





High Transcriptional Error Rates Vary as a Function of Gene Expression Level

Kendra M. Meer ^{1,3,†}, Paul G. Nelson ^{1,†}, Kun Xiong ^{2,4}, and Joanna Masel ^{1,*}

¹Department of Ecology & Evolutionary Biology, University of Arizona

²Department of Molecular & Cellular Biology, University of Arizona

³Present address: Computational Bioscience Program, University of Colorado Anschutz Medical Campus, Aurora, CO

⁴Present address: Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT

†These authors contributed equally to this work.

*Corresponding author: E-mail: masel@email.arizona.edu.

Accepted: December 12, 2019

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, for example *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 with *E. coli* non-C→U errors, are biological.

Key words: nearly neutral theory, weak selection, Cir-Seq, RNA editing, phenotypic mutation.

Introduction

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 (Lynch 2007; Good and Desai 2014). 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, because 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 nonfunctional 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 (Petrov and Hartl 2000; Lynch and Marinov 2015). Second, there is the opportunity cost of not using ribosomes to make other gene products (Dekel and Alon 2005; Scott et al. 2014; Kafri et al. 2016). 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, whereas 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 *S. 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 than that of *Buchnera aphidicola* (4.7×10^{-5}) (Traverse and Ochman 2016b). *Buchnera* is a highly mutationally degraded species in which the drift barrier is an obstacle to the maintenance of fidelity in many other important cellular functions (McCutcheon and Moran 2012); this high rate in *Buchnera* may thus indicate that the drift barrier forms an obstacle even to global solutions (Xiong et al. 2017). All these error rates except for that of *C. elegans* (Gout et al. 2013), were estimated using Cir-Seq (Acevedo and Andino 2014), and should therefore be comparable, although sample preparation techniques differ in vulnerability to deamination.

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

Here, we test whether selection is able to maintain locally lower transcriptional error rates in highly expressed genes.

Selection to purge deleterious mutations is generally more effective in highly expressed genes, as evidenced, for example, by stronger codon bias (Duret and Mouchiroud 1999; Cutter and Charlesworth 2006; Sharp et al. 2010; Ran et al. 2014), which lowers translational error rates (Zhang et al. 2016). Somatic mutations (Frigola et al. 2017), alternative transcriptional start sites (Xu et al. 2019), post-transcriptional modifications (Liu and Zhang 2018a, 2018b), alternative mRNA polyadenylation (Xu and Zhang 2018), and translation errors (Mordret et al. 2019) also occur at lower rates at sites where they are likely to have larger effects. We similarly predict that because high mistranscription rates matter more for highly expressed genes, highly expressed genes should evolve a lower rate of mistranscription. We make this prediction for *E. coli*, where mistranscription rates are globally high and thus so is local selection pressure. In contrast, we do not expect a relationship between expression level and mistranscription rate in *S. cerevisiae*, where mistranscription rates are globally much lower.

Results and Discussion

Mistranscription rate data in *E. coli* were taken from Traverse and Ochman (2016a), who used Cir-Seq (Acevedo and Andino 2014) to distinguish mistranscription events from sequencing errors. Within the largest and highest-quality batch of their data (see Materials and Methods section), data from four experimental conditions (minimal vs. rich media, and midlog vs. stationary phase) were sometimes analyzed separately and sometimes pooled. Mistranscription rates are much higher for C→U substitutions: $\sim 10^{-4}$ rather than $\sim 10^{-5}$ for other mistranscription types. Because C→U changes are more sensitive to preparation artifacts (Chen et al. 2014), that is they may not be mistranscription errors, we excluded them from most of our analysis.

To further ensure the data quality, we exclude “hotspot” nucleotide sites experiencing significantly ($P < 10^{-9}$) more errors of one type than expected from our model fitted as described below. This eliminates recent mutations, inaccurate mapping of reads to the genome, or other artifacts of the experiment or pipeline, as well as any sites subject to programmed post-transcriptional RNA editing. We excluded 5 protein-coding and 2,390 noncoding sites that met this “hotspot” criteria for at least one experimental condition. The high rate of apparent mistranscription hotspots in noncoding genes has been interpreted (Traverse and Ochman 2016a) as a consequence of *E. coli* having multiple polymorphic rRNA operons, making mapping of reads inaccurate. We therefore restrict our analysis to protein-coding genes.

We modeled the number of errors observed per nucleotide site as count data, using a generalized linear model. The number of errors expected is the product of the number of observations of that nucleotide site, and the modeled mistranscription rate, the latter a linear function of log protein

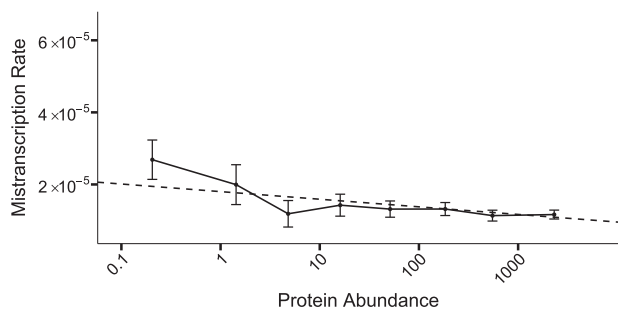


Fig. 1.—Highly expressed *E. coli* genes are subject to lower mistranscription rates. The dashed line shows the equation (1) model applied to the 11 non-C→U substitution types, in which both condition and substitution type affect the intercept but not the slope, plotted as a weighted average over conditions 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 shows the pooled data, binned by protein abundance as described in the Materials and Methods section, and plotted according to mean protein abundance and the mean and 95% CI of the mistranscription rate within each bin. Data were divided into ten 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.

abundance, experimental condition, and substitution type (see Materials and Methods section). The dependence on protein abundance (fig. 1; slope of 0 rejected from eq. 1 model with $P = 2 \times 10^{-14}$) supports our prediction from drift barrier theory, a result that gets slightly stronger if we omit our hot-spot removal procedure. The 11 non-C→U substitution types have substantially different mistranscription rates (supplementary fig. S1, Supplementary Material online); fitting different intercepts for each type (while leaving their slopes the same) is strongly supported for inclusion in our equation (1) model ($P = 2 \times 10^{-16}$).

Different intercepts for different experimental conditions are also supported, in addition ($P = 1.5 \times 10^{-3}$). Fitting different slopes for each experimental condition only marginally improves the fit relative to our equation (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 (supplementary fig. S2, Supplementary Material online).

Standing out from results on all non-C→U error types in supplementary figure S1, Supplementary Material online, and shown in figure 2, is the fact that G→A errors depend more strongly on protein abundance than other error types do ($P = 3 \times 10^{-4}$, eq. 2 as improvement on eq. 1). A separate model fit to G→A error data only, gives a slope of -2.9×10^{-6} (95% CI of -1.6×10^{-6} to -4.2×10^{-6}) with \log_{10} protein abundance, that is there are 1.6–4.2 fewer G→A errors per million G transcription events per 10-fold increase in expression, against a background of about 20–

40 errors per million G transcription events. To ensure that the nonzero slope of figure 1 is not driven solely by G→A errors, we repeated the analysis for the ten error types, that is excluding both C→U and G→A (fig. 2, right). This yields a slope of -8.4×10^{-7} ($P = 1 \times 10^{-7}$) with \log_{10} protein abundance, with a 95% confidence interval (CI) corresponding to 0.4 and 1.2 fewer expression errors per million opportunities per 10-fold increase in expression.

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

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

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

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 nonzero slope on a C→U equivalent of eq. 2). This confirms that the protocol of Gout et al. (2017) succeeded in avoiding

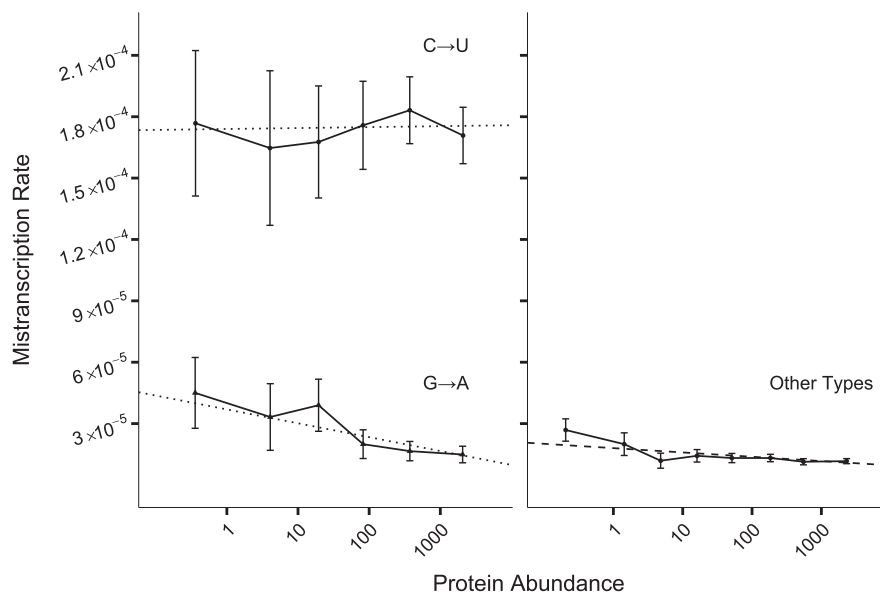


FIG. 2.—C→U errors in *E. coli* are mostly artifacts, G→A depend most strongly on protein abundance, but the other ten error types also show dependence. Dotted lines (left) show linear models with both the slope and intercept fitted separately for each error type using data pooled across all four conditions; for a comparison of all 12 error types, see [supplementary figure S1, Supplementary Material](#) online. The C→U slope is not different from 0 ($P = 0.91$). The dashed line (right) shows an equation (1) model in which the slope is the same across all ten 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 Materials and Methods section, plotted according to mean protein abundance within each bin; error bars show 95% CI. Data were divided into eight 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.

deamination events during sample preparation (which should not depend on protein abundance), where that of *Traverse* and *Ochman* (2016a) did not.

Discussion

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

The dependence of the *E. coli* mistranscription rate on the strength of selection (as reflected by protein abundance), but not the *S. cerevisiae* mistranscription rate, is consistent with proposed drift barrier explanations (*Rajon* and *Masel* 2011; *McCandlish* and *Plotkin* 2016; *Xiong* et al. 2017). In particular, *E. coli* is smaller and is generally accepted to have a larger effective population size than *S. cerevisiae*. *Escherichia coli* also has fewer loci, occurring within 4,453 genes in K-12 (*Riley* 2006) compared with 5,178 genes in *S. cerevisiae* (*Engel* et al. 2014), which makes it easier to evolve robustness at each one. What is more, the average *E. coli* mRNA produces about 540 proteins out of a total of 2.5×10^6 per cell (*Lu* et al. 2007), that is 0.02% of the proteome, which is twice as much as the average yeast mRNA producing 5,600 proteins out of a total of 5×10^7 per cell (*Lu* et al. 2007), that is 0.01% of the proteome. Although a typical yeast mRNA has a longer half-life and so makes proteins over a longer time (6.7 vs. 27.4 min; *Siwiak* and *Zielenkiewicz* 2013), the magnitude of this should not be enough to counteract all other factors making local solutions easier to evolve in *E. coli*.

We have shown that local mistranscription rates vary in a systematic way on a per-gene basis, but have not determined the mechanisms by which expression error rates vary. Mistranscription rates are affected by local sequence characteristics such as long mononucleotide repeats (*Ackermann* and *Chao* 2006; *Gu* et al. 2010) and at the gene level by

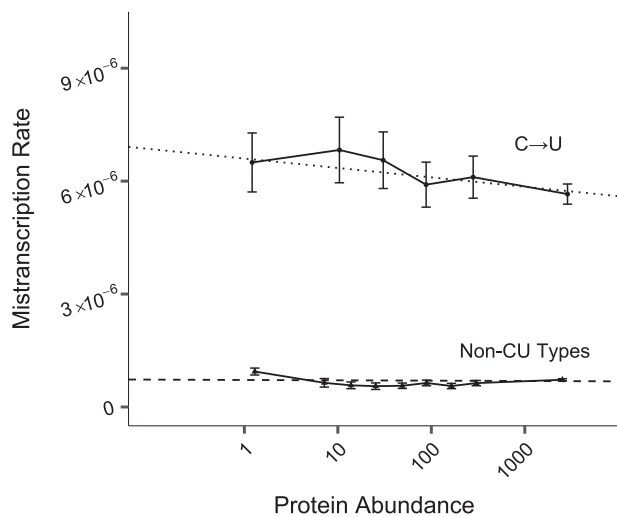


Fig. 3.—In *S. cerevisiae*, only C→U mistranscription errors depend on protein abundance. Dashed line shows a linear model fitted to a pooled data set of the 11 non-C→U substitution types. Dotted line shows a linear model fitted to C→U data alone. To display non-C→U model fit, we took a weighted average 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 Materials and Methods section, 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 eight 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 ten 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 figure S3, Supplementary Material](#) online.

the presence or absence of specific RNA polymerase subunits (Thomas et al. 1998; Walmacq et al. 2009) or transcription factors (Irvin et al. 2014; Roghanian et al. 2015; Bubunencko et al. 2017). Our finding that G→A errors depend more strongly on expression than do other error types in *E. coli* suggests that *GreA*, which specifically reduces G→A transcription errors (Traverse and Ochman 2018), may be a likely mechanistic candidate.

We have also shown that the local mistranscription rates even of highly expressed *E. coli* genes are higher than the global mistranscription rate in *S. cerevisiae*, suggesting that *E. coli* genes are somehow more robust to the consequences of mistranscription than are *S. cerevisiae* genes. However, the robustness associated with *E. coli*'s global solution is not so complete as to eliminate selection for locally lower mistranscription rates in the genes subject to the strongest selection, leading to the trend detected here.

Materials and Methods

Scripts used in these analyses are available at <https://github.com/MasellLab/Meer-et-al-Transcriptional-Error-Rates>

Escherichia coli Mistranscription Data

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

Traverse and Ochman (2016a) data were obtained in multiple batches (referred to as “replicates” in their data tables), with results reported only on two of the batches. Batch no. 2 had approximately half as much data and twice the error rate of batch no. 1, so we restrict our analysis to batch no. 1 only. Combining the data from each of the four experimental conditions (minimal vs. rich media, and midlog vs. stationary phase) within batch no. 1 effectively yielded 15,742,204 protein-coding sites and 189,376 noncoding sites, where “site” is used here as shorthand for condition × nucleotide locus, that is to describe the set of reads of a nucleotide locus within just one experimental condition.

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

Mutations occurring during the Cir-Seq experiment, inaccurate mapping of reads to the genome, or other artifacts of the experiment or pipeline can result in the appearance of mistranscription “hot spots” that are best removed. We calculated the likelihoods of seeing that many or more errors for each of the 18,530,088 possible errors being analyzed, using a significance cutoff of 10^{-9} to ensure that only $10^{-9} \times 18,530,088 = 0.02$ possible errors are falsely excluded, or potentially more if there is genuine biological variation in mistranscription rates beyond that captured by our linear model. We calculated likelihoods from a cumulative binomial distribution based on the number of reads at that site and the rate of error expected at that site from our model. When a possible error was excluded with likelihood $< 10^{-9}$, we excluded the entire nucleotide locus (i.e. all three possible substitutions in all four conditions). We performed an iterative

procedure, first fitting a model of constant error rate for all non-C→U errors and a separate error rate for C→U errors, using expectations from this model to exclude outliers, then using the cleaned-up data to develop a more sophisticated error rate model of all conditions/substitution types, and using the revised expectations from this model to update which loci should be excluded etc. until convergence. In the final iteration, one or more possible errors were determined to be an outlier at 5 protein-coding and 2,390 noncoding loci. For protein-coding outliers, we excluded all possible errors at each of the 5 outlier loci, that is up to 60 possible errors (3 possible errors at 5 loci in 4 conditions). Some sites had no transcript reads in some conditions, resulting in only 48 rather than 60 possible errors being excluded by this procedure, leaving 17,983,341 possible errors in protein-coding transcript regions for analysis. Excluding C→U substitutions, due to their significantly higher error rate and likelihood of occurring post-transcriptionally, further reduced this to 16,466,559 non-C→U possible errors for analysis.

Saccharomyces cerevisiae Mistranscription Data

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

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

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

Protein Abundance Data

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

GroEL Client Status

We labeled the 1,929,741 possible errors associated with 252 *E. coli* proteins as having GroEL client status, based on the

identification of those proteins by Kerner et al. (2005) as specific interactors with the GroEL chaperonin.

Statistical Model

We modeled the error rate at site i within gene j as a linear function of the log-abundance of protein j , that is

$$\frac{E_i}{R_i} = \rho + \beta \ln(\text{Abundance}_j),$$

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

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

$$E_i \sim R_i + R_i \log_{10}(\text{Abundance}_j) + \varepsilon_{\text{Poisson}}.$$

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

$$E_i \sim \text{type} : R_i + \text{cond} : R_i + R_i \log_{10}(\text{Expression}_j) + \varepsilon_{\text{Poisson}}. \quad (1)$$

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

$$E_i \sim \text{type} : R_i + \text{cond} : R_i + GA : R_i \log_{10}(\text{Expression}_j) + \varepsilon_{\text{Poisson}}. \quad (2)$$

P -values associated with adding or removing terms to equation (1) or (2) models were obtained using the ANOVA command with the χ^2 option to compare nested models in R , as given throughout the text, sometimes manually correcting the number of degrees of freedom.

Data Binning

We binned data by protein abundance for visualization and comparison with the fitted models. All possible errors were sorted by the abundance value of the corresponding protein. Bin boundaries were evenly spaced along our log-abundance axis between the 5% quantile and the 95% quantile, with data beyond these quantiles included in the edge bins. For each bin, one point was plotted with y -value equal to the mean and 95% CI of the mistranscription rate and an x -value equal to the geometric mean of protein abundance. The number of mistranscription errors observed is expected to follow a

binomial distribution with r trials, each with probability p of an error. We thus estimated a standard error of $\sqrt{(1 - \hat{p})\hat{p}/r}$, where r is the total number of reads within the bin and \hat{p} is the observed error frequency within the bin. To generate the 95% CI we multiplied this standard error by 1.96. To keep standard errors for low-abundance bins reasonably low, data from several low-abundance bins were combined.

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

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by the Undergraduate Biology Research Program at the University of Arizona, the John Templeton Foundation (39667, 60814), and the National 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.

Literature Cited

- Acevedo A, Andino R. 2014. Library preparation for highly accurate population sequencing of RNA viruses. *Nat Protoc.* 9(7):1760–1769.
- Ackermann M, Chao L. 2006. DNA sequences shaped by selection for stability. *PLoS Genet.* 2(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(1):179–187.
- Chen G, et al. 2014. Cytosine deamination is a major cause of baseline noise in next-generation sequencing. *Mol Diagn Ther.* 18:587–593.
- Cutter AD, Charlesworth B. 2006. Selection intensity on preferred codons correlates with overall codon usage bias in *Caenorhabditis remanei*. *Curr Biol.* 16(20):2053–2057.
- Dekel E, Alon U. 2005. Optimality and evolutionary tuning of the expression level of a protein. *Nature* 436(7050):588–592.
- Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet.* 10:715–724.
- 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(8):4482–4487.
- Engel SR, et al. 2014. The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *Genome Res.* 24(3):389–398.
- Frigola J, et al. 2017. Reduced mutation rate in exons due to differential mismatch repair. *Nat Genet.* 49(12):1684–1692.
- 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(2):680–685.
- Good BH, Desai MM. 2014. Deleterious passengers in adapting populations. *Genetics* 198(3):1183–1208.
- Gout J-F, et al. 2013. Large-scale detection of in vivo transcription errors. *Proc Natl Acad Sci U S A.* 110(46):18584–18589.
- Gout J-F, et al. 2017. The landscape of transcription errors in eukaryotic cells. *Sci Adv.* 3(10):e1701484.
- Gu T, et al. 2010. Avoidance of long mononucleotide repeats in codon pair usage. *Genetics* 186(3):1077–1084.
- Irvin JD, et al. 2014. A genetic assay for transcription errors reveals multilayer control of RNA polymerase II fidelity. *PLoS Genet.* 10(9):e1004532.
- Kafri M, MetzI-Raz E, Jona G, Barkai N. 2016. The cost of protein production. *Cell Rep.* 14(1):22–31.
- Kerner MJ, et al. 2005. Proteome-wide analysis of chaperonin-dependent protein folding in *Escherichia coli*. *Cell* 122(2):209–220.
- Liu Z, Zhang J. 2018a. Human C-to-U coding RNA editing is largely non-adaptive. *Mol Biol Evol.* 35(4):963–969.
- Liu Z, Zhang J. 2018b. Most m⁶A RNA modifications in protein-coding regions are evolutionarily unconserved and likely nonfunctional. *Mol Biol Evol.* 35(3):666–675.
- Lu P, et al. 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol.* 25(1):117–124.
- Lynch M. 2007. *The origins of genome architecture*. Sunderland: Sinauer Associates.
- Lynch M, Marinov GK. 2015. The bioenergetic costs of a gene. *Proc Natl Acad Sci U S A.* 112(51):15690–15695.
- McCandlish DM, Plotkin JB. 2016. Transcriptional errors and the drift barrier. *Proc Natl Acad Sci U S A.* 113(12):3136–3138.
- McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol.* 10(1):13–26.
- Mordret E, et al. 2019. Systematic detection of amino acid substitutions in proteomes reveals mechanistic basis of ribosome errors and selection for translation fidelity. *Mol Cell* 75(3):427–441.
- Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* 246(5428):96–98.
- Petrov DA, Hartl DL. 2000. Pseudogene evolution and natural selection for a compact genome. *J Hered.* 91(3):221–227.
- Rajon E, Masel J. 2011. The evolution of molecular error rates and the consequences for evolvability. *Proc Natl Acad Sci U S A.* 108(3):1082–1087.
- Ran W, Kristensen DM, Koonin EV. 2014. Coupling between protein level selection and codon usage optimization in the evolution of bacteria and Archaea. *mBio* 5(2):e00956–14.
- Riley M. 2006. *Escherichia coli* K-12: a cooperatively developed annotation snapshot—2005. *Nucleic Acids Res.* 34(1):1–9.
- Roghianian M, Zenkin N, Yuzenkova Y. 2015. Bacterial global regulators DksA/ppGpp increase fidelity of transcription. *Nucleic Acids Res.* 43(3):1529–1536.
- Scott M, Klumpp S, Mateescu EM, Hwa T. 2014. Emergence of robust growth laws from optimal regulation of ribosome synthesis. *Mol Syst Biol.* 10(8):747.
- Sharp PM, Emery LR, Zeng K. 2010. Forces that influence the evolution of codon bias. *Phil Trans R Soc B* 365(1544):1203–1212.
- Siwiak M, Zielenkiewicz P. 2013. Transimulation—protein biosynthesis web service. *PLoS One* 8(9):e73943.
- Thomas MJ, Platas AA, Hawley DK. 1998. Transcriptional fidelity and proofreading by RNA polymerase II. *Cell* 93(4):627–637.
- Tomala K, Korona R. 2013. Evaluating the fitness cost of protein expression in *Saccharomyces cerevisiae*. *Genome Biol Evol.* 5(11):2051–2060.
- Traverse CC, Ochman H. 2016a. Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *Proc Natl Acad Sci U S A.* 113(12):3311–3316.
- Traverse CC, Ochman H. 2016b. Correction for Traverse and Ochman, Conserved rates and patterns of transcription errors across bacterial

- growth states and lifestyles. *Proc Natl Acad Sci U S A*. 113:E4257–4258.
- Traverse CC, Ochman H. 2018. A genome-wide assay specifies only *GreA* as a transcription fidelity factor in *Escherichia coli*. *G3*. 8:2257–2264.
- Vicario S, Moriyama EN, Powell JR. 2007. Codon usage in twelve species of *Drosophila*. *BMC Evol Biol*. 7(1):226.
- Wagner A. 2007. Energy costs constrain the evolution of gene expression. *J Exp Zool B* 308B(3):322–324.
- Walmacq C, et al. 2009. Rpb9 subunit controls transcription fidelity by delaying NTP sequestration in RNA polymerase II. *J Biol Chem*. 284(29):19601–19612.
- Wang M, et al. 2015. Version 4.0 of PaxDB: protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* 15(18):3163–3168.
- Wilke CO. 2013. Common errors in statistical analyses. *The Serial Mentor*: August 18 2013. Available from: <https://serialmentor.com/blog/2013/8/18/common-errors-in-statistical-analyses>; last accessed December 20, 2019.
- Xiong K, McEntee JP, Porfirio DJ, Masel J. 2017. Drift barriers to quality control when genes are expressed at different levels. *Genetics* 205(1):397–407.
- Xu C, Park J-K, Zhang J. 2019. Evidence that alternative transcriptional initiation is largely nonadaptive. *PLoS Biol*. 17(3):e3000197.
- Xu C, Zhang J. 2018. Alternative polyadenylation of mammalian transcripts is generally deleterious, not adaptive. *Cell Syst*. 6(6):734–742.
- Zhang D, et al. 2016. The effect of codon mismatch on the protein translation system. *PLoS One* 11(2):e0148302.

Associate editor: George Zhang