The Importance of Nonsense Errors: Estimating the Rate and Implications of Drop-Off Errors during Protein Synthesis

Alexander L. Cope[1,2,3,+,*], Denizhan Pak[4], and Michael A. Gilchrist[5,6,*]

[1]Department of Genetics, Rutgers University, Piscataway, New Jersey, United States
[2]Human Genetics Institute of New Jersey, Rutgers University, Piscataway, New Jersey, United States
[3]Robert Wood Johnson Medical School, Rutgers University, Piscataway, New Jersey, United States
[+]Current: Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee, United States
[4]Luddy School of Informatics, Indiana University, Bloomington, Indiana, United States
[5]Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, Tennessee, United States
[6]National Institute of Mathematical and Biological Synthesis, University of Tennessee, Knoxville, Tennessee, United States
[*]Email: mikeg@utk.edu, alexander.cope@vanderbilt.edu

September 5, 2024

Abstract

The process of mRNA translation is both energetically costly and relatively error-prone compared to transcription and replication. Nonsense errors during mRNA translation occur when a ribosome drops off a transcript before reaching a stop codon, resulting in energetic investment in an incomplete and likely non-functional protein. Nonsense errors impose a potentially significant energy burden on the cell, making it critical to quantify their frequency and energetic cost. Here, we present a model of ribosome movement for estimating protein production, elongation, and nonsense error rates from high-throughput ribosome profiling data. Applying this model to an exemplary ribosome profiling dataset in *S. cerevisiae*, we find that nonsense error rates vary between codons, in conflict with the general assumption of uniform rates across sense codons. Using our parameter estimates, we find multiple lines of evidence that selection against nonsense errors is a prominent force shaping coding-sequence evolution, including that nonsense errors place an energetic burden on cells comparable to ribosome pausing. Our results indicate greater consideration should be given to the impact of nonsense errors in shaping coding-sequence evolution.

Introduction

Protein synthesis, i.e., the transcription and translation of mRNAs, is one of the most energetically costly cellular processes (Wagner, 2005). For example, 66% of ATP is dedicated to protein synthesis in *E. coli* (Russell and Cook, 1995). While both transcription and translation are critical to protein synthesis, the cost

of translation is estimated to be 125 times the cost of transcription (Lynch and Marinov, 2015). Translation involves both direct (GTPs used to charge tRNAs and drive each elongation cycle) and indirect costs (overhead cost of the translation infrastructure such as the ribosome and tRNAs) (Li, 1987, Bulmer, 1991, Gilchrist, 2007, Gilchrist *et al.*, 2015). Furthermore, mRNAs generally undergo multiple rounds of translation. Given the energetic cost of mRNA translation and the finite energy budget of a cell, genotypes reducing the cost of protein production are expected to have a selective advantage that scales with a gene's average protein production rate. A classic example of this is the bias of highly expressed genes toward synonymous codons recognized by more abundant tRNA is hypothesized to be an adaptation due to natural selection to reduce the indirect cost of ribosome pausing (Bulmer, 1991, Shah and Gilchrist, 2011, Gilchrist *et al.*, 2015, Frumkin *et al.*, 2018).

In addition to being energetically costly, mRNA translation is relatively error-prone compared to DNA replication and transcription (Drummond and Wilke, 2009) and translation errors potentially result in non-functional proteins. Much of the literature has focused on the impact of codons being misread by the ribosome, resulting in the insertion of the incorrect amino acid into the growing peptide chain (i.e., a missense error) (Akashi, 1994, Drummond *et al.*, 2006, Drummond and Wilke, 2008, Mordret *et al.*, 2019). The incorrect amino acid sequence may result in a misfolded protein that is potentially corrected via interactions with chaperone proteins (Warnecke and Hurst, 2010, 2011). In contrast, a nonsense error (a.k.a. premature termination errors or processivity errors) results in an incomplete protein (Kurland, 1992, Eyre-Walker, 1996). As an incomplete protein is expected to have little if any functional capacity and is unlikely to be rescued via chaperones, nonsense errors may contribute substantially to the overall cost of mRNA translation (Gilchrist, 2007). Based on work in *E. coli* and *S. cerevisiae*, nonsense errors are estimated to occur at a frequency of approximately 1 every 10,000 codons (Tsung *et al.*, 1989, Jørgensen *et al.*, 1993, Sin *et al.*, 2016, Duc and Song, 2018, Awad *et al.*, 2024).

A simple exponential waiting time model of ribosome elongation where two possible outcomes are a nonsense error or successful elongation implies the probability of a nonsense error at a given codon depends on both its nonsense error rate (i.e., the background rate at which a nonsense error event is triggered) and its elongation rate. It is well-established that ribosome elongation rates generally vary across codons, often due to differences in the availability of the corresponding tRNA and wobble effects (dos Reis *et al.*, 2004, Kramer *et al.*, 2010, Joshi *et al.*, 2018, Weinberg *et al.*, 2016, Chou *et al.*, 2017, Wu *et al.*, 2019). As a result, slower codons are expected to have greater nonsense error probabilities even if nonsense error rates are uniform across codons (Gilchrist and Wagner, 2006, Gilchrist, 2007, Yang *et al.*, 2018). Although the

assumption of uniform nonsense error rates is likely unrealistic, theoretical studies making this assumption while still allowing ribosome elongation rates to vary between codons predict the probability of a ribosome completing translation varies substantially across genes as a function of amino acid sequence, protein length, and codon usage bias (Gilchrist and Wagner, 2006, Gilchrist *et al.*, 2009, Bonnin *et al.*, 2017). Importantly, studies in both *E. coli* and *S. cerevisiae* indicate sense codons differ in their capacity to pair with release factors (Freistroffer *et al.*, 2000, Chavatte *et al.*, 2003), often by mimicking the structural mechanism of stop codon recognition upon release factor binding (Svidritskiy *et al.*, 2018). This limited biochemical evidence suggests nonsense error rates differ across codons which could amplify or dampen the differences in codon-specific elongation probabilities. If codons vary in their nonsense error probabilities, then it follows that protein-coding sequences can evolve to reduce the cost of nonsense errors via codon usage. As before, the selective advantage for synonymous genotypes reducing the cost of nonsense errors should increase with a gene's average protein production rate. Thus, we expect to see increasing evidence of adaptation in codon usage bias to reduce the cost of protein production with a gene's average protein production rate and a lineage's effective population size N_e (Charlesworth, 2009).

Ribosome profiling or Ribo-seq is a powerful technique for quantifying translation at codon-level resolution across the transcriptome (Ingolia *et al.*, 2009). However, ribosome profiling does not allow for directly tracking ribosome movement along a transcript, requiring even simple models to extract information from the sequenced reads. Previous attempts to quantify nonsense errors from ribosome profiling data have typically averaged over codons and genes, ignoring variation in elongation, nonsense, and protein production rates (Sin *et al.*, 2016, Awad *et al.*, 2024). To explicitly account for this variation, we extend our previous model of nonsense errors and ribosome movement along an mRNA transcript (Gilchrist and Wagner, 2006, Gilchrist, 2007) to estimate codon-specific elongation and nonsense error rates using a high-quality ribosome profiling dataset in *S. cerevisiae* (Weinberg *et al.*, 2016). Using our estimates, we evaluate and contextualize the estimated distribution and costs of nonsense errors across the *S. cerevisiae* transcriptome. Perhaps unsurprisingly, we find compelling evidence that nonsense error rates and probabilities differ across codons and amino acids. As genes differ in amino acid usage, protein length, and codon usage, our results indicate that the probability a ribosome completes translation varies across CDS (interquartile range $0.87 - 0.95$). Using simulation, we identify multiple lines of evidence that the yeast genome shows extensive adaptation to reduce the cost of nonsense errors. Our evidence includes a bias towards nonsense error-prone codons towards the 5'-ends of transcripts and an avoidance of nonsense error-prone codons in highly expressed genes. Consistent with previous theoretical studies relying upon tRNA-based proxies of elongation rates (Gilchrist

and Wagner, 2006, Gilchrist *et al.*, 2009), we find approximately 60% of genes in *S. cerevisiae* exhibit signals of adaptation to reduce the cost of nonsense errors. Despite these adaptions to reduce the cost of nonsense errors, we estimate nonsense errors impose an energetic burden comparable to, if not greater than, ribosome pausing. Although nonsense errors are believed to have a lower probability than missense errors, because a preponderance of nonsense errors, but only a fraction of missense errors, are likely to disrupt a protein's functionality, our work suggests natural selection against nonsense errors plays a substantial, important, and underappreciated role in protein sequence evolution.

Materials and Methods

We are interested in calculating the probability of observing a ribosome footprint (RFP) (i.e., a mapped sequencing read) at a codon within a transcript as measured via ribosome profiling. We assume there is a pool of RFP generated from the transcriptome and that the mRNAs in this pool are close to steady-state in terms of ribosome initiation and completion of translating a transcript. Below we give an overview of our model formulation and usage, with additional details found in Supplemental Materials. Definitions of all model parameters can be found in Table 1. After defining and fitting our model, we test if, how, and why nonsense error rates vary across codons and amino acids. We assess how well these parameters recapitulate the sharp drop in ribosome densities at the 5'-end of the transcript, often referred to as the "5'-ramp region." Next, we examine how variation in nonsense error probabilities and codon usage shape the probabilities that a ribosome completes translation across the transcriptome. Given that synonymous codons vary in their probability of triggering a nonsense error, we test if variation in codon usage within and across genes is consistent with natural selection against nonsense errors. Finally, we determine if observed codon usage patterns in genes reduce the energetic burden of nonsense errors and how these costs compare to the direct and indirect costs of peptide elongation.

Ribosome Pausing and Nonsense Error Model

Below we outline the assumptions behind our ribosome PAusing and NonSense Error (PANSE) model and the formal likelihood function they lead to.

PANSE Model Assumptions

Briefly, PANSE makes the following assumptions. For a given gene g, m_g represents its equilibrium mRNA transcript concentration within a cell and κ_g represents the average ribosome initiation rate for each of its mRNAs. Gene g consists of n_g codons and the ribosomes' elongation and nonsense error waiting times at codon position i are $v_{g,i}$ and $w_{g,i}$, respectively. We assume the elongation waiting time $w_{g,i}$ varies between sites such that $w_{g,i} \sim \text{Gamma}(\alpha_i, \lambda_i)$ where the shape parameter α_i and the rate parameter λ_i can vary between the 61 sense codons. Although the nonsense error waiting time $v_{g,i}$ likely varies with codon and position, for simplicity we focus solely on the variation between codons and, thus, treat v_i as a singular parameter value (i.e., is the same value across all codons of type i) rather than a random variable. We expect any variation in v_i that does exist to be absorbed into the variation of $w_{g,i}$. Based on these assumptions, it is possible to model the distribution of footprints as coming from a negative binomial distribution, i.e. $Y_{g,i} \sim \text{NB}(k, p)$ with $k = \alpha_c$, $p = \frac{\kappa_g m_g \,\text{E}[\sigma_g(i-1)]}{\lambda'_c + \kappa_g m_g \,\text{E}[\sigma_g(i-1)]}$, where $\lambda'_c = \lambda_c Z/Y$. The composite parameter λ'_c is a rescaling of the codons specific rate parameter λ_c by the ratio of the partition function for the RFP distribution Z the footprints are sampled from to the total number of observed footprints in a dataset $Y = \sum_{g,i} Y_{g,i}$. Thus, Z/Y is a measure of the sampling efficiency of the experiment.

Given $Y_{g,i}$ represents the observed number of ribosome footprints (RFP) at codon i of gene g, it follows that the likelihood of observing $Y_{g,i}$ is,

$$\text{L}\left(\alpha_c, \lambda'_c, \kappa'_g \middle| Y_g\right) = \prod_{i=1}^{n_c} \frac{\Gamma\left(\alpha_c + Y_{g,i}\right)}{\Gamma\left(\alpha_c\right)} \left(\frac{\kappa'_g \,\text{E}\left[\sigma_g(i-1)\right]}{\lambda'_c + \kappa'_g \,\text{E}\left[\sigma_g(i-1)\right]}\right)^{Y_{g,i}} \left(1 - \frac{\kappa'_g \,\text{E}\left[\sigma_g(i-1)\right]}{\lambda'_c + \kappa'_g \,\text{E}\left[\sigma_g(i-1)\right]}\right)^{\alpha_c} \tag{1}$$

where $\kappa'_g = \kappa_g \times m_g$ is a rescaling of the mRNA$_g$ specific ribosome initiation rate κ_g by the density of mRNA$_g$ transcripts within the cell, m_g and $f_w(\alpha_i, \lambda_i)$ represents the PDF of the Gamma distribution for the ribosome waiting times codon i. Assuming independence between elongation steps, it follows that the probability of a ribosome successfully elongating at codons 1 through $i-1$ is,

$$\text{E}\left[\sigma_g(i-1)\right] = \prod_{j=1}^{i-1} \left(\int_0^\infty \frac{v_j}{w_{g,j} + v_j} f_w\left(\alpha_j, \lambda_j\right) dw_{g,j}\right). \tag{2}$$

Because the nonsense error waiting time v_i between sense codons is likely to be several orders of magnitude greater than a given $w_{g,i}$ and we will be working with our likelihood function on the log scale, we can greatly speed up our evaluation of the log of Eqn. (1) by approximating $\log\left(\text{E}\left[\sigma_g(i)\right]\right)$ using a 2nd order Taylor

Series approximation around a mean nonsense error rate for all codons $\overline{1/v} = 0$. Doing so gives,

$$\begin{aligned}\log(\mathrm{E}\left[\sigma_g(i-1)\right]) &= \sum_{j=1}^{i}\left[-\frac{\alpha_j}{\lambda_j v_j} + \frac{\alpha_j}{{\lambda_j}^2 v_j^2} + \frac{1}{2}\left(\frac{\alpha_j}{\lambda_j v_j}\right)^2\right] + O\left(\overline{1/v}^3\right) \\ &= \sum_{j=1}^{i}\left[-\frac{\overline{w}_j}{v_j} + \frac{\mathrm{Var}\left(w_j\right)}{v_j^2} + \frac{1}{2}\left(\frac{\overline{w}}{v_j}\right)^2\right] + O\left(\overline{1/v}^3\right).\end{aligned}$$

Where $\overline{w}_j = \alpha_j/\lambda_j$ and $\mathrm{Var}[w_j] = \alpha_j/{\lambda_j}^2$ are the expected waiting time and variance in waiting time for codon j, respectively.

Parameter	Definition	Units
$w_{g,i}$	Ribosome elongation waiting time at codon position i in gene g	1/t
$\overline{w}_c$	Expected ribosome elongation waiting time for codon c	1/t
$v_{g,i}$	Ribosome nonsense error waiting time at codon position i in gene g	1/t
α_c, λ_c	Shape and rate parameter for distribution of waiting times for codon c . The rate parameter is inversely related to average wait time, i.e. for codon c $E[w_{g,i}] = \alpha_c/\lambda_c$	-
m_g	Density of mRNA transcripts for gene g in cytosol.	1/Vol
M_g	Observed mRNA counts from RNA-Seq data.	1/Vol
κ_g	Rate constant determining ribosome initiation rate per mRNA. Function of diffusion of ribosomes, mRNA, and other factors.	$1/t$
ϕ_g	Target protein functionality production rate under a given set of conditions. Equal to $m_g\kappa_g\sigma_g(n_g)$ which is initiation rate discounted by the expected functionality of a protein.	1/t
κ'_g	Rate of initiation of protein synthesis under a given set of conditions. Equal to $m_g\kappa_g$.	1/t
$\sigma_g(i-1)$	Probability ribosome reaches codon i by successfully translating the previous $i-1$ codons of codon g.	
$\mathrm{E}\left[\sigma_g(i-1)\right]$	Expectation of $\sigma_g(i-1)$ used when fitting model to data.	
$p_{g,i}$	Probability a ribosome is found at position i on an mRNA transcript from gene g when translation initiation and completion of mRNA is at steady state.	
$P_{g,i}$	Probability of observing a footprint for position i of mRNA from gene g.	
P_g^c	Probability of observing a footprint for codon c of mRNA from gene g.	
Z	Partition function which scales the codon footprint sampling probabilities $P_{g,i}$ summed across i and g equals 1.	
n_g^c	Number of codons type c in mRNA of gene g.	
$Y_{g,i}$	Number of RFP observed for position i in gene g	
Y_g^c	Number of RFP observed for codon c in gene g	
Y	Totalnumber of RFP in dataset.	
λ'_c	Composite parameter equal to $\lambda_c Z/Y$	
π_j	Prior probability for parameter j.	

Table 1: Table of model parameters.

Analysis of ribosome profiling data

Raw ribosome profiling for *S. cerevisiae* reads were downloaded from the Sequence Read Archive (SRA Run Accession: SRR1049521) (Weinberg *et al.*, 2016). These data were generated using a flash-freeze protocol to halt ribosome elongation, which avoids many of the technical artifacts caused by the use of cycloheximide (Lareau *et al.*, 2014, Weinberg *et al.*, 2016). The raw reads were processed using the **riboviz 2** workflow (Cope *et al.*, 2022). Aligned reads of length 28-30 were extracted and assigned to codons in the A-site assuming a 15 nucleotide offset relative to the 5'-end of the read (Ingolia *et al.*, 2009). A flat file was created that includes the number of ribosome footprints (or counts) at each codon within a gene. This file was used as the primary input into our implementation of PANSE. We employed numerous filtering strategies to remove genes that may violate the assumptions of the model (see Supplementary Material: Filtering of processed ribosome profiling data and Figure S1). This ultimately led to a final dataset with 3,112 genes, approximately 50% of the protein-coding sequences in *S. cerevisiae*.

Fitting the Pausing and Nonsense Error Model

The PANSE model was fit to the codon-level ribosome footprint counts across the approximately 3,000 genes included in the final data set using a Markov Chain Monte Carlo (MCMC) simulation. The MCMC was run for 50,0000 iterations, keeping every 5th iteration to reduce autocorrelation. Convergence was assessed by comparing the results of two separate MCMCs that were started at random points in parameter space. Posterior means and 95% highest density intervals (HDIs) were calculated for each parameter of PANSE based on the MCMC traces. Consistent with our previous work (Gilchrist *et al.*, 2015), gene-specific initiation rates κ'_g are assumed to follow a lognormal distribution with a mean initiation rate of 1. This is accomplished by fixing the mean of the lognormal distribution to be $-s_{\kappa'_g}/2$, where $s_{\kappa'_g}$ is the standard deviation of the lognormal distribution. The elongation waiting time parameters α, λ and NSE rate (i.e., $1/v_c$) are all assumed to have a flat prior with ranges (0,100), (0,100) and (1e-100,1e-1) on the natural scale. We note these are particularly broad distributions, but we wanted to ensure that adequate parameter space could be explored. We fit PANSE (1) assuming uniform NSE rates across codons and (2) allowing NSE rates to vary across codons. To determine if there is support for variation in NSE rates across codons, these two model fits were compared using the Deviance Information Criterion (DIC) (Spiegelhalter *et al.*, 2002).

The ribosome profiling data we used were prepared with a flash-freeze protocol to halt elongation, but there is still an observable increase in ribosome density in the first 200 codons of the transcript. Although the increased density was less so than observed in ribosome profiling measurements made with cycloheximide, it

remains unclear if this increased density reflects true biology or a technical artifact (Weinberg *et al.*, 2016). Indeed, previous work found that nonsense error probabilities estimated from the first 200 codons using the same ribosome profiling data as in this study were biologically implausible, as they would imply that ribosomes hardly ever complete translation (Duc and Song, 2018). As it is plausible that ribosome counts for the first 200 codons are impacted by unknown technical biases, we fit PANSE masking the ribosome counts for this region in the likelihood calculation for each gene. However, we did include these codons in our calculation of the expected probability of elongation up to codon i $E[\sigma_g(i)]$.

We compared parameter estimates of PANSE to (1) empirical data such as mRNA abundances and elongation waiting times (Weinberg *et al.*, 2016), and (2) parameter estimates from the Ribosomal Overhead Cost version of the Stochastic Evolutionary Model of Protein Production Rates (ROC-SEMPPR), which estimates protein production rates ϕ and codon-specific selection coefficients $\Delta\eta$ from codon usage bias patterns (Gilchrist *et al.*, 2015). See previous work regarding parameter estimation with ROC-SEMPPR for *S. cerevisiae* (Gilchrist *et al.*, 2015, Cope and Shah, 2022).

Analyzing variation in nonsense error rates $1/v_i$ across codons

To better understand the variation in nonsense errors (NSE) rates across codons, we performed a linear regression of the posterior mean NSE rates against numerous properties of the codon and amino acid. This includes the identity of the nucleotide at each of the three codon positions (A, G, C, or T, with A as the reference class), the number of stop codons 1 nucleotide mutation away from the codon (0, 1, or 2), and the physicochemical properties of the resulting amino acid (charged, hydrophobic, neutral, other) on each codon. We weighted each codon in the regression by the standard deviations estimated from the MCMC traces.

Quantifying variation in translation completion probabilities $\sigma_g(n_g)$ across genes

Posterior means for elongation waiting times w_i and v_i were used for each of the 61 sense codons to calculate the translation completion probability $\sigma_g(n_g)$. This allowed us to apply this formula to all genes in the *S. cerevisiae* genome, regardless of whether or not it was included in the model fit. We assessed how variation in translation completion probabilities $\sigma_g(n_g)$ varied across genes as a function of length and gene expression measured as mRNA abundances (in units of RPKM) taken from previous work (Weinberg *et al.*, 2016). Additionally, we compared how well our inferences of translation completion probabilities $\sigma_g(n_g)$ made directly from empirical ribosome profiling data compared to theoretical estimates based on simulations from a Totally Asymmetric Exclusion Process (TASEP) model of translation (Bonnin *et al.*, 2017).

Quantifying elongation probability variation within and across genes

As before, $\overline{w}_i$ and v_i represent the expected elongation and nonsense error waiting times at codon i, respectively. Given the fact there are only two possible outcomes, elongation or a nonsense error), it follows that the probability a ribosome successfully elongates the growing peptide chain at codon i is,

$$\Pr[\text{ribosome elongates peptide at codon } i] = \frac{v_i}{\overline{w}_i + v_i}.$$

The probability a ribosome reaches codon i by successfully elongating at the previous $i-1$ codons is,

$$\sigma_{i-1}(\vec{c}) = \prod_{j=1}^{i-1} \frac{v_j}{\overline{w}_j + v_j}$$

Thus, the probability a ribosome reaches and successfully translates the final codon n_g is $\sigma_g(n_g)$, where n_g is the number of codons found in gene g. We note that to calculate $\sigma_g(n_g)$, we use the posterior mean of $\overline{w}_i$

To evaluate how elongation probabilities vary with codon position i, we calculated the average probability a ribosome elongates at a given position as the geometric mean of the elongation probability across genes. We then regressed the log(Position i) against the average elongation probability to quantify how the elongation probability changes as a function of position. To test if the observed slope was greater than expected under the null model of no selection against nonsense errors, we generated 1000 different permuted sets of genes for each of 3 different possible nulls: (1) the synonymous codons of an amino acid were permuted within a gene (randomized by amino acid), (2) amino acids and codons were permuted within a gene (randomized by CDS), and (3) codons were permuted across genes. The slope estimated from the real set of genes was then compared to each of the 3 null distributions, with the p-value= $(k+1)/(1000+1)$, where k is the number of occurrences in which a slope from a permuted set of genes was greater than the slope from the real set of genes. A similar analysis was performed on the real data after binning genes based on mRNA abundances into the lower quartile (low expression), interquartile (moderate expression), and upper quartile (high expression) on the log scale.

Determining 5' vs 3' biases of codon usage

To determine the average position of a codon across genes, all positions within a gene were enumerated and normalized to be relative to the length of the gene, i.e., the relative position of codon i in a gene with n_g codons is i/n_g. In doing so, the positions of all codons are mapped to the range [0,1]. A codon with no

position-specific biases (i.e., is just as likely to be found near the 5'-end as the 3'-end) is expected to have an average position across genes of 0.5.

Calculating the cost of nonsense errors

To better understand the potential importance of nonsense errors in mRNA translation, we calculated the expected cost of producing a complete, functional protein (in terms of the number of hydrolyzed NTPs) E[NTP/protein] from a given transcript that includes the impact of nonsense errors. For simplicity, we exclude the cost of amino acid synthesis from our calculations.

Conceptually, we break the cost of protein synthesis into two overlapping sets of categories: direct vs. indirect costs and fixed vs. variable costs. Direct costs include the NTP used in assembling the small and large ribosome subunits on the mRNA and each elongation step by the ribosome, whereas indirect costs are based on the synthesis cost of the mRNA translation infrastructure. In terms of fixed and variable costs, fixed costs are the direct and indirect costs of translating a protein in the absence of a nonsense error, i.e., costs incurred every round of successful translation. Variable costs are the expected direct and indirect costs of mRNA translation that are wasted when a nonsense error occurs – these costs are variable because they are only incurred if a nonsense error occurs. As a result, the total cost of producing a particular protein E[NTP/protein] is the sum of fixed and variable costs, each of which is comprised of direct and indirect costs.

Using the same underlying model of ribosome movement we used to estimate the elongation waiting times $\overline{w}$ and NSE waiting times v, the expected cost in units of NTP bonds E[NTP/protein] of translating a given

sequence of codons $\vec{c}$ as

$$\begin{aligned}
\text{Fixed Cost} &= \text{Cost of synthesizing a complete protein} \\
&= \overbrace{a_1 + a_2 n_g}^{\text{Fixed Direct Costs}} + \overbrace{C\left(w_0 + \sum_{j=1}^{i-1} \overline{w}_j\right)}^{\text{Fixed Indirect Costs}} \\
\text{Variable Cost} &= \text{Expected Number of NSEs/Complete protein} \times \text{Expected cost of a NSE} \\
&= \frac{1}{\sigma_{n_g}(\vec{c})} \times \sum_{i=1}^{n_g} \left[(a_1 + a_2 i) + C\left(w_0 + \sum_{j=1}^{i-1} \overline{w}_j\right)\right] \sigma_{i-1}(\vec{c}) \frac{\overline{w}_i}{\overline{w}_i + v_i} \\
\text{Total Cost} &= \text{Fixed Cost} + \text{Variable Cost} \\
&= \left[(a_1 + a_2\, n_g) + C\left(w_0 + \sum_{i=1}^{n_g} \overline{w}_i\right)\right] \\
&\qquad + \frac{1}{\sigma_{n_g}(\vec{c})} \sum_{i=1}^{n_g} \left[(a_1 + a_2 i) + C\left(w_0 + \sum_{j=1}^{i-1} \overline{w}_j)\right)\right] \sigma_{i-1}(\vec{c}) \frac{\overline{w}_i}{\overline{w}_i + v_i} \qquad (3)
\end{aligned}$$

Where $a_1 = 4\text{NTP}$ and $a_2 = 4\text{NTP}$ are the direct cost of translation initiation and elongation, respectively in terms of hydrolyzed phosphate bonds (Gilchrist, 2007).

Combining the contribution of direct costs to the fixed and variable costs of protein synthesis yields,

$$\text{Direct Costs} = \overbrace{a_1 + a_2 n_g}^{\text{Fixed Direct Costs}} + \overbrace{\frac{1}{\sigma_{n_g}(\vec{c})} \times \sum_{i=1}^{n_g} (a_1 + a_2 i)\, \sigma_{i-1}(\vec{c}) \frac{\overline{w}_i}{\overline{w}_i + v_i}}^{\text{Variable Direct Costs}} \qquad (4)$$

Similarly, combining the contribution of indirect costs to the fixed and variable costs of protein synthesis yields,

$$\text{Indirect Costs} = \overbrace{C\left(w_0 + \sum_{j=1}^{n_g} \overline{w}_j\right)}^{\text{Fixed Indirect Costs}} + \overbrace{\frac{1}{\sigma_{n_g}(\vec{c})} \times \sum_{i=1}^{n_g} \left(w_0 + \sum_{j=1}^{i-1} \overline{w}_i\right) \sigma_{i-1}(\vec{c}) \frac{\overline{w}_i}{\overline{w}_i + v_i}}^{\text{Variable Indirect Costs}} \qquad (5)$$

The term $w_0 + \sum_{j=1}^{i} \overline{w}_j$ represents the <u>indirect</u> cost of ribosome pausing up to codon i. Because RPF data lacks information on the absolute rate of ribosome elongation, we rescaled our estimates of pausing times such that the average elongation rate across the 61 sense codons was ~ 9.3 codon/sec or, equivalently, $\bar{\overline{w}} =\sim 0.108$ sec/codon. Assuming the average mRNA has a coding sequence of 400 codons, the same average

elongation rate as above, and that at any given time 80% of the ribosome population is engaged in mRNA translation (Weinberg *et al.*, 2016), it follows that the expected value of w_0 is ~ 10.8 sec (see Supplementary Material for more details).

The parameter C converts the indirect cost of ribosome pausing (in units of seconds) into their equivalent costs in terms of NTP and has the units of NTP/sec. Although we are unaware of any empirical estimates of C, we use two different approaches to estimate C that vary by less than one order of magnitude. One estimate of C is based on selection coefficients estimated from a ROC-SEMPPR analysis of the yeast genome and yields $C_{\text{ROC}} = 0.76$ NTPs/sec). The second estimate of C is based on the assembly cost (in NTPs) and average lifespan (in seconds) of the ribosome and yields $C_{t_{1/2}} = 5.5$ NTPs/sec. See Supplementary Materials for detailed descriptions of these calculations. Given these two independent estimates of C are the only ones we have, we will treat C_{ROC} and $C_{t_{1/2}}$ as lower and upper bounds on C, respectively. Thus the average cost of ribosome pausing is approximately 0.108sec/codon $\times\, C = 0.082$ or 0.59NTP/codon for C_{ROC} and $C_{t_{1/2}}$, respectively, suggesting indirect costs are a fraction of the direct cost of 4 NTP/codon.

In addition to using our cost estimates directly, we can also use them to test the hypothesis that the ordering of synonymous codons along a transcript shows evidence of adaptation to reduce the cost of nonsense errors. To do so, we generated a null distribution of the expected cost of mRNA translation for each gene when there is no adaptation to reduce the cost of nonsense errors by permuting the order of its synonymous codon within the coding sequence. These permuted sequences had the exact same amino acid usage, ordering, and codon usage as the sequence observed in the genome, and only differed amongst the ordering of synonyms within each amino acid. For each permutation, we calculated its expected protein production cost using Eq (3). We then compared the mean protein production cost of our population of permutated sequences to that of the protein production cost of the observed sequence. We then scored each observed sequence as having a protein production cost either greater than (0) or less than (1) the expected cost under the null hypothesis. To test whether the observed sequences are biased towards lower costs than expected under the null hypothesis, we performed a binomial test ($H_0 : p = 0.5$). This allowed us to test if the number of sequences with an expected cost less than the mean expected cost across the permuted sequences was greater than expected by chance. Because the gross cost of nonsense errors scales with the protein production rate of a gene, we performed a logistic regression to test if the probability the observed gene has a lower cost than expected increases with gene expression.

Results

PANSE was fit to *S. cerevisiae* ribosome profiling data (Weinberg *et al.*, 2016). We compared parameter estimates to appropriate empirical and theoretical proxies, finding them in good agreement (Figure 1). Both mRNA abundances and tRNA gene copy number (tGCN) are common empirical proxies for translation initiation rates and elongation waiting times, respectively (Gilchrist, 2007, dos Reis *et al.*, 2004, Shah and Gilchrist, 2010, Weinberg *et al.*, 2016, Bonnin *et al.*, 2017). We expected PANSE estimates of gene-specific translation initiation rates κ'_g and codon-specific elongation waiting times w_c (where c is a codon) to be well-correlated with these empirical proxies. Indeed, translation initiation rates κ'_g and independent RNA-Seq-based estimates of mRNA abundances are strongly correlated (Weinberg *et al.*, 2016) (Figure 1A, Spearman rank correlation $R = 0.93, p < 2.2E-16$). PANSE estimates of elongation waiting times w_c and estimates based on tGCNs (including effects of wobble base pairing) are reasonably well-correlated (Figure 1B, Spearman rank correlation $R = 0.53, p = 9.7E-06$).

Theoretical biophysical models rooted in population genetics principles effectively estimate parameters relevant to the evolution of codon usage bias (Gilchrist, 2007, Shah and Gilchrist, 2011, Wallace *et al.*, 2013, Gilchrist *et al.*, 2015). One such model, ROC-SEMPPR, estimates protein production rates ϕ and selection coefficients $\Delta\eta$ solely from genome-wide patterns of codon usage bias. ROC-SEMPPR assumes differences in natural selection on codon usage are due to differences in elongation waiting times $\overline{w}$ between synonyms (Shah and Gilchrist, 2011, Gilchrist *et al.*, 2015). We obtained parameter estimates from a previous application of ROC-SEMPPR to the *S. cerevisiae* protein-coding sequences (Cope and Shah, 2022). We note protein production rates $\phi = \kappa'_g \sigma_g(n_g)$, but ROC-SEMPPR ignores translation errors such that $\sigma_g(n_g) = 1$. As expected, initiation rates κ'_g amd protein production rates ϕ are strongly correlated (Figure 1C, Spearman rank correlation $R = 0.65, p < 2.2E-16$). Similarly, elongation waiting times $\overline{w}$ between synonymous codons and selection coefficient $\Delta\eta$ are strongly correlated (Figure 1D, Spearman rank correlation $R = 0.81, p < 1.5E-08$).

NSE rates vary across codons

Studies often assume a uniform (background) NSE rate $1/v$ (Gilchrist and Wagner, 2006, Gilchrist *et al.*, 2009, Sin *et al.*, 2016, Bonnin *et al.*, 2017), but targeted studies suggest that NSE rates may vary across codons (Freistroffer *et al.*, 2000, Chavatte *et al.*, 2003). We found that the model allowing for variation in NSE rates (not just NSE probabilities) across codons is 227 DIC units better than the model assuming

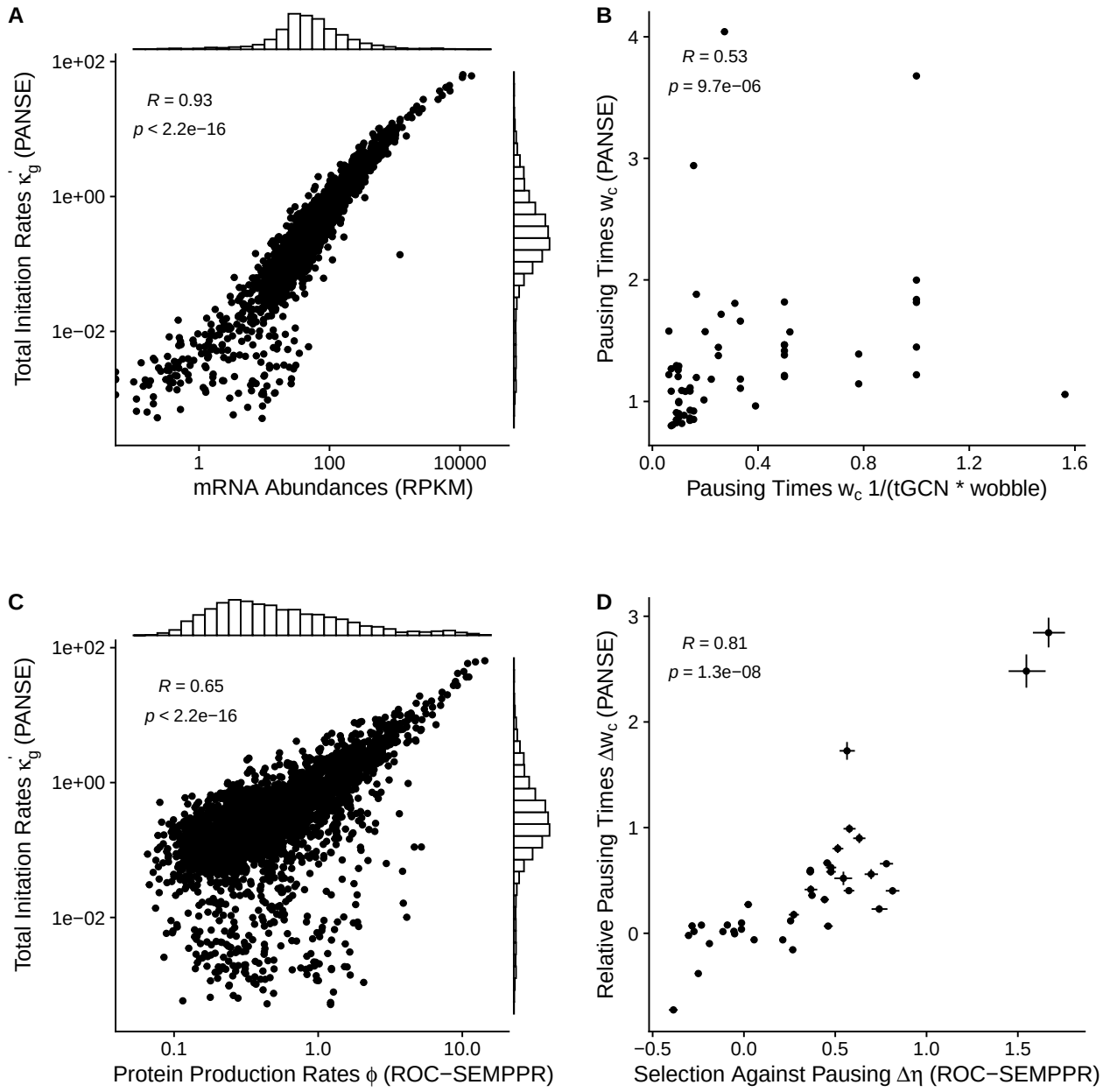

Figure 1: Comparing PANSE gene-specific initiation rates and codon-specific elongation waiting times to empirical and codon-based estimates. Histograms on x and y-axes represent the distributions of the relevant variables. (A) RNA-seq estimates of mRNA abundances (RPKM) (a common proxy for translation initiation rates) and PANSE estimates of initiation rates κ'_g. (B) tRNA gene copy numbers (tGCN) based estimates of elongation waiting times and PANSE estimates of elongation waiting times. (C) ROC-SEMPPR estimates of protein production rates ϕ_g (based on differences in codon usage) and PANSE estimates of initiation rates κ'_g. (D) ROC-SEMPPR estimates of selection coefficients $\Delta\eta$ and relative elongation waiting times Δw_c. Selection coefficients and elongation waiting times were set relative to the alphabetically last codon for each amino acid.

uniform NSE rates, indicating strong support for differences in the NSE rates across codons. NSE rates varied over multiple orders of magnitude, ranging from 8.65×10^{-6} to 2.13×10^{-3} (Figure 2A). The mean NSE probabilities across codons were on the order of 10^{-4}, consistent with previous estimates in *E. coli* and yeast (Tsung *et al.*, 1989, Jørgensen *et al.*, 1993, Sin *et al.*, 2016, Duc and Song, 2018, Awad *et al.*, 2024).

To identify codon and amino acid properties associated with higher NSE rates, we regressed the individual nucleotides at each position (1st, 2nd, or 3rd), the number of stop codon neighbors a sense codon has at each position, and the physicochemical properties of the resulting amino acid for each codon (i.e., whether the amino acid is charged, hydrophobic, neutral, or other) (Figure 2B). We note that we treated the number of stop codon neighbors as a categorical variable, as it is always either 0, 1, or 2. As this is regressing multiple categorical variables against the $\log_{10}$(NSE rate), each coefficient represents a change relative to a reference class defined by the intercept. First, we find that codons with cytosine (C) at the first position of the codon or guanine (G) at the second or third position tend to have higher NSE rates compared to a codon with an adenine (A, the reference class). One might expect that having more stop codon neighbors – the number of stop codons that are a single nucleotide mutation away from a codon – increases the NSE rate of a codon. Indeed, we find that codons with a single stop codon neighbor at the 3rd position of the codon have an NSE rate over 1 order of magnitude greater than codons with no stop codon neighbors at the third position ($\beta = 1.35$, $p = 0.002$). Surprisingly, no such relationship is observed for codons with 2-stop codon neighbors at the third position ($\beta = 0.50$, $p = 0.518$). Even more surprisingly, codons with a single stop codon neighbor at the second position have overall lower NSE rates than codons with no stop codon neighbors at the second position ($\beta = -1.278$, $p = 0.004$). We find no evidence that the physicochemical properties of the amino acid have any impact on the NSE rates of a codon (Figure 2B).

The "5'-ramp" cannot be explained by NSE alone

Ribosome profiling measurements consistently exhibit higher ribosome densities at the 5'-end relative to the remainder of the transcript (Weinberg *et al.*, 2016). We simulated ribosome profiling data based on the PANSE model using the parameters estimated from the real data. We observe a moderate correlation between the real and simulated ribosome counts (Spearman rank correlation $R = 0.56$, $p < 2.2E-16$). On a position-by-position basis, the average log-fold difference between the real and simulated data was approximately -0.007 (One-sample t-test $p = 4.885e-12$, Figure S2), suggesting our model slightly overestimates the number of counts by about 0.7%, on average. We observe clear discrepancies between the metagene profiles of the real and simulated data (Figure 3A). The drop in ribosome densities along the first 200 codons is far more

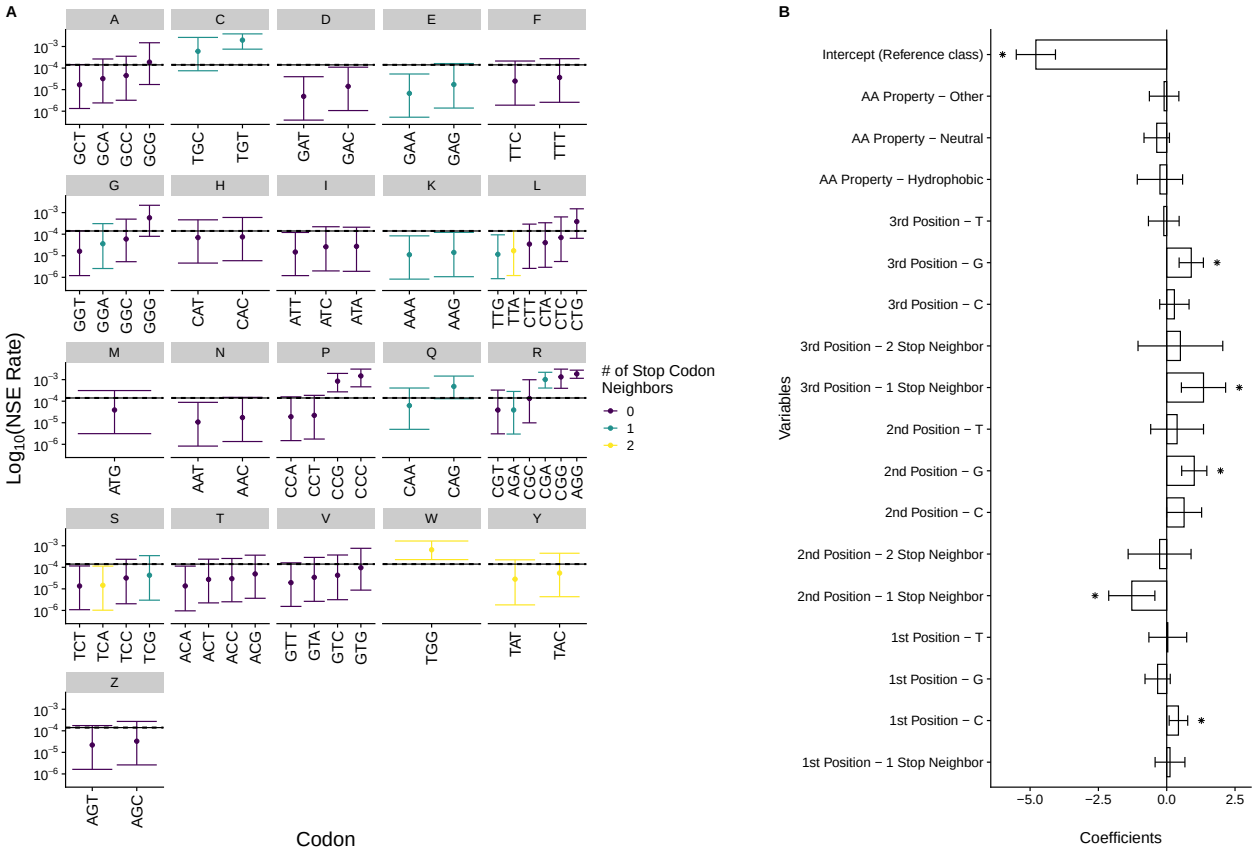

Figure 2: (A) Posterior means and 95% highest density intervals (HDIs) of $\log_{10}$(NSE Rates) estimated for each codon. Colors indicate the number of nucleotide mutations away a codon is from a stop codon. Solid and dashed black lines indicate the NSE Rate posterior mean and 95% HDI for a model fit sharing information across codons (i.e., assuming no variation exists in NSE Rates.) (B) Regression coefficient estimates from a weighted multiple regression of codon properties and the $\log_{10}$(NSE Rates). Variables include nucleotide identity (A, G, C, or T) at each position in the codon, the number of stop codon neighbors at each position, and the physicochemical properties of the resulting amino acid (Charged, Hydrophobic, Neutral, or Other). As we are working with categorical independent variables, the intercept reflects the mean The intercept reflects the mean NSE rate of the reference class of codon Error bars represent 95% confidence intervals. An * indicates statistical significance ($p < 0.05$.).

drastic for the real data, suggesting the 5'-ramp cannot be entirely due to nonsense errors.

The discrepancy between real and simulated data raises the question of what NSE rates are necessary to generate the dramatic drop in ribosome density at the 5'-end. We fit PANSE to only the first 200 codons to determine if the parameters estimated were biologically plausible. Although initiation rates κ'_g and elongation waiting times $\overline{w}$ are in good agreement with estimates from the post-200 codon fits (Figure S3), the NSE rates are significantly greater, with a mean NSE rate of approximately 0.003 across codons. This would lead to an NSE probability of approximately 80% over the first 200 codons, which seems unrealistically high. Taken together, these results indicate the "5'-ramp" in ribosome profiling data is, at best, only partially due

to nonsense errors.

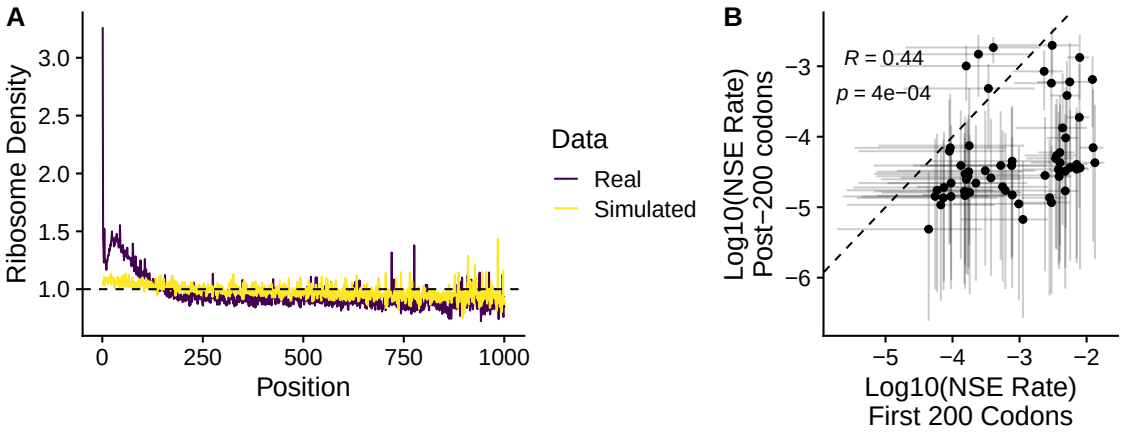

Figure 3: The effects of increased ribosome density at the 5'-end of transcripts. (A) Comparing metagene profiles for real ribosome profiling data (purple) and simulated ribosome profiling data based on the PANSE model (yellow). (B) Comparison of NSE rate estimates from the first 200 codons and the remainder of the gene. Colors indicate the number of single mutations in a codon that could produce a stop codon (i.e., the number of stop codon neighbors).

The probability that translation is completed varies greatly between transcripts

As codons have different NSE probabilities, variation in codon usage across protein-coding sequences is expected to lead to variation in the probability of a ribosome completing translation $\sigma_g(n_g)$. Across all genes, the median probability of a ribosome completing translation $\sigma_g(n_g)$ is approximately 0.92 (interquartile range 0.87 – 0.95). A key factor determining the probability of a ribosome completing translation is the length of a transcript. Even if the probability of an NSE at any given codon is rare, longer transcripts afford more opportunities for an NSE to occur. The length of a transcript and its probability of experiencing a nonsense error $1-\sigma_g(n_g)$ are highly correlated, as expected (Figure 4A, Spearman rank correlation $R = 0.91$, $p < 2.2E-16$). Unsurprisingly, the *S. cerevisiae* proteins YHR099W, YKR054C, and YLR106C are the only transcripts considered with a greater than 50% chance of a ribosome experiencing a nonsense error. This likely has little to do with codon usage, as their respective transcripts have > 3700 codons, such that there are many opportunities for any given ribosome to drop off during elongation.

A previous study estimated the probability a ribosome completes translation using a Totally Asymmetric Exclusion Process (TASEP) simulation parameterized by data taken from polysome profiling data (transla-

tion initiation rates), tRNA counts (codon-specific elongation waiting times), and a single NSE rate estimate taken from *E. coli* (Bonnin *et al.*, 2017). We compared our estimates of translation completion probability based on a model fit to empirical data $\sigma_g(n_g)$ to theoretical expectations based on this TASEP simulation (referred to as "drop-off resilience"), finding them to be highly correlated (Figure 4B, Spearman rank correlation $R = 0.95$, $p < 2.2E-16$). We still observed a high correlation between the empirical and theoretical estimates when conditioning on transcript length (partial Spearman rank correlation, $R_{partial} = 0.74$, $p < 2.2E-16$), indicating transcript length is not the only cause of the similarity between the two estimates of translation completion probabilities. Estimates of $\sigma_g(n_g)$ from PANSE are generally greater than expected based on these TASEP simulations, which assumed a uniform NSE rate for all codons. The NSE rate used for the TASEP simulations was taken from a study that implicitly assumed the drop in ribosome density along transcripts was solely due to nonsense errors (Sin *et al.*, 2016). As we and others have shown, the 5'-ramp region inflates estimates of NSE rates (Duc and Song, 2018). Regardless, the high correlation between $\sigma_g(n_g)$ and theoretical expectations suggests our model generally captures the across-transcript variability in the probability a ribosome completes translation based solely on empirical data.

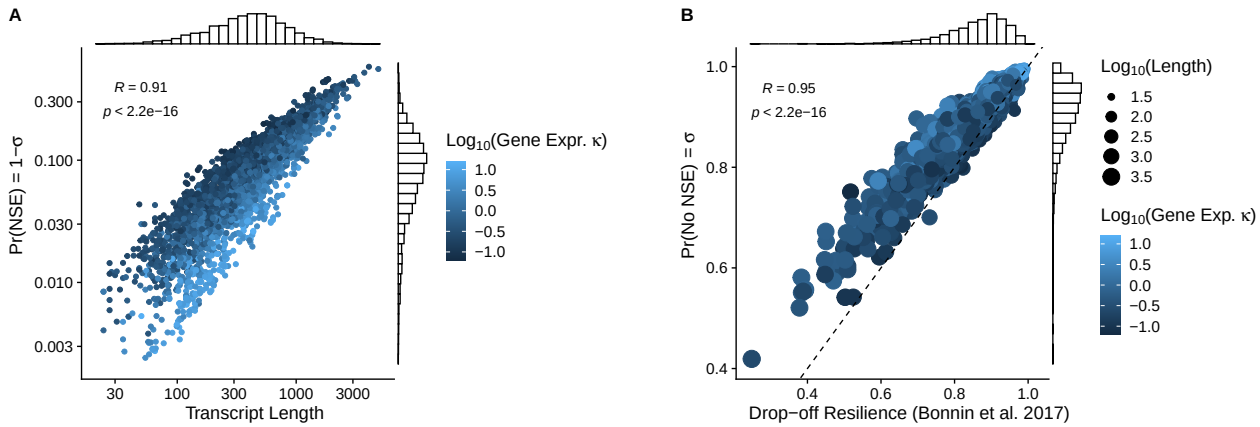

Figure 4: Variation in the probability of a single ribosome completing translation $\sigma_g(n_g)$. (A) Relationship between length and σ_g, with colors representing the ROC-SEMPPR protein production rate ϕ for each transcript. (B) Comparison of σ_g with simulation-based estimates of drop-off resilience (Bonnin *et al.*, 2017). Colors represent κ_g from ROC-SEMPPR (i.e. ϕ) for each transcript and the size of the point represents the transcript's length.

Evidence for adaptation to reduce NSEs

As protein synthesis is energetically costly, natural selection is expected to result in genome-level patterns consistent with adaptations to reduce the costs of NSEs. We specifically examine two key adaptations:

(1) reduced frequency of codons with higher NSE probabilities moving along the transcript in the 5' to 3' direction and (2) an anti-correlation between the frequencies of codons with higher NSE probabilities and gene expression.

Evidence that adaptation increases with position

As the energetic investment in producing a protein increases as amino acids are added to the peptide chain, natural selection against NSEs is expected to be weakest near the start of mRNA translation, resulting in position-specific patterns of codon usage in which nonsense error-prone codons are biased toward the 5'-end of the mRNA transcript (Eyre-Walker, 1996, Qin *et al.*, 2004, Gilchrist and Wagner, 2006, Gilchrist, 2007, Gilchrist *et al.*, 2009). After determining the average position for each codon across transcripts, normalized between 0 (the start codon) and 1 (the stop codon), we find a negative correlation between NSE probabilities and a codon's average relative position in the transcript (Spearman rank correlation $R = -0.34$, $p = 0.0075$, Figure 5A). This indicates codons with higher NSE probabilities tend to be located upstream in the transcript relative to codons with lower NSE probabilities. Codons with lower NSE probabilities tend to show little position-specific bias, as indicated by the clustering of these codons around the normalized position 0.5 i.e., they are equally likely to be found near the 5'-end as the 3'-end of the transcript. In contrast, codons with higher NSE probabilities generally have an average normalized position less than 0.5, indicating a bias towards the 5'-end of the transcript. Tryptophan codon TGG and both cysteine codons TGC/T are exceptions to this general trend, despite all three codons being a single nucleotide mutation away from stop codons. We emphasize that the first 200 codons of each transcript were not considered in the actual parameter estimation (i.e., were not considered in the likelihood calculation).

Natural selection to reduce NSEs is expected to strengthen along transcripts, thereby increasing the probability of successful elongation along transcripts. Except for the first 10 codons, the probability of elongation increases along the transcript before appearing to plateau (Figure 5B). This increase in the probability of elongation as a function of codon position is consistent with adaptations to reduce the energetic cost of NSEs. Regarding the apparent decrease in the probability of elongation for the first 10 codons, we note that these results are qualitatively consistent with previous findings that observed a decrease in codon bias immediately following the start codon, followed by a gradual increase (Qin *et al.*, 2004). The change in the probability of elongation observed in the real protein-coding sequences is much greater than observed for the null expectations generated by permuting codons (Figure 5B and Figure S4). This is consistent with natural selection against NSEs being generally weaker at the 5'-end.

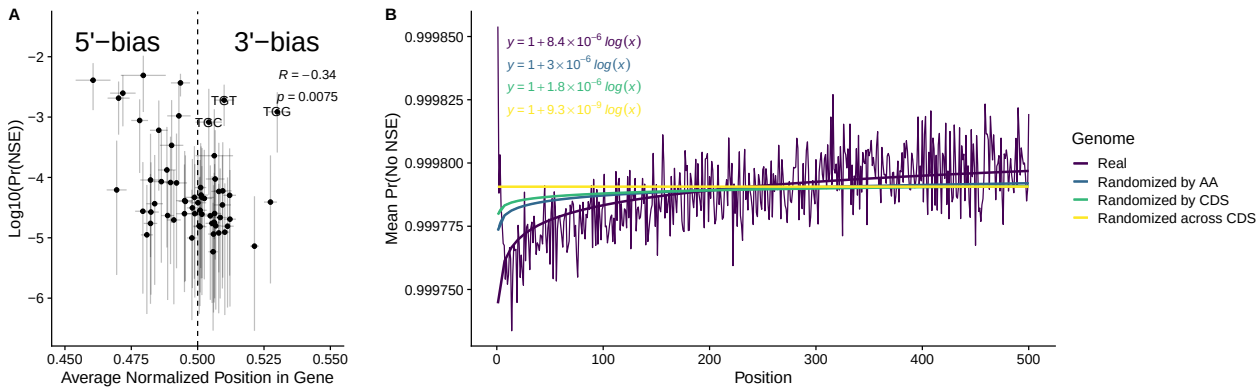

Figure 5: (A) Average position of codons in a gene as a function of NSE probabilities ($\log_{10}$ scale. An average position < 0.5 indicates a codon is more likely to be found at the 5'-end of the transcript. Similarly, an average position > 0.5 is more likely to be found at the 3'-end. The dashed line indicates the expected average position of a codon if has no bias towards the 5' or 3' end. (B) Geometric mean of the probability of successful translation by the ribosome up to the 500th codon across all transcripts for the real protein-coding sequences (purple). Solid lines and equations represent the linear regressions relating the log(Position) of a codon to the change in probability of elongation (i.e., not an NSE) for the real sequences and the various nulls. For the null regression lines, the mean slope and the mean intercept across the 1000 permuted sequences were used.

Evidence that adaptation against NSEs increases with gene expression

As highly expressed genes generally undergo more rounds of translation (i.e., take up a larger portion of the cell's energy budget), a lower probability of completing translation in a highly expressed gene will lead to more wasted NTP. Thus, selection against NSEs should increase with gene expression and with highly expressed genes having the greatest probability of completing translation $\sigma_g(n_g)$. In our comparisons of transcript length and the probability an NSE occurs, we observed a clear gradient in this relationship: highly-expressed genes tend to have lower probabilities of experiencing an NSE compared to transcripts of similar length but lower expression levels (Figure 4A). Using mRNA abundances (Weinberg *et al.*, 2016) as an independent measure of gene expression, we find a positive correlation (Spearman rank correlation $R = 0.31$, $p < 2.2E-16$) between the probability of completing translation $\sigma_g(n_g)$ and gene expression (Figure 6A). This analysis does not control for transcript length, indicating that regardless of length, highly expressed genes tend are more likely to be successfully translated. By controlling for transcript length using a partial correlation, we found that the correlation between gene expression and the probability of completing translation $\sigma_g(n_g)$ increased (partial Spearman rank correlation $R_partial = 0.49$, $p < 2.2E-16$). As expected when natural selection is acting against NSEs, selection against NSEs will be generally stronger in a high-expression gene compared to a moderate- or low-expression gene of similar length.

ROC-SEMPPR assumes selection (specifically, selection against translation inefficiency) on codon usage

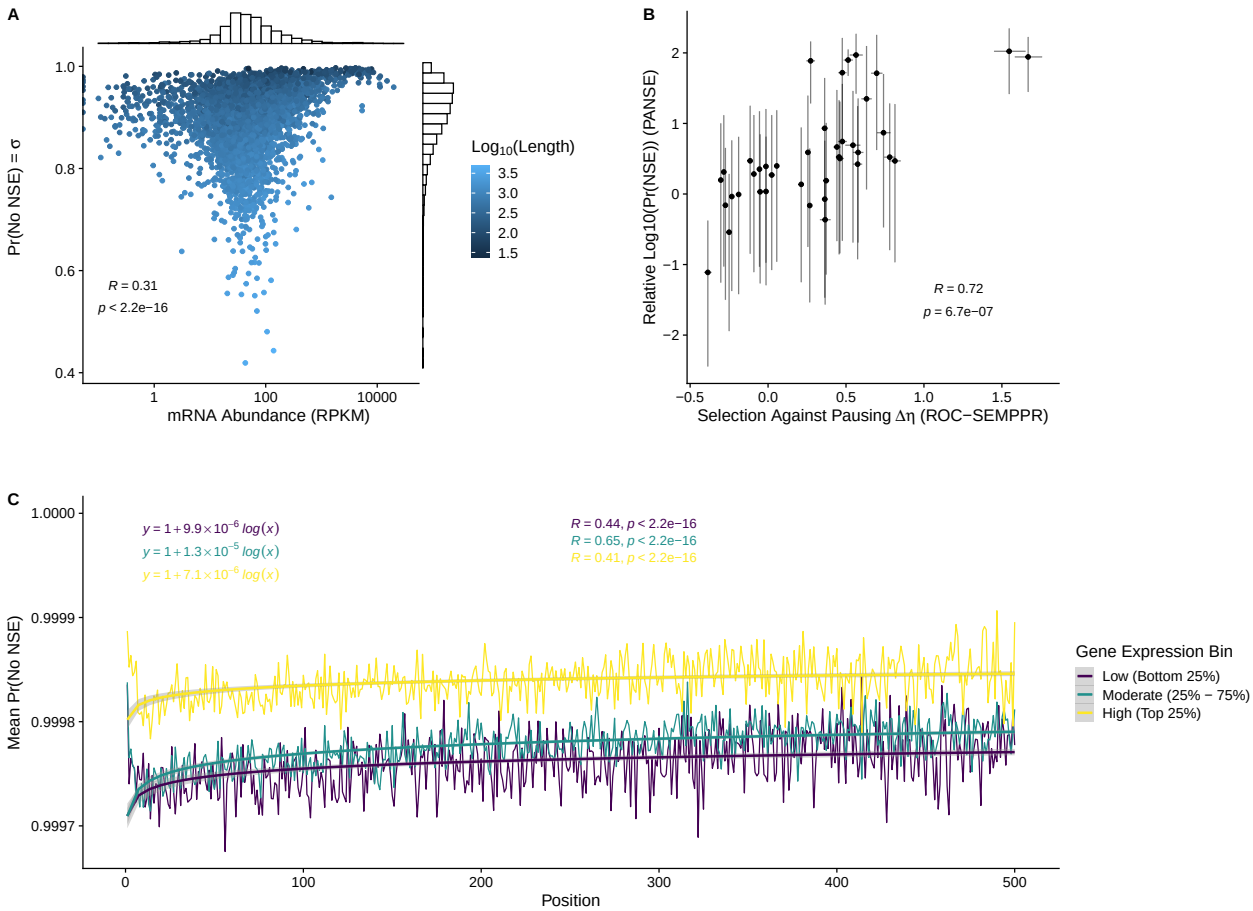

Figure 6: (A) Comparing gene expression (mRNA abundance RPKM) with the probability of a ribosome completing translation $\sigma_g(n_g)$. Histograms on the x and y axes represent the distributions of the corresponding variables. Colors indicate the length of the protein-coding sequence.(B) Comparing relative NSE rates among synonymous codons to selection coefficients $\Delta\eta$ from ROC-SEMPPR. A higher value of $\Delta\eta$ indicates a codon that is disfavored by selection relative to a reference codon (the alphabetically last codon for each amino acid). Error bars represent the 95% HDIs. Colors indicate the number of stop codons that can be reached via a single mutation to the codon. (C) Same as in Figure 5(B) but separating genes into bins based on mRNA abundances (RPKMs). Bins represent the quartiles. Solid lines represent the linear regression relating codon log(position) to the probability of elongation.

is invariant within a gene, but codons with higher NSE probabilities are expected to be avoided in high-expression genes due to the increased energetic costs of frequent translation errors. Under selection against NSEs, we expect ROC-SEMPPR's selection coefficients $\Delta\eta$ to be correlated with relative NSE probabilities (i.e., between synonymous codons). We find that $\Delta\eta$ and differences in NSE probabilities are well-correlated (Spearman rank correlation $R = 0.72, p = 6.7E-07$), indicating that synonymous codons with higher NSE probability are avoided in high-expression genes.

Consistent with selection against synonymous codon with higher NSE probabilities in high-expression

genes, we observe the probability of successful elongation by position is greater in high-expression genes (Figure 6C), indicating these genes are better adapted to reduce the cost of nonsense errors. Independent of gene expression, we observe the probability of successful elongations increases along the transcripts.

The energetic costs of NSEs are likely substantial

Although NSEs are rarer than missense errors (Jørgensen *et al.*, 1993), they may be more costly due to the high probability that the truncated protein is non-functional. We calculated the expected cost of mRNA translation (in terms of NTP) E[NTP/protein](c) (see Materials and Methods) that accounts for the direct ATP cost of translation initiation and peptide elongation, the indirect overhead cost due to ribosome pausing, and the direct and indirect costs of NSEs. The direct and indirect costs associated with translation initiation and elongation are the fixed costs as these will be incurred every time a functional protein is produced. In contrast, direct and indirect costs associated with NSEs as are variable costs as these will only be incurred <u>if</u> an NSE occurs (see Eqn. 3). For the cost of ribosome pausing during translation C, we focus on a rough estimate based on ribosome production costs and half-lives $C_{t_{1/2}} = 5.5$NTP/s as this estimate is more directly tied to ribosome assembly (see Materials and Methods). Unsurprisingly, the majority of NTP used during mRNA translation is associated with fixed direct costs of translation initiation and ribosome elongation (Figure 7A); however, this is not expected to impact the evolution of codon usage because fixed direct costs do not vary across codons. Studies on the evolution of codon usage have primarily focused on the cost of pausing, either explicitly or implicitly. Based on $C_{t_{1/2}} = 5.5$ NTP/s, the fixed indirect cost (i.e., the cost of ribosome pausing) is generally greater than variable costs (both variable direct and variable indirect costs) (Figure 7B). However, variable costs (i.e., the direct and indirect costs associated with an NSE) are usually within an order magnitude of the fixed indirect costs (i.e., the total cost of ribosome pausing for producing a single protein) for the majority of genes under consideration ($\approx 86\%$). We note this is likely a conservative estimate of variable costs. Using $C_{ROC} = 0.7$ NTP/s, which is based on comparing ROC-SEMPPR's selection coefficients $\Delta\eta$ to ribosome waiting times w, we found the variable costs to be generally greater than the fixed indirect costs (Figure S5). Consistent with adaptation to reduce the cost of translation, indirect costs (both fixed and variable) generally decrease as a function of total energetic flux (Figure S6). We suspect our estimates of C_{ROC} and $C_{t_{1/2}}$ reflect reasonable bounds on the true energetic cost of ribosome pausing. Combined, our results suggest the costs of NSEs are comparable to the cost of ribosomal pausing.

As natural selection to reduce energetic costs is expected to be stronger in high-expression genes, we

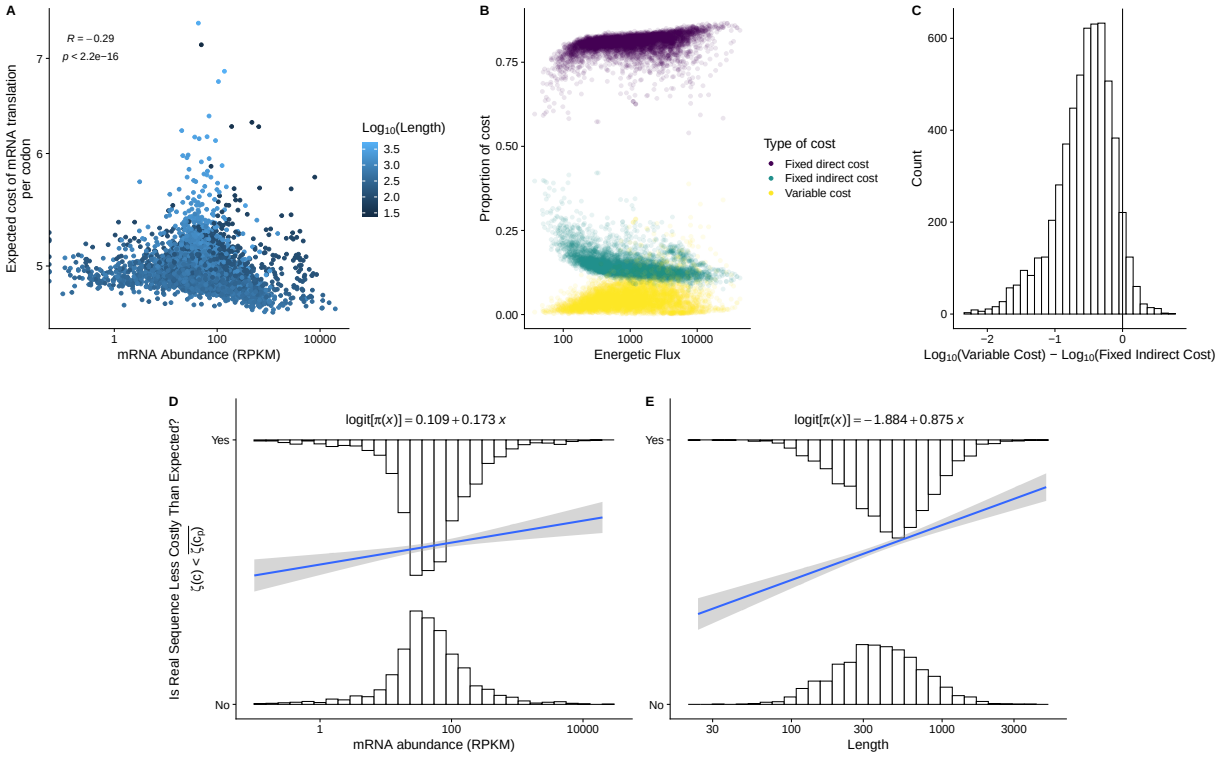

Figure 7: Expected cost of mRNA translation in terms of NTP across genes. (A) Relationship between gene expression and the expected cost of translation per codon. (B) Relative costs of translation in terms of direct costs (translation initiation and peptide elongation), indirect costs (overhead costs of ribosome pausing), and variable costs. Genes are ordered based on total energetic flux: the product of the total costs and protein production rates ϕ (as estimated by ROC-SEMPPR). (C) The log fold-difference between the variable costs (i.e. the direct and indirect cost of NSEs) and the fixed indirect cost of elongation (i.e., ribosome pausing). The mean log fold difference is -0.56 (One-sample t-test $p < 2.2E-16$). (D) Relationship between gene expression and the probability of whether or not a gene has a lower observed energetic cost than expected based on the permuted nulls. Histograms represent the distribution of gene expression for genes with and without evidence of adaptation against nonsense errors. The solid line and the equation (slope: represent the result of a logistic regression. The logistic regression slope was statistically significant ($p = 0.00024$), but the intercept was not ($p = 0.19$). (E) Same as in (D), but with length as the predictor variable. Both the logistic regression slope ($p < 2.2E-16$) and intercept ($p = 2.4E-12$) were statistically significant.

expect the cost of a gene to decrease as gene expression increases. Consistent with this expectation, we observe the expected cost per codon (i.e. E[NTP/protein]$(c)/n$, where n is the number of codons) is negatively correlated with gene expression (Figure 7C). To test for evidence of adaptation to reduce the cost of NSEs, we generated a null distribution for the expected translation cost for each protein-coding sequence based on 1000 permuted sequences of synonymous codons within the gene. These permutations could be viewed as nulls reflecting the absence of natural selection against NSEs, as permutations do not alter the codon-order

invariant fixed translation costs. As codon usage for *S. cerevisiae* is at selection-mutation-drift equilibrium (Shah and Gilchrist, 2011), we are effectively testing against a null that assumes adaptive codon usage via natural selection that varies with gene expression but not with position within a gene e.g., ribosome pausing. We found the true expected cost of a sequence $\mathrm{E[NTP/protein]}(c)$ is less than the mean expected cost of the permuted sequence $\overline{\mathrm{E[NTP/protein]}(c_p)}$ for 59% of protein-coding sequences, which is greater than the 50% expected by random chance (two-sided binomial test, $p < 2.2E-16$). The average difference between the true cost and the mean cost of the permuted sequences is approximately 4 NTPs, roughly the same cost as initiating another ribosome or the direct cost of translation elongation.

We have already seen evidence that high-expression genes are better adapted to reduce the cost of NSEs. Furthermore, as NSEs are more likely to occur in longer genes, gene length may be another predictor of adaptation to reduce the cost of NSEs. By classifying protein-coding sequences as either those that were better adapted relative to the permuted sequences (i.e., 1 if $\mathrm{E[NTP/protein]}(c) < \overline{\mathrm{E[NTP/protein]}(c_p)}$ and 0 otherwise), we find highly expressed and longer genes are more likely to be adapted to reduce the cost of NSEs (Figure 7D,E).

Discussion

The impact of translation errors on protein evolution has been a major focus for the last 3 decades, with most of this work focused on the impact of missense errors (Kurland, 1992, Akashi, 1994, Drummond *et al.*, 2006, Drummond and Wilke, 2008, Yang *et al.*, 2012). Recent advances in mass spectrometry technology and proteome bioinformatics make it possible to detect missense errors on a proteome-scale (Mordret *et al.*, 2019). Due to the robustness of the genetic code, missense errors may not necessarily lead to a non-functional protein. In contrast, nonsense errors are likely to almost always result in a non-functional protein. There is no current high-throughput technology to directly identify the location of a nonsense error at a particular codon in a transcript. However, high-throughput ribosome profiling allows for a steady-state measurement of the "translatome," including the ability to assign ribosomes to particular codons (Ingolia *et al.*, 2009). We developed a model of ribosome movement during mRNA translation to quantify codon-specific nonsense error (NSE) rates and probabilities from ribosome profiling data.

Applying our model to an exemplary measurement for *S. cerevisiae* (Weinberg *et al.*, 2016), we found evidence that NSE rates (sometimes referred to as the "background NSE rate") vary across codons. Importantly, prior work generally assumed the probability of an NSE occurring at any given codon was solely

due to variation in the elongation rates of the codons, with the NSE rate assumed uniform across codons (Gilchrist and Wagner, 2006, Gilchrist, 2007, Bonnin *et al.*, 2017). In contrast, we observed that NSE rates vary over multiple orders of magnitude (8.65×10^{-6} to 2.13×10^{-3}), suggesting intrinsic properties of codons that alter their propensity to trigger an NSE. Indeed, codons a single nucleotide away from a stop codon (i.e., has a stop codon neighbor) at the 3rd nucleotide position (cysteine codons TGC/T and tryptophan codons TGG) have a greater NSE rate, on average, compared to codons that do not have a 3rd position stop codon neighbor. The high NSE rate of the tryptophan codon TGG is particularly notable in the context of more direct studies which label this codon a cryptic stop codon due to its ability to bind with eRF1 (Chavatte *et al.*, 2003). Surprisingly, this was not the case for the amino acid tyrosine, for which both of its codons (TAC/T) each have 2 stop codon neighbors (TAA/TAG) – although the effect of having 2-stop codon neighbors at the 3rd nucleotide is positive compared to a codon with no stop neighbors, it was not statistically significant ($p = 0.52$). Consistent with our results, previous work found eRF1 binds TAC only weakly (Chavatte *et al.*, 2003). Furthermore, codons with a stop codon neighbor at the second nucleotide position have, on average, lower NSE rates. Taken together, our results indicate being a single nucleotide away from a stop codon is not in and of itself sufficient to increase the NSE rates of codons. Indeed, other mechanisms have been found that trigger premature translation termination (Zaher and Green, 2011, Chiabudini *et al.*, 2014, Presnyak *et al.*, 2015, Chadani *et al.*, 2017). For example, previous work found that a codon-anticodon mismatch (i.e., a missense error) for the tRNA in ribosome P-site was detrimental to the accuracy decoding of the codon in the A-site and increased the probability of a release factor binding to a sense codon (Zaher and Green, 2011, Nguyen *et al.*, 2023). In this case, the probability of an NSE at any given codon is not independent, but a function of the missense error rate of its immediate upstream neighbor, among other things.

Overall, we observed that simulated ribosome counts across codons and genes generated from PANSE were generally in good agreement with the real data beyond the 5'-ramp region (first 200 codons), indicating PANSE adequately models the underlying processes shaping ribosome profiling data. We emphasize our model does not explicitly account for various potential factors hypothesized to impact local elongation rates, from upstream basic amino acids in the ribosome tunnel to downstream mRNA secondary structure (Duc and Song, 2018, Tunney *et al.*, 2018, Erdmann-Pham *et al.*, 2020). We note the correlation between the real and simulated data may be inflated due to the latter being generated from parameters estimated from the former, but also attenuated (i.e., biased toward 0) due to the inherent noise in any sequencing data (Sokal and Rohlf, 1995). Regarding the impact of noise, previous work found generally poor agreement between the ribosome footprint counts at individual codons across independent ribosome profiling experiments (Wright

et al., 2020). Future extensions of the model will benefit from explicitly accounting for noise in ribosome profiling data by combining information across experiments.

Significant efforts have been made to understand the increased ribosome density at the 5'-end of transcript observed universally in ribosome profiling data. In combination with the increased frequency towards slow codons at the 5'-end, previous work hypothesized slow translation was favored at the 5'-end of transcripts to prevent downstream ribosome queuing ("the 5'-ramp hypothesis") (Tuller *et al.*, 2010). This hypothesis has been challenged as either a technical artifact of ribosome profiling (Weinberg *et al.*, 2016) or a statistical artifact (Sejour *et al.*, 2024). Although our work here is insufficient to directly address the 5'-ramp hypothesis, selection against nonsense errors could lead to similar patterns of intragenic codon usage as expected under the 5'-ramp hypothesis. However, we can say that nonsense errors likely do not entirely explain the dramatic drop observed in ribosome density at the 5'-end. Given that the expected nonsense error rates are unrealistically high, we suspect (but cannot confirm) that the increased ribosome density at the 5'-ends results from an experimental artifact or a decrese in codon decoding at some codons (Weinberg *et al.*, 2016).

We found numerous lines of evidence consistent with adaptation to reduce the cost of nonsense errors in intragenic codon usage patterns. Generally, selection against nonsense errors is expected to be weakest at the 5'-end of a gene, as a nonsense error occurring early in translation will waste fewer NTPs. As a result, the codon usage of 5'-ends is expected to be less adapted than regions further along in the gene. Indeed, codons with lower nonsense error probabilities tend to be found closer to the 5'-ends across all genes. Consistent with this avoidance of higher NSE probability codons at the 5'-end, we find that the probability of successful elongation generally increases along a transcript. We note that some codons, such as both cysteine codons (TGC/TGT) and the tryptophan codon TGG have high NSE probabilities, but show no bias towards the 5'-end. This likely reflects other factors shaping amino acid usage as a function of position, such as protein structure.

In general, high-expression genes tend to avoid codons more prone to NSEs. As a result, high-expression genes generally have a higher probability of completing mRNA translation and a lower expected total energetic cost of a protein. Furthermore, selection coefficients $\Delta\eta$ estimated via the population genetics model ROC-SEMPPR (generally assumed to reflects differences in ribosome pausing times) (Gilchrist *et al.*, 2015) are well-correlated with relative differences in NSE probabilities, consistent with high-expression genes avoiding codons more prone to the NSE rates. Importantly, selection coefficients estimated via ROC-SEMPPR will average over any form of selection pressure that correlates with gene expression (Cope and Gilchrist, 2022) – we cannot say to what extent these selection coefficients reflect selection for translation efficiency as

opposed to selection against nonsense errors. Interestingly, moderately expressed genes showed the greatest change along transcripts in the probability of a ribosome successfully elongating compared to high and low-expression genes. Assuming selection on codon usage is at least partially due to selection against NSEs, this may seem counter-intuitive at first glance. However, if selection against nonsense errors is strongest in high-expression genes, then selection may be shaping codon usage at the 5'-end. As a result, high-expression genes would be well-adapted against NSEs even at the 5'-end. In contrast, low-expression genes are expected to be under the weakest selection against NSEs, resulting in a lower probability of completion overall that changes very little as a function of position. Moderate-expression genes have may high enough protein production levels such that selection against NSEs is expected to be more effective than in low-expression genes, but not so much that the 5'-end is well-adapted. We speculate this explains the greater increase in the probability of elongation along transcripts in moderate-expression genes compared to high and low-expression genes.

Although studies on codon usage bias have predominantly focused on selection against ribosome pausing, recent theoretical work concluded the indirect cost of elongation due to ribosome pausing may be less of a factor shaping codon usage bias than other factors, including NSEs (Erdmann-Pham *et al.*, 2020). Based on a simple model for approximating the energetic costs (in NTPs) for each gene, we found the costs of NSEs comparable (i.e., within an order of magnitude) to the fixed indirect costs of ribosome pausing across the majority of the transcriptome. The majority (59%) of genes in *S. cerevisiae* exhibit signals consistent with adaptations to reduce the cost of NSEs. However, this is likely a conservative estimate of the proportion of genes adapted to reduce the cost of NSEs because our permutation test does not represent the expected codon frequencies sans natural selection. More accurately, our test reflects adaptation to reduce the cost of NSEs, in which the position of a codon is relevant, to adaptation to reduce a cost invariant with position e.g., ribosome pausing. Importantly, natural selection against NSEs and pausing are correlated, as both are impacted by codon-specific ribosome waiting times. Although a logistic regression revealed the probability of a gene being adapted to reduce the cost of NSEs increased with gene expression, this was not the case for many high-expression genes (Figure 7D). As discussed previously, high-expression genes have a relatively low probability of NSEs even at the 5'-end (Figure 6C), such that permuting synonymous codons within a gene likely has a small impact on the energetic cost for a significant portion of high-expression genes. Even so, the expected energetic cost savings of adaptation against NSEs is approximately 4 NTPs – roughly the same as the cost of initiating a new ribosome for translation or a single round of elongation. Noting that mRNA translation is generally initiation-limited, adaptation against nonsense errors is expected to result in 1 fewer initiation events to produce a functional protein. We note that these estimates were based on many

assumptions and parameter estimates from previous model fits. As a result, we may be also underestimating the cost of NSEs.

Our model of ribosome movement PANSE is applicable to any ribosome profiling measurement. This allows for investigations into the impact of different growth conditions on nonsense error probabilities (Sin *et al.*, 2016, Awad *et al.*, 2024). With the number of species with ribosome profiling measurements growing, it is possible to perform comparative analyses of nonsense error rates and probabilities. Furthermore, using the approaches outlined here, it is possible to quantify the impact of natural selection against nonsense errors on protein-coding sequence evolution across species and how this may change with e.g., effective population size N_e.

Conclusion

By applying a model of mRNA translation to an exemplary ribosome profiling dataset in *S. cerevisiae*, we found multiple lines of evidence that nonsense errors play a significant role in protein-coding evolution that has largely been underappreciated. Overall, 59% of the protein-coding sequences in *S. cerevisiae* exhibit signals of adaptation to reduce the cost of nonsense errors. Natural selection to reduce the cost of ribosome pausing has been the predominant hypothesis to explain codon usage bias, but if the cost of nonsense errors is comparable, if not greater than the cost of pausing, then this hypothesis must be updated or revised. Further consideration of the impact and consequences of nonsense errors is critical for understanding the evolution of codon usage bias, which has been observed to varying degrees across all taxa.

Data Availability

Raw sequencing reads were obtained from the SRA (SRR1049521) and processed using riboviz 2 `https://github.com/riboviz/riboviz/`. The version of AnaCoDa used for model fits can be found at `https://github.com/acope3/RibModelFramework/tree/develop`. Processed data and scripts for performing model fits and subsequent analyses are available at `https://github.com/acope3/Yeast_Nonsense_Error_Analysis`.

Acknowledgements

We thank Premal Shah, Matt Pennell, Edward Wallace, and Josh Schraiber for helpful discussions throughout the course of this project and the writing of this manuscript.

Funding Statement

This work was supported by the NIH-funded Rutgers INSPIRE IRACDA Postdoctoral Program (grant #GM093854 to ALC).

Conflict of interest statement.

None declared.

References

Akashi, H. 1994. Synonymous codon usage in drosophila melanogaster: natural selection and translational accuracy. *Genetics*, 136.

Awad, S., Valleriani, A., and Chiarugi, D. 2024. A data-driven estimation of the ribosome drop-off rate in s. cerevisiae reveals a correlation with the genes length. *NAR Genomics and Bioinformatics*, 6: 36.

Bonnin, P., Kern, N., Young, N. T., Stansfield, I., and Romano, M. C. 2017. Novel mrna-specific effects of ribosome drop-off on translation rate and polysome profile. *PLOS Computational Biology*, 13: e1005555.

Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129: 897–907.

Chadani, Y., Niwa, T., Izumi, T., Sugata, N., Nagao, A., Suzuki, T., Chiba, S., Ito, K., and Taguchi, H. 2017. Intrinsic ribosome destabilization underlies translation and provides an organism with a strategy of environmental sensing. *Molecular Cell*, 68: 528–539.e5.

Charlesworth, B. 2009. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10: 195–205.

Chavatte, L., Frolova, L., Laugâa, P., Kisselev, L., and Favre, A. 2003. Stop codons and ugg promote efficient binding of the polypeptide release factor erf1 to the ribosomal a site. *Journal of Molecular Biology*, 331: 745–758.

Chiabudini, M., Tais, A., Zhang, Y., Hayashi, S., Wölfle, T., Fitzke, E., and Rospert, S. 2014. Release factor erf3 mediates premature translation termination on polylysine-stalled ribosomes in saccharomyces cerevisiae. *Molecular and Cellular Biology*, 34.

Chou, H. J., Donnard, E., Gustafsson, H. T., Garber, M., and Rando, O. J. 2017. Transcriptome-wide analysis of roles for trna modifications in translational regulation. *Molecular Cell*, 68: 978–992.e4.

Cope, A. L. and Gilchrist, M. A. 2022. Quantifying shifts in natural selection on codon usage between protein regions: a population genetics approach. *BMC Genomics*, 23: 408.

Cope, A. L. and Shah, P. 2022. Intragenomic variation in non-adaptive nucleotide biases causes underestimation of selection on synonymous codon usage. *PLOS Genetics*, 18: e1010256.

Cope, A. L., Anderson, F., Favate, J., Jackson, M., Mok, A., Kurowska, A., Liu, J., Mackenzie, E., Shivakumar, V., Tilton, P., Winterbourne, S. M., Xue, S., Kavoussanakis, K., Lareau, L. F., Shah, P., and

Wallace, E. W. 2022. riboviz 2: a flexible and robust ribosome profiling data analysis and visualization workflow. *Bioinformatics*, 38: 2358–2360.

dos Reis, M., Savva, R., and Wernisch, L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Research*, 32: 5036–5044.

Drummond, D. A. and Wilke, C. O. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134: 341–352.

Drummond, D. A. and Wilke, C. O. 2009. The evolutionary consequences of erroneous protein synthesis. *Nature Reviews Genetics*, 10: 715–724.

Drummond, D. A., Raval, A., and Wilke, C. O. 2006. A single determinant dominates the rate of yeast protein evolution. *Molecular Biology and Evolution*, 23: 327–337.

Duc, K. D. and Song, Y. S. 2018. The impact of ribosomal interference, codon usage, and exit tunnel interactions on translation elongation rate variation. *PLoS Genetics*, 14: e1007166.

Erdmann-Pham, D. D., Duc, K. D., and Song, Y. S. 2020. The key parameters that govern translation efficiency. *Cell Systems*, 10: 183–192.

Eyre-Walker, A. 1996. Synonymous codon bias is related to gene length in escherichia coli: Selection for translational accuracy? *Mol. Biol. Evol.*, 13: 864–872.

Freistroffer, D. V., Kwiatkowski, M., Buckingham, R. H., and Ehrenberg, M. 2000. The accuracy of codon recognition by polypeptide release factors. *Proceedings of the National Academy of Sciences of the United States of America*, 97.

Frumkin, I., Lajoie, M. J., Gregg, C. J., Hornung, G., Church, G. M., and Pilpel, Y. 2018. Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proceedings of the National Academy of Sciences of the United States of America*, 115: E4940–E4949.

Gilchrist, M. A. 2007. Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Molecular Biology and Evolution*, 24: 2362–2372.

Gilchrist, M. A. and Wagner, A. 2006. A model of protein translation including codon bias, nonsense errors, and ribosome recylcing. *Journal of Theoretical Biology*, 239: 417–434.

Gilchrist, M. A., Shah, P., and Zaretzki, R. 2009. Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. *Genetics*, 183: 1493–1505.

Gilchrist, M. A., Chen, W. C., Shah, P., Landerer, C. L., and Zaretzki, R. 2015. Estimating gene expression and codon-specific translational efficiencies, mutation biases, and selection coefficients from genomic data alone. *Genome Biology and Evolution*, 7: 1559–1579.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., and Weissman, J. S. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324: 218–223.

Joshi, K., Bhatt, M. J., and Farabaugh, P. J. 2018. Codon-specific effects of trna anticodon loop modifications on translational misreading errors in the yeast saccharomyces cerevisiae. *Nucleic Acids Research*, 46: 10331–10339.

Jørgensen, F., Adamski, F. M., Tate, W. P., and Kurland, C. G. 1993. Release factor-dependent false stops are infrequent in escherichia coli. *Journal of Molecular Biology*, 230.

Kramer, E. B., Vallabhaneni, H., Mayer, L. M., and Farabaugh, P. J. 2010. A comprehensive analysis of translational missense errors in the yeast saccharomyces cerevisiae. *RNA*, 16: 1797–1808.

Kurland, C. G. 1992. Translational accuracy and the fitness of bacteria. *Annual Review of Genetics*, 26: 29–50.

Lareau, L. F., Hite, D. H., Hogan, G. J., and Brown, P. O. 2014. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mrna fragments. *eLife*, 2014.

Li, W. H. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *Journal of Molecular Evolution*, 24: 337–345.

Lynch, M. and Marinov, G. K. 2015. The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences of the United States of America*, 112: 15690–15695.

Mordret, E., Dahan, O., Asraf, O., Rak, R., Yehonadav, A., Barnabas, G. D., Cox, J., Geiger, T., Lindner, A. B., and Pilpel, Y. 2019. Systematic detection of amino acid substitutions in proteomes reveals mechanistic basis of ribosome errors and selection for translation fidelity. *Molecular Cell*, 75: 427–441.

Nguyen, H. A., Hoffer, E. D., Fagan, C. E., Maehigashi, T., and Dunham, C. M. 2023. Structural basis for reduced ribosomal a-site fidelity in response to p-site codon–anticodon mismatches. *Journal of Biological Chemistry*, 299.

Presnyak, V., Alhusaini, N., Chen, Y. H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K. E., Graveley, B. R., and Coller, J. 2015. Codon optimality is a major determinant of mrna stability. *Cell*, 160: 1111–1124.

Qin, H., Wu, W. B., Kreitman, J. M. C. M., and Li, W. 2004. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics*, 168: 2245–2260.

Russell, J. B. and Cook, G. M. 1995. Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiological Reviews*, 59: 48–62.

Sejour, R., Leatherwood, J., Yurovsky, A., and Futcher, B. 2024. Enrichment of rare codons at 5' ends of genes is a spandrel caused by evolutionary sequence turnover and does not improve translation. *eLife*, 12.

Shah, P. and Gilchrist, M. 2011. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proceedings of the National Academy of Sciences of the United States of America*, 108: 10231–10236.

Shah, P. and Gilchrist, M. A. 2010. Effect of correlated trna abundances on translation errors and evolution of codon usage bias. *PLoS Genetics*, 6: 1–9.

Sin, C., Chiarugi, D., and Valleriani, A. 2016. Quantitative assessment of ribosome drop-off in e. coli. *Nucleic Acids Research*, 44: 2528–2537.

Sokal, R. R. and Rohlf, F. J. 1995. *Biometry - The Principles and Practices of Statistics in Biological Research*. W.H. Freeman, 3rd edition.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, A. V. D. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64: 583–616.

Svidritskiy, E., Demo, G., and Korostelev, A. A. 2018. Mechanism of premature translation termination on a sense codon. *Journal of Biological Chemistry*, 293: 12472–12479.

Tsung, K., Inouye, S., and Inouye, M. 1989. Factors affecting the efficiency of protein synthesis in escherichia coli: Production of a polypeptide of more than 6000 amino acid residues. *Journal of Biological Chemistry*, 264: 4428–4433.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141: 344–354.

Tunney, R., McGlincy, N. J., Graham, M. E., Naddaf, N., Pachter, L., and Lareau, L. F. 2018. Accurate design of translational output by a neural network model of ribosome distribution. *Nature Structural and Molecular Biology*, 25: 577–582.

Wagner, A. 2005. Energy constraints on the evolution of gene expression. *Molecular Biology and Evolution*, 22: 1365–1374.

Wallace, E. W. J., Airoldi, E. M., and Drummond, D. A. 2013. Estimating selection on synonymus codon usage from noisy experimental data. *Molecular Biology and Evolution*, 30: 1438–1453.

Warnecke, T. and Hurst, L. D. 2010. Groel dependency affects codon usage - support for a critical role fo misfolding in gene evolution. *Molecular Systems Biology*, 6: 1–11.

Warnecke, T. and Hurst, L. D. 2011. Error prevention and mitigation as forces in the evolution of genes and genomes. *Nature Reviews Genetics*, 12: 875–881.

Weinberg, D. E., Shah, P., Eichhorn, S. W., Hussmann, J. A., Plotkin, J. B., and Bartel, D. P. 2016. Improved ribosome-footprint and mrna measurements provide insights into dynamics and regulation of yeast translation. *Cell Reports*, 14: 1787–1799.

Wright, G., Rodriguez, A., Li, J., Clark, P. L., Milenković, T., and Emrich, S. J. 2020. Analysis of computational codon usage models and their association with translationally slow codons. *PLOS ONE*, 15: e0232003.

Wu, C. C. C., Zinshteyn, B., Wehner, K. A., and Green, R. 2019. High-resolution ribosome profiling defines discrete ribosome elongation states and translational regulation during cellular stress. *Molecular Cell*, 73: 959–970.e5.

Yang, J.-R., Liao, B.-Y., Zhuang, S.-M., and Zhang, J. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America*, 109: E831–40.

Yang, Y., Gu, Q., Zhang, Y., Sasaki, T., Crivello, J., O'Neill, R. J., Gilbert, D. M., and Ma, J. 2018. Continuous-trait probabilistic model for comparing multi-species functional genomic data. *Cell Systems*, 7: 208–218.e11.

Zaher, H. S. and Green, R. 2011. A primary role for release factor 3 in quality control during translation elongation in escherichia coli. *Cell*, 147: 396–408.