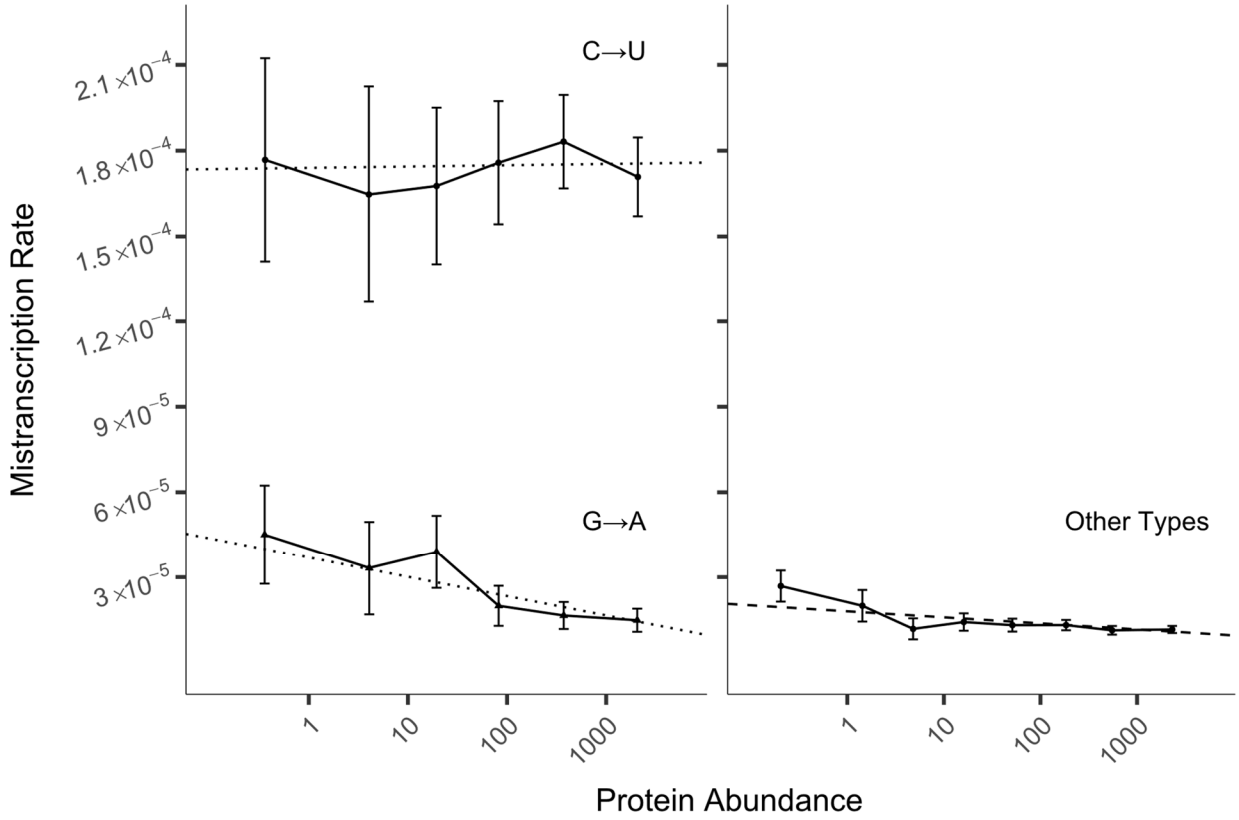

Figure 2. C→U errors in *E. coli* are mostly artifacts, G→A depend most strongly on protein abundance, but the other 10 error types also show dependence. Dotted lines (left) show linear models with both the slope and intercept fitted separately for each error type using data pooled across all four conditions; for a comparison of all 12 error types, see supplemental Figure S1. The C→U slope is not different than 0 (p=0.91). The dashed line (right) shows an Eq. 1 model in which the slope is the same across all 10 error types (non-C→U, non-G→A). To display this model, we averaged the intercept over the four conditions, weighted according to the numbers of reads in each condition. 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.

Standing out from results on all non C→U error types in Figure S1, and shown in Figure 2, is the fact that G→A errors depend more strongly on protein abundance than other error types do (p=3×10^{-4}, Eq. 2 as improvement on Eq. 1). A separate model fit to G→A error data only, gives a

177 slope of -2.9×10^{-6} (95% CI of -1.6×10^{-6} to -4.2×10^{-6}) with $\log_{10}$ protein abundance, i.e. there are
178 1.6 to 4.2 fewer G→A errors per million G transcription events per 10-fold increase in
179 expression, against a background of about 20-40 errors per million G transcription events. To
180 ensure that the non-zero slope of Figure 1 is not driven solely by G→A errors, we repeated the
181 analysis for the 10 error types, i.e. excluding both C→U and G→A (Figure 2, right). This yields a
182 slope of -8.4×10^{-7} ($p=1\times10^{-7}$) with $\log_{10}$ protein abundance, with a 95% confidence interval
183 corresponding to 0.4 and 1.2 fewer expression errors per million opportunities per 10-fold
184 increase in expression.
185

186 Traverse and Ochman (2016a) reported that mistranscription errors were more commonly
187 synonymous (32%) than would be predicted if errors occurred at random across the genome
188 (24%). When controlling for the effects of substitution type, condition, and protein abundance
189 in our Eq. 2 model of mistranscription rates, the synonymous vs. non-synonymous status of the
190 potential mistranscription error did not predict the error rate (p = 0.89). Indeed, following our
191 data processing and quality filters, the overall frequency with which a mistranscription error
192 was synonymous was 23.4%, suggesting that the previously reported excess of synonymous
193 mistranscription events was due to data quality issues. In any case, whatever molecular
194 mechanism is responsible for variation in mistranscription rates, it seems to act at the level of
195 the gene rather than at the level of the nucleotide site.
196

197 Molecular chaperones play a critical role in mitigating the harm from mistranscription by
198 reducing misfolding. Genes that are chaperone clients might tolerate higher mistranscription
199 rates. Alternatively, sensitivity to mistranscription might select both for a lower
200 mistranscription rate and chaperone use. We found no support for either hypothesis; adding an
201 intercept term for *GroEL* chaperonin use was not a significant improvement on top of our Eq. 2
202 model (p = 0.085). We also tested other predictors including gene length, absolute position of a
203 locus (number of nucleotides from the start of gene), and relative position of a locus (absolute
204 position / total gene length), but neither slope nor intercept were significantly different from 0
205 (i.e. p>0.05) for any of the three metrics.

206

207 As discussed in the Introduction, Cir-Seq data on the yeast *S. cerevisiae* indicates a much lower
208 mistranscription rate than *E. coli* (Gout et al. 2017), suggesting that it uses a global solution,
209 reducing site-specific selection pressures on mistranscription rates. We therefore do not predict
210 a relationship between gene expression and local mistranscription rate in this species, and do
211 not find one for the 11 non C→U substitution types (Figure 3 bottom; p=0.2 in our Eq. 1 model
212 controlling for substitution type as a fixed effect).

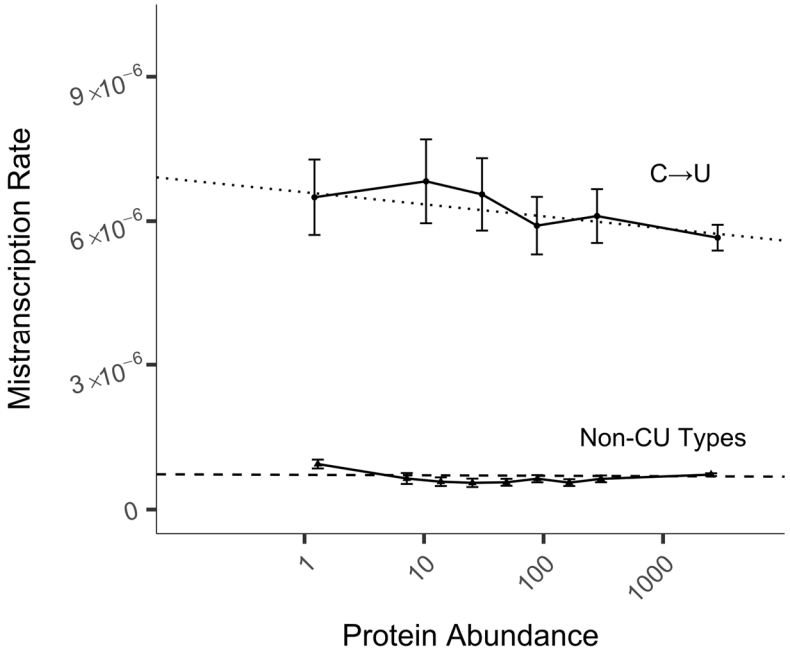

Figure 3. In *S. cerevisiae*, only C→U mistranscription errors depend on protein abundance. Dashed line shows a linear model fitted to a pooled dataset of the 11 non-C→U substitution types. Dotted line shows a linear model fitted to C→U data alone. To display non-C→U model fit, we took a weighted average of the intercept over substitution types as a function of the frequencies of opportunity to occur. Solid line shows 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. C→U 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. For non-C→U data, two out of 10 bins were pooled. Note that mistranscription rate is per possible error, so total mistranscription rate per nucleotide is around three times larger. All 12 error types are shown separately in Supplementary Fig. S3.

However, C→U substitutions, which occur at much higher rates than other substitution types and hence are subject to more selection even in *S. cerevisiae*, are less frequent for highly abundant proteins (Fig. 3 top; p= 0.006 for non-zero slope on a C→U equivalent of Eq. 2). This confirms that the protocol of Gout et al. (2017) succeeded in avoiding deamination events during sample preparation (which should not depend on protein abundance), where that of Traverse & Ochman (2016a) did not.

231

232 The high rate of mistranscription errors in *E. coli* came as a surprise to many (Traverse and
233 Ochman 2016a, b). This naturally raises the hypothesis that it is the data that are in error. While
234 the Cir-Seq technique is effective in preventing sequencing errors from inflating estimated
235 mistranscription rates (Acevedo and Andino 2014), it does not eliminate artifacts of the sample
236 preparation and analysis such as mutations occurring during the Cir-Seq experiment, nor
237 inaccurate mapping of reads to the genome. While these could artificially inflate estimated
238 mistranscription rates, we are not aware of any plausible mechanism by which the degree of
239 such inflation would be a function of protein abundance. Our results thus confirm the credibility
240 of the data, and hence of the statement that *E. coli* has a strikingly high non-C→U
241 mistranscription rate. After applying our quality filters, we calculate the total rate of all non-
242 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.*
243 *cerevisiae*, we calculate from the data of Gout et al. (2017) a non-C→U mistranscription rate of
244 2.3×10^{-6}, or 3.5×10^{-6} with the C→U error type included.

245

246 The dependence of the *E. coli* mistranscription rate on the strength of selection (as reflected by
247 protein abundance), but not the *S. cerevisiae* mistranscription rate, is consistent with proposed
248 drift barrier explanations (McCandlish and Plotkin 2016; Rajon and Masel 2011; Xiong et al.
249 2017). In particular, *E. coli* is smaller and is generally accepted to have a larger effective
250 population size than *S. cerevisiae*. *E. coli* also has fewer loci, occurring within 4453 genes in K-12
251 (Riley et al. 2006) compared to 5178 genes in *S. cerevisiae* (Engel et al. 2014), which makes it

252 easier to evolve robustness at each one. What is more, the average *E. coli* mRNA produces
253 about 540 proteins out of a total of 2.5×10^6 per cell (Lu et al. 2007), i.e. 0.02% of the
254 proteome, which is twice as much as the average yeast mRNA producing 5600 proteins out of a
255 total of 5×10^7 per cell (Lu et al. 2007), i.e. 0.01% of the proteome. While a typical yeast mRNA
256 has a longer half-life and so makes proteins over a longer time (6.7 vs. 27.4 minutes; Siwiak and
257 Zielenkiewicz 2013), the magnitude of this should not be enough to counteract all other factors
258 making local solutions easier to evolve in *E. coli*.

259

260 We have shown that local mistranscription rates vary in a systematic way on a per-gene basis,
261 but have not determined the mechanisms by which expression error rates vary.
262 Mistranscription rates are affected by local sequence characteristics such as long
263 mononucleotide repeats (Ackermann and Chao 2006; Gu et al. 2010) and at the gene level by
264 the presence or absence of specific RNA polymerase subunits (Thomas et al. 1998; Walmacq et
265 al. 2009) or transcription factors (Bubunenko et al. 2017; Irvin et al. 2014; Roghanian et al.
266 2015). Our finding that G→A errors depend more strongly on expression than do other error
267 types in E. coli suggests that GreA, which specifically reduces G→A transcription errors
268 (Traverse and Ochman 2018), may be a likely mechanistic candidate.

269

270 We have also shown that the local mistranscription rates even of highly expressed *E. coli* genes
271 are higher than the global mistranscription rate in *S. cerevisiae*, suggesting that *E. coli* genes are
272 somehow more robust to the consequences of mistranscription than are *S. cerevisiae* genes.
273 However, the robustness associated with *E. coli*'s global solution is not so complete as to

274 eliminate selection for locally lower mistranscription rates in the genes subject to the strongest
275 selection, leading to the trend detected here.

276

277 **Methods**

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

280

281 ***E. coli mistranscription data***

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

291

292 Traverse and Ochman (2016a) data were obtained in multiple batches (referred to as
293 "replicates" in their data tables), with results reported only on two of the batches. Batch #2 had

294 approximately half as much data and twice the error rate of batch #1, so we restrict our
295 analysis to batch #1 only. Combining the data from each of the four experimental conditions
296 (minimal vs. rich media, and midlog vs. stationary phase) within batch #1 effectively yielded
297 15,742,204 protein-coding sites and 189,376 non-coding sites, where "site" is used here as
298 shorthand for condition×nucleotide locus, i.e. to describe the set of reads of a nucleotide locus
299 within just one experimental condition.

300

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

308

309 Mutations occurring during the Cir-Seq experiment, inaccurate mapping of reads to the
310 genome, or other artifacts of the experiment or pipeline can result in the appearance of
311 mistranscription "hot spots" that are best removed. We calculated the likelihoods of seeing that
312 many or more errors for each of the 18,530,088 possible errors being analyzed, using a
313 significance cutoff of 10^{-9} to ensure that only 10^{-9}×18,530,088=0.02 possible errors are falsely
314 excluded, or potentially more if there is genuine biological variation in mistranscription rates

315 beyond that captured by our linear model. We calculated likelihoods from a cumulative
316 binomial distribution based on the number of reads at that site and the rate of error expected
317 at that site from our model. When a possible error was excluded with likelihood < 10^{-9}, we
318 excluded the entire nucleotide locus (i.e. all three possible substitutions in all four conditions).
319 We performed an iterative procedure, first fitting a model of constant error rate for all non
320 C→U errors and a separate error rate for C→U errors, using expectations from this model to
321 exclude outliers, then using the cleaned-up data to develop a more sophisticated error rate
322 model of all conditions/substitution types, and using the revised expectations from this model
323 to update which loci should be excluded etc. until convergence. In the final iteration, one or
324 more possible errors was determined to be an outlier at 5 protein-coding and 2,390 non-coding
325 loci. For protein-coding outliers, we excluded all possible errors at each of the 5 outlier loci, i.e.
326 up to 60 possible errors (3 possible errors at 5 loci in 4 conditions). Some sites had no transcript
327 reads in some conditions, resulting in only 48 rather than 60 possible errors being excluded by
328 this procedure, leaving 17,983,341 possible errors in protein-coding transcript regions for
329 analysis. Excluding C→U substitutions, due to their significantly higher error rate and likelihood
330 of occurring post-transcriptionally, further reduced this to 16,466,559 non-C→U possible errors
331 for analysis.

332

333 ***S. cerevisiae mistranscription data***

334 Similarly pre-processed transcript data were obtained from Gout et al. (2017), who recorded
335 how many times each nucleotide locus was observed in the S288C reference genome (GenBank

336 accession: GCA_000146045.2), to which the wild-type BY4741 strain used in their experiment is
337 very closely related. Only one experimental condition was used in this study. Using the same
338 methodology as for the *E. coli* data, we used the accession to assign nucleotide sites to the
339 5,983 protein-coding nuclear gene regions based on the annotated 'start' and 'stop' positions.
340 This process identified 8,853,931 nucleotide loci within annotated protein-coding ORFs,
341 resulting in 26,561,793 possible errors for analysis.

342

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

347

348 C→U errors were also identified as having a substantially higher error rate in the yeast data
349 (1.8×10^{-5} versus 2.3×10^{-6} for other mistranscription types), and were excluded from some
350 analyses, resulting in 17,394,875 non-C→U possible errors.

351

352 ***Protein abundance data***

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

354

355 ***GroEL client status***

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

359

360 ***Statistical model***

361 We modeled the error rate at site *i* within gene *j* as a linear function of the log-abundance of
362 protein *j*, i.e.

363 $$\frac{\mathrm{E}_i}{\mathrm{R}_i} = \rho + \beta \ln(Abundance_j)$$

364 where E_i is the number of reads containing a particular error and R_i is the total number of
365 reads at that nucleotide site.

366

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

368 $$E_i \sim R_i + R_i \log_{10}(Abundance_j) + \varepsilon_{Poisson}$$

369 The observed number of errors E_i has the properties of count data, and so can be modeled as a
370 sample from a Poisson distribution. We fitted the statistical model above using a generalized
371 linear model function in R (glm, stats package), specifying the family of the model as
372 "poisson(link = identity)". For *E. coli*, experimental condition and type of error (excluding C→U)
373 were added as fixed effects to yield:

374 $$E_i \sim type : R_i + cond : R_i + R_i \log_{10}(Expression_j) + \varepsilon_{Poisson} \quad (1)$$

375

376 Slope as a function of expression level can also be made dependent on type and/or condition.
377 For *S. cerevisiae*, the condition term does not apply, and expression was not supported as
378 predictive in the model. For *E. coli*, a separate slope for G→A errors was supported, yielding

379 $$E_i \sim type : R_i + cond : R_i + GA : R_i \log_{10}\left(Expression_j\right) + \varepsilon_{Poisson} \quad (2)$$

380

381 P-values associated with adding or removing terms to Eq. 1 or Eq. 2 models were obtained
382 using the anova command with the Chisq option to compare nested models in R, as given
383 throughout the text, sometimes manually correcting the number of degrees of freedom.

384

385 ***Data Binning***

386 We binned data by protein abundance for visualization and comparison to the fitted models. All
387 possible errors were sorted by the abundance value of the corresponding protein. Bin
388 boundaries were evenly spaced along our log-abundance axis between the 5% quantile and the
389 95% quantile, with data beyond these quantiles included in the edge bins. For each bin, one
390 point was plotted with y-value equal to the mean and 95% confidence interval of the
391 mistranscription rate and an x-value equal to the geometric mean of protein abundance. The
392 number of mistranscription errors observed is expected to follow a binomial distribution with r
393 trials, each with probability p of an error. We thus estimated a standard error of $\sqrt{(1-\hat{p})\hat{p}/r}$,
394 where r is the total number of reads within the bin and $\hat{p}$ is the observed error frequency

395 within the bin. To generate the 95% confidence interval we multiplied this standard error by
396 1.96. To keep standard errors for low-abundance bins reasonably low, data from several low-
397 abundance bins were combined.

398

399 Binned data is shown for the purpose of illustrating that it is appropriate to log-transform
400 protein abundance before using it as a linear predictor of error rate. Note that it is normal for
401 the edge bins to depart from the linear trend (Wilke 2013), and thus the linearity of the fit
402 should be judged within the central region of the relationship.

403

404 **Acknowledgements:** This work was supported by the Undergraduate Biology Research Program
405 at the University of Arizona, the John Templeton Foundation [39667, 60814], and the National
406 Institutes of Health (GM104040). We thank Christophe Herman for helpful discussions and
407 Charles Traverse and Jean-François Gout for making their preprocessed data readily available to
408 us.

409

410 **References**

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

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

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

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

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

420 Dekel E, Alon U. 2005. Optimality and evolutionary tuning of the expression level of a protein. Nature.
421 436:588-592.

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

424 Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in
425 Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci U S A. 96:4482-4487. doi:
426 10.1073/pnas.96.8.4482

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

429 Frigola J, et al. 2017. Reduced mutation rate in exons due to differential mismatch repair. Nat Genet.
430 49:1684-1692. doi: 10.1038/ng.3991

431 Geiler-Samerotte KA, et al. 2011. Misfolded proteins impose a dosage-dependent fitness cost and trigger
432 a cytosolic unfolded protein response in yeast. Proc Natl Acad Sci U S A. 108:680-685. doi:
433 10.1073/pnas.1017570108

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

436 Gout J-F, et al. 2017. The landscape of transcription errors in eukaryotic cells. Sci Adv. 3:e1701484. doi:
437 10.1126/sciadv.1701484

438 Gout J-F, et al. 2013. Large-scale detection of in vivo transcription errors. Proc Natl Acad Sci U S A.
439 110:18584-18589. doi: 10.1073/pnas.1309843110

440 Gu T, et al. 2010. Avoidance of long mononucleotide repeats in codon pair usage. Genetics. 186:1077-
441 1084.

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

444 Kafri M, Metzl-Raz E, Jona G, Barkai N. 2016. The Cost of Protein Production. Cell Rep. 14:22-31. doi:
445 10.1016/j.celrep.2015.12.015

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

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

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

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

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

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

457 McCandlish DM, Plotkin JB. 2016. Transcriptional errors and the drift barrier. Proc Natl Acad Sci U S A.
458 113 3136-3138. doi: 10.1073/pnas.1601785113

459 McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol.
460 10:13-26. doi: 10.1038/nrmicro2670

461 Mordret E, et al. 2019. Systematic Detection of Amino Acid Substitutions in Proteomes Reveals
462 Mechanistic Basis of Ribosome Errors and Selection for Translation Fidelity. Molecular Cell. 75:427-441.
463 doi: 10.1016/j.molcel.2019.06.041

464 Ohta T. 1973. Slightly Deleterious Mutant Substitutions in Evolution. Nature. 246:96-98. doi:
465 10.1038/246096a0

466 Petrov DA, Hartl DL. 2000. Pseudogene evolution and natural selection for a compact genome. J Hered.
467 91:221-227.

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

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

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

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

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

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

480 Siwiak M, Zielenkiewicz P. 2013. Transimulation - Protein Biosynthesis Web Service. PLoS ONE.
481 8:e73943. doi: 10.1371/journal.pone.0073943

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

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

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

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

491 Traverse CC, Ochman H. 2018. A Genome-Wide Assay Specifies Only GreA as a Transcription Fidelity
492 Factor in *Escherichia coli*. G3: Genes, Genomes, Genetics. 8:2257-2264.

493 Vicario S, Moriyama EN, Powell JR. 2007. Codon usage in twelve species of Drosophila. BMC Evol Biol.
494 7:226. doi: 10.1186/1471-2148-7-226

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

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

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

501 Wilke CO. 2013. Common errors in statistical analyses. The Serial Mentor: August 18 2013. Available
502 from: https://serialmentor.com/blog/2013/8/18/common-errors-in-statistical-analyses

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

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

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

509 Zhang D, et al. 2016. The Effect of Codon Mismatch on the Protein Translation System. PLoS ONE.
510 11:e0148302. doi: 10.1371/journal.pone.0148302

511