Slow Translation of the 5' ends of genes may be a non-adaptive consequence of their rapid evolutionary turn-over and slow loss of rare codons.

Richard Sejour, Janet Leatherwood, and Bruce Futcher
Dept. of Microbiology and Immunology, Stony Brook University, Stony Brook, NY 11794-5222
B. Futcher, corresponding author, bfutcher@gmail.com

Abstract

Previously, Tuller et al. found that the first 30 to 100 codons of the genes of yeast and other eukaryotes are slightly enriched for rare codons, and are presumably translated somewhat slowly. They argued, based on informatics, that this initial slow translation "ramp" was adaptive; and that it increased the efficiency of translation by provided a queuing mechanism to prevent ribosome collisions and traffic jams. Today, the translational speeds of the various codons are known, and there is a strong correlation between rare codon usage and slow translation speed. We have re-examined the slow translation ramp. We confirm the finding of Tuller et al. that a region of slow translation exists. However, we also find that the 5' ends of yeast genes are poorly conserved, suggesting that they are unstable in evolution and turn over relatively rapidly. When a new 5' end forms *de novo*, it is likely to include codons that would otherwise be rare. Because evolution has had a relatively short time to select against these codons, 5' ends are typically slightly enriched for rare, slow codons. Opposite to the expectation of Tuller et al., we show by direct experiment that genes with slowly translated 5' ends are translated poorly compared to genes with rapidly translated 5' ends. Thus we conclude that slow 5' translation may be a "spandrel"; it is a non-adaptive consequence of something else, in this case the turnover of 5' ends in evolution.

Introduction

Tuller et al. (Tuller et al., 2010) were interested in the idea a slow translational ramp at the beginning of a gene might queue ribosomes in an orderly way, thereby preventing ribosome traffic jams and collisions. However, at that time, translation speeds for the 61 sense codons were not known from direct measurement. Therefore, as a proxy for codon translation speed, Tuller et al. devised a proxy speed measurement based on the tRNA-adaptation index (tAI). The assumption of this is that codons recognized by more common tRNAs will be translated faster. Using this proxy, Tuller et al. indeed found that in yeast and many other eukaryotes, the first 30 to 100 codons, depending on the species, are enriched for rare codons (which are the codons for which tRNAs are rare), and are presumably translated slowly. The size of the effect is small (the initial rate of translation is about 3% less than the average rate, Tuller et al. Fig. 2C), but it is very highly statistically significant.

Since then, several workers have directly measured the speed of translation of individual codons using ribosome profiling data (cite). With such data in hand, we re-visited the issues addressed by Tuller et al. (Tuller et al., 2010). On the one hand, we confirm the finding of Tuller et al. that there is an initial 5' region of slow translation. On the other hand, various aspects of the data led us to an alternative hypothesis, namely that the 5' ends were turning over relatively rapidly in evolution; that these 5' ends were therefore relatively young; and that selection had not yet fully succeeded in removing all of the rare, slow codons initially present in the de novo 5' ends. We did a direct experimental test of the effects of slow or fast initial translation. We found that slow initial translation resulted in lower protein production than fast initial translation. This continued to be true even when we placed ribosome collision sites inside the reporter gene. Thus slow initial translation is likely a spandrel (Gould & Lewontin, 1979), a non-adaptive consequence of something else, in this case the instability of 5' ends in evolution.

Results

Measurements of codon translation speed reveal slow initial translation

Tuller et al. (Tuller et al., 2010) used codon-specific tRNA abundance as a proxy to estimate speed of translation of codons. Since then, analysis of ribosome profiling data has yielded direct measurements of the translation speed of individual codons (Dao Duc & Song, 2018; Gardin et al., 2014; Gritsenko, Hulsman, Reinders, & de Ridder, 2015; Lareau, Hite, Hogan, & Brown, 2014; Sharma et al., 2019; Tunney et al., 2018; Wang, McManus, & Kingsford, 2017). Accordingly, we have repeated some of the work of Tuller et al., but using the Ribosome Residence Time (RRT, Methods and Materials, Table S1) (Gardin et al., 2014) of each of the 61 sense codons as a measure of translation speed.

Using the RRT, translation rates were measured in sliding windows across all coding ORFs from *S. cerevisiae*. In our analysis below, the start codon was omitted because it is constant across genes and is an unusually "fast" codon. Confirming Tuller et al., we find that the first 30-70 codons had significantly lower translation speeds than the rest of the gene (Fig. 1). We found that translation is slowest among the first 40 codons compared to windows of 30, 50, 60, or 70 codons (data not shown—we did not test longer distances). Therefore, our analyses focus on the first 40 codons; we call this the "Slow Initial Translation" region, or SIT. Although the tendency towards slow translation near the beginning of the gene is very highly statistically significant, the size of the effect is small. When we compare the first 40 codons of a gene to the rest of the same gene, we find that the average difference in translation speed is about 1.73% (similar to (Tuller et al., 2010)). For comparison, different codons can vary in translation speed by somewhat more than 200% (Table S1).

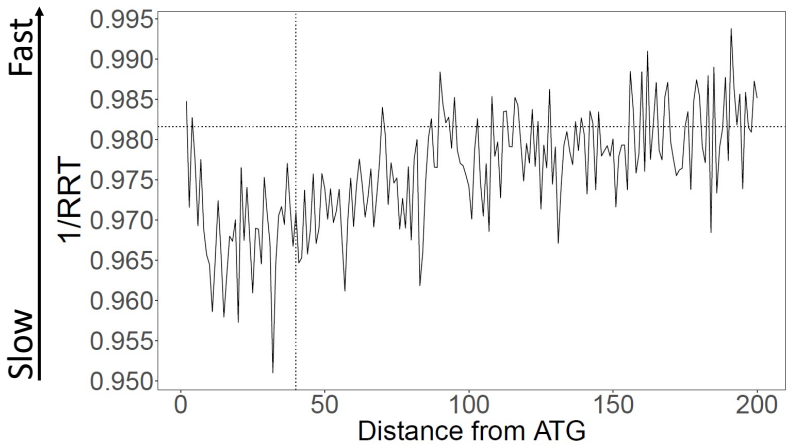

Figure 1: Measurement of translation speed across 5737 *S. cerevisiae* ORFs confirms slow initial translation (SIT). Ribosome Residence Time (RRT) was used as a measure of codon-specific translation speed. The horizontal line indicates the average RRT across all ORFs. The difference in speed between the SIT and body is about 1.7%.

Although on average genes are translated slowly near their 5' ends, there is considerable gene-to-gene variability. Fig. S1 shows the distribution of initial translation speed for all yeast ORFs. About 60% of genes have a Slow Initial Translation (SIT) region, while the remainder have Fast Initial Translation (FIT).

Rare (slow) codons are enriched within the first 40 codons.

Rare codons tend to be slow codons, and vice versa (Gardin et al. 2014) (Fig. 2). But the correlation is not perfect—other things being equal, A/T-rich codons tend to be faster than G/C-rich codons, and codons with a wobble base at the third position tend to be faster than codons with the cognate base (Gardin et

al., 2014). To better understand why gene beginnings are more slowly translated, we examined the relative usage of each of the 61 sense codons in the first 40 codons (Fig. 3).

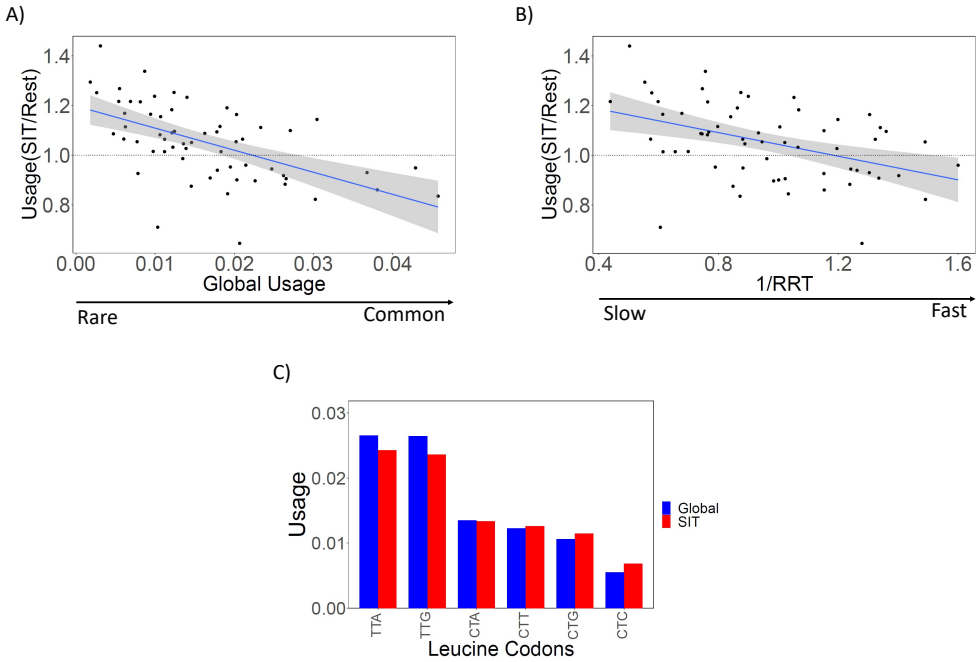

Figure 2: Slower initial translation (SIT) is caused by an enrichment of rarer and slower codons. A) Each dot represents one codon. Usage(SIT/Rest) is the ratio of the codon usage in the SIT (first 40 codons) relative to the rest of the gene. Relative to the body of genes, the first 40 codons are enriched for codons that are less common ($r= -0.561$; $p<0.001$). B) Similarly, there is a tendency for slower codons to be overrepresented in the first 40 codons ($r= -4.35$; $p<0.001$). C) Rare codons are still rare in the slow initial translation region (SIT), just not as rare. The usages of CTT, CTG, and CTC are higher in the SIT relative to the body of genes. However, even in the SIT, their usages are lower than for the common codons. This is true for rare and common codons for all amino acids.

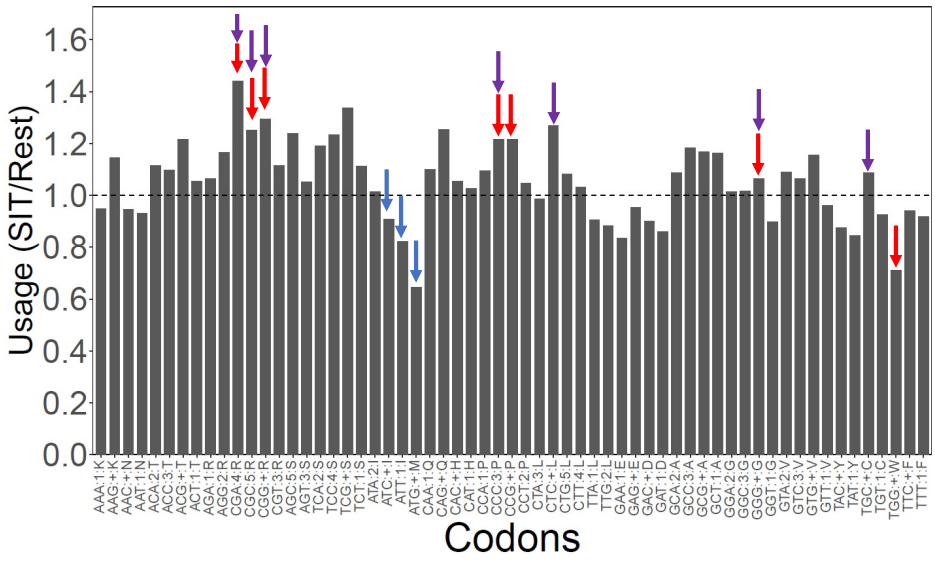

Figure 3: Relative codon usage in slow initial translation (SIT) region. Numbers on the x-axis (1,2,3,4,5, or +) indicate the codon usage, with 1 the most frequent, and "+" the rarest. Red arrows show the seven slowest codons (CCG, CGA, CGG, GGG, CGC, CCC, TGG), purple arrows show the seven rarest codons (CGG, CGC, CGA, TGC, CCG, CTC, GGG), and blue shows Start and possible Start codons (ATG, ATA, ATC).

In the first 40 codons, we saw notable depletion of AUG (by nearly 50%), ATT, and ATC (Fig. 3). These are all fast codons, and their depletion would lead to slower translation. But AUG is the usual Start codon, and ATT and ATC can also be used as Start codons. Possibly these codons are depleted to reduce the possibility of translation initiating at the wrong place. We considered whether the SIT might be due to depletion of AUG. We recalculated translation speeds after assigning AUG a neutral speed. On the one hand, this did reduce the difference in translation speed between the SIT and the rest of the gene by about 13%, but even so, the slow initial translation was highly significant. Thus, bias against Start codons contributes to, but does not fully explain, the SIT.

We also saw enrichments of all six arginine codons (likely because of their use in N-terminal signal sequences, data not shown), especially the slow, rare codons CGA, CGC, and CGG. Slow, rare codons for proline (CCC, CCG), leucine (CTC), glycine (GGG) and cysteine (TGC) were also enriched (Fig. 3).

Rare codons are rare in the first 40 codons, just not as rare.

Although the proportion of rare, slow codons in the SITs is relatively higher than the proportion of rare codons in the body of genes, rare codons are still rare compared to more common synonymous codons (e.g., Fig. 2C). That is, rare, slow codons are still strongly disfavored in the SIT, though they are less strongly disfavored than elsewhere. This was true for all rare codons. One can look at this as weak selection in favor of rare, slow codons in the SIT; alternatively, one can look at this as reduced selection against rare slow codons.

Why are there more rare, slow codons in N-terminal regions?

Our analysis agrees with that of Tuller that there is a slight relative enrichment of rare, slow codons near the 5' ends of coding regions. Tuller interpret this as an adaptation—they believe there is selection for slow codons near the 5' end to enhance efficiency of translation. But there are other possibilities.

We have noticed (see below) that the N-termini of yeast genes are often poorly conserved, and homologous genes often vary at the N-termini between different closely-related species. This suggests a different idea: N-termini are unstable and variable in evolution. They form *de novo* from new DNA sequence, and so all codons may initially occur at similar frequencies. Since these N-termini are, on average, younger than the remainder of the gene, selection has worked on them for a shorter time. Therefore selection against rare, slow codons is less complete, and N-termini, due to their relative youth, still retain some extra rare codons.

In this idea, the slight excess of rare codons near N-termini is not adaptive; it is not at all explained by selection. Instead, in the words of Gould and Lewontin (Gould & Lewontin, 1979), it is a spandrel. It is a non-adaptive by-product of something else, in this case the evolutionary instability of N-termini. We call this the "N-terminal instability" hypothesis. This idea is explored below.

Poor 5' conservation is a feature of many yeast genes.

To illustrate the idea, we picked several example genes where the N-termini are poorly conserved amongst yeast. We used protein-protein BLAST at NCBI to blast several query yeast genes against species of the subphylum *Saccharomycotina* (but having subtracted out all sequences from *Saccharomyces cerevisiae*). Fig. 4 shows that for these chosen example genes, the middle portions of the proteins are highly conserved, but the N-termini are not.

In the next section, we ask whether the findings of Fig. 4 can be generalized to most or all proteins of *S. cerevisiae*.

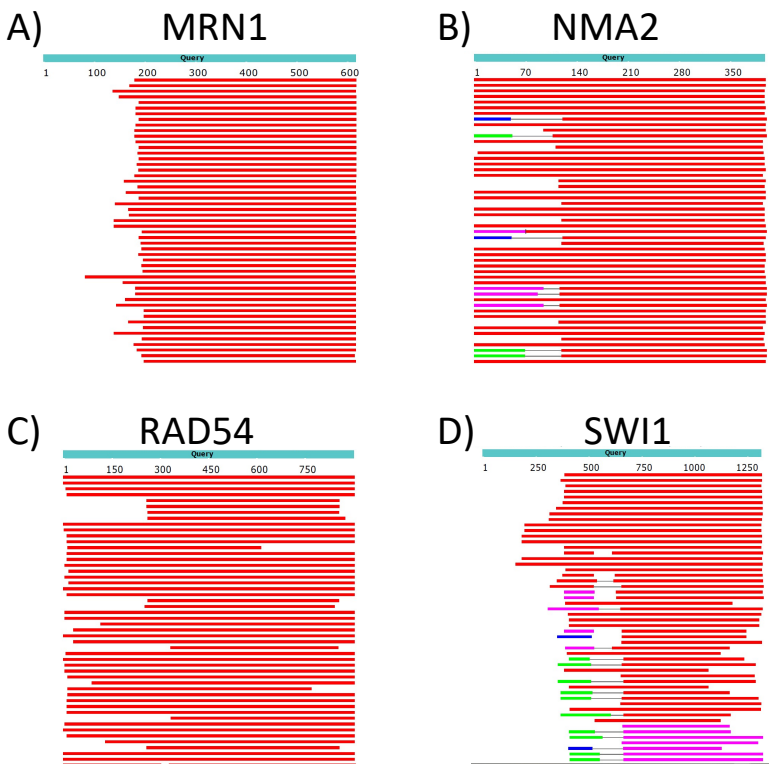

Figure 4. The N-termini of proteins are relatively variable. NCBI BLAST of example proteins was performed online again *Saccharomycotina* (taxid:147537) excluding *Saccharomyces* (taxid:4930). All other parameters were set to default. NCBI uses a color-coded spectrum of homology, where red indicates highest homology between orthologues, and white indicates no homology.

Method of scoring N-terminal conservation, and rationale for using Saccharomycotina.

We continued our investigations into the relationship between N-termini, protein conservation, and translation speed using a more systematic approach. We designed an algorithm to run local protein-protein Blast for all *S. cerevisiae* genes against a database consisting of the *Saccharomycotina* subphylum, omitting *Saccharomyces cerevisiae*. *Saccharomycotina* was chosen because almost every gene from *S. cerevisiae* has a recognizable, conserved homolog in almost every species in *Saccharomycotina*, and yet the evolutionary distances are long enough that there is considerable sequence variability. We excluded species of *Saccharomyces* as they are too closely related to *S. cerevisiae*, and they are also very numerous in the database, and would overwhelm results from the other members of *Saccharomycotina*. We believe, however, that this particular choice of subphylum does not greatly affect the final result.

Conservation at the N-terminus was calculated as the weighted proportion of yeast species with sequence matches (a match by default Blast parameters) beginning in the first 40 amino acids. The lowest conservation score is 0 (no hits in the first 40 amino acids), whereas the highest conservation score is 40 indicating that every species had a match (default Blast parameters) starting at the first amino acid (Table 1, Supplementary Table 1). The length of genes is correlated with the N-termini conservation score ($r=-0.34$; $p<0.001$)—that is, short genes tended to have more conserved N-termini.

Protein	Total Hits	Hits in the first 40 amino acids	Query Start	Conservation Score
PCA1	68	0	237-726	0
NSR1	67	0	61-169	0
SNX41	66	35	12-152	10.2
BNA2	54	54	17-33	20.8
TRP3	69	69	2-16	30.2
RPS8B	66	66	1	40

Table 1: Example conservation scores. Scores were calculated as described in Table S1. Scores ranged from 0 to 40, with proteins that had sequence hits only outside of the first 40 amino acids getting a value of 0, and proteins where all orthologs had homology beginning at the first amino acid a score of 40. Note that "homology" is by BLAST defaults, and does not imply 100% identity. For example, *PCA1* and *NSR1* both had many yeast orthologs with high homology, but none of the homology was within the first 40 amino acids. Conversely, *RPS8B* had the highest possible score (40) with each of the 66 orthologs having homology beginning at the first amino acid. While all 54 orthologues of *BNA2* had homology beginning in the first 40 amino acids, homology began between the 17th and 33rd amino acid positions. This approach was applied for all yeast proteins that returned high-homology hits, creating a spectrum of protein conservation (Fig. 5).

N-termini are variable and poorly conserved.

We measured protein conservation across more than three thousand *S. cerevisiae* proteins with orthologues among 419 closely related yeast species from *Saccharomycotina*. For each protein, we developed a conservation score for the first 40 amino acids to represent the N-termini, and an equivalent conservation score for the middle 40 amino acids, representing the body of the protein.

Strikingly, the N-termini of *S. cerevisiae* orthologs had conservation scores that were much lower than the middle of the same orthologs (Fig. 5). The first 40 amino acids had a flat distribution of protein conservation scores, with a prominent peak at 0, indicating high levels of variability amongst these orthologs. That is, many of the orthologs had no detectable homology with the first 40 amino acids of the *cerevisiae* protein. Conversely, and in striking contrast, the middle 40 amino acids of the orthologs were very highly conserved, with conservations scores peaking at 40, the highest possible score. It Is evident that for the middle 40 amino acids, the vast majority of orthologs, over every species, had a region of high homology to the *S. cerevisiae* protein, whereas this is not true for the N-termini.

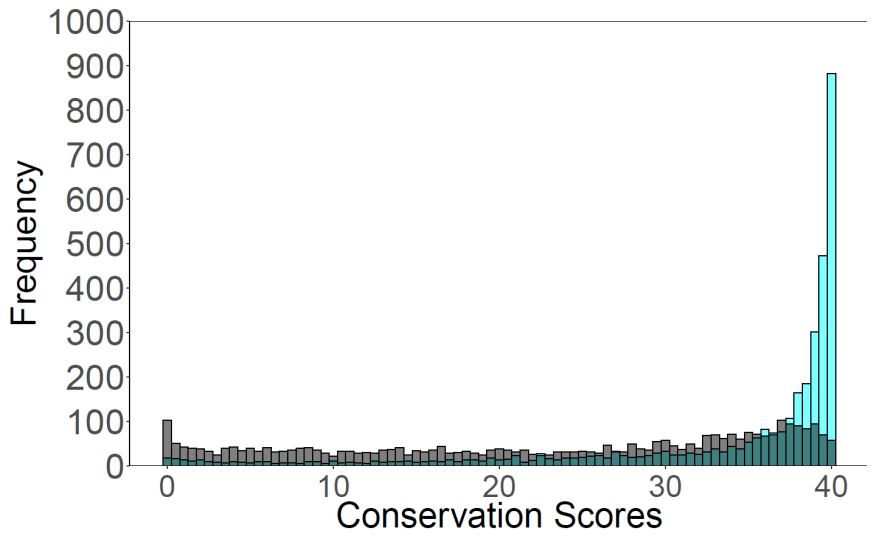

Figure 5. Conservation at N-termini is low. Protein BLAST was done for every *S. cerevisiae* protein against the *Saccharomycotina* subphylum. For each gene, conservation scores for the first 40 amino acids are shown as the gray distribution, and conservation scores for the middle 40 amino acids are shown as the cyan distribution.

This result is highly consistent with the idea that N-termini are highly variable, whilst middles are conserved. Thus, we speculate that the N-termini are relatively young in evolution and have been formed from new nucleic acid sequences likely to include rare codons. Selection has not yet had enough time to fully remove these rare codons.

5' translation speeds positively correlate with 5' conservation scores.

If the "N-terminal Instability" hypothesis is true, and slow 5' translation is caused by protein instability of 5' ends, then there should be a correlation (in both directions) between slow translation, and poor N-terminal conservation. Our model predicts the newest N-termini to have the lowest conservation as well as the rarest codons, which corresponds with the slowest translation.

To test this, we first calculated the ratio of conservation of the first 40 codons to the conservation of the middle 40 codons for each protein. The lower the ratio, the more variable the N-terminus is compared to the middle of the gene.

We then chose the 10% of genes with the lowest conservation ratios; the 10% of genes with the middle-most conservation ratios, and the 10% of genes with the highest conservation ratios. We then found the relative initial translation speed for each of these three groups. As shown in Fig. 6A, the genes with the most poorly conserved N-termini also had the slowest initial translation, while the genes with the most conserved N-termini had the fasted initial translation.

We also looked at the correlation in the other direction (Fig. 6B). Results are less clear cut, but nevertheless there is a trend towards the genes with the slowest initial translation having the lowest N-terminal conservation ratios. Thus, overall, there is a strong correlation between N-terminal instability (i.e., newness in evolution) and slow initial translation (i.e., the presence of slow/rare codons).

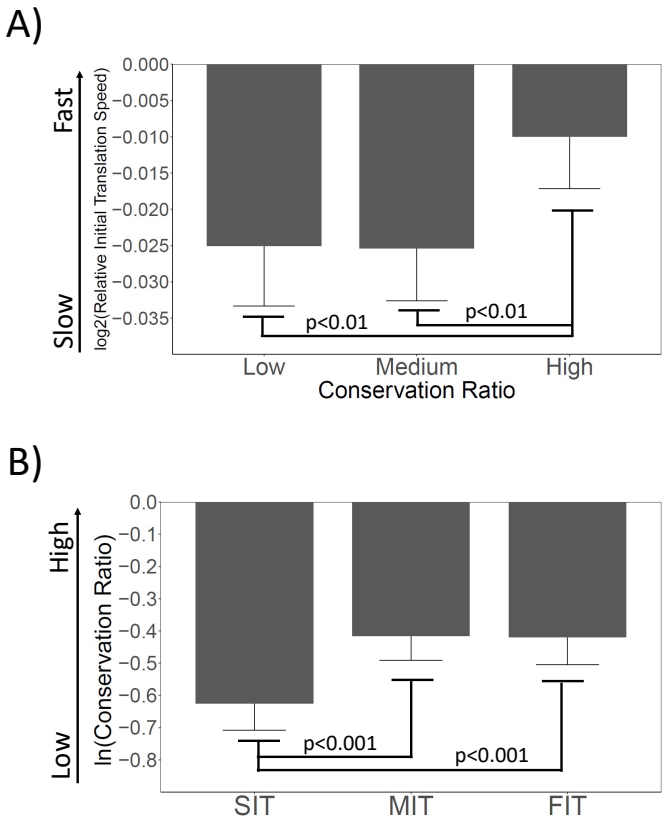

Figure 6. Slow initial translation and poor N-terminal conservation are correlated. A) Genes chosen to have the lowest N-terminal conservation also have the slowest N-terminal translation. B). Genes chosen to have the slowest N-terminal translation also have the lowest N-terminal conservation. SIT, Slow Initial Translation; MIT, Medium Initial Translation; FIT, Fast Initial Translation.

Slow initial translation fails to predict ribosome density and gene expression.

In the Tuller "Ramp" hypothesis, genes with the highest ribosome occupancy would presumably be in the most danger of ribosome collisions, and would therefore presumably have the most pronounced SITs. To test this theory, we used the Arava et al. 2003 dataset which measured the density of ribosomes on all *S. cerevisiae* mRNAs (Arava et al., 2003). We selected genes corresponding with the 10% highest, 10% middle-most, and 10% lowest ribosome densities. Contrary to the expectation from the Tuller et al. "Ramp" theory, the genes with the highest ribosome densities had the fastest initial translation, whereas the genes with the lowest ribosome densities had the slowest initial translation. Specifically, mRNAs with the highest ribosome densities had an average initial translation speed that was about 44% faster than genes with the lowest ribosome densities (SIT score of 0.0193 vs 0.0278 respectively, smaller numbers are faster). High-ribosome density mRNAs had an average initial translation speed about 8% faster than mRNAs with the middle ribosome densities (SIT score of 0.0193 vs 0.021). While these findings are opposite to the expectation of the "Ramp" theory, they are consistent with the spandrel "N-terminal Instability" theory, because genes with high ribosome density would be subject to more intense selection against slow codons, thus leading to faster 5' ends.

Similarly, in the "Ramp" hypothesis, highly-expressed genes would presumably be in more danger of ribosome collisions than poorly-expressed genes, and would therefore have the most pronounced SITs. To test this, we used the Lipson et al. 2009 transcriptome dataset which measured the number of mRNA transcripts for *S. cerevisiae* genes (Lipson et al., 2009). Again, we chose the genes with the 10% most mRNA transcripts, 10% middle-most, and 10% fewest transcripts (but above a floor of 10 transcripts). Contrary to the "Ramp" hypothesis, we found that genes with the highest expression had a mean SIT score

0.0166, more than 40% faster than the middle-most genes (SIT score = 0.0247). The high expression genes were also marginally faster than the poorly expressed genes (SIT scores 0.0166 vs 0.0171). The most poorly-expressed genes may not be representative, as they could be genes that are essentially off, but highly expressed under some other condition.

Experimentally, slow initial translation does not increase gene expression; the opposite is true.
Tuller et al. hypothesized that slow initial translation was adaptive, and improved efficiency of translation and improved gene expression, possibly by minimizing ribosome collisions. However, in our "N-terminal Instability" hypothesis, the slow initial translation is a spandrel that is not generated by selection and is not adaptive. It would not necessarily have any effect on gene expression, but, if anything, one might expect that slow translation would reduce gene expression. In this section, we present direct experimental results regarding the effect of slow, medium, and fast initial translation on gene expression.

We used a gene expression reporter closely based on EKD1024 (Brule, Dean, & Grayhack, 2016) (Methods and Materials). In this construct, GFP is the test reporter, but for accuracy it is normalized against a divergently transcribed red fluorescent protein (Fig. S3). Thus, GFP expression is reported as a GFP/RFP ratio. Although the test reporter is GFP, the N-terminal region of this particular protein is derived from yeast *HIS3*, not GFP, and has little if any effect on the fluorescence of the GFP fused downstream. We used synonymous slow, medium, or fast codons to recode the first 40 amino acids of this GFP reporter to generate three reporters with slow, medium, or fast translation over the first 40 amino acids. The amino acid sequences of the three constructs were identical. The slow, medium, and fast RRT values over the first 40 codons were 1.20, 1.04, and 0.93, respectively. That is, this SIT is slower than most naturally-occurring SITs, and this FIT is faster than most naturally occurring FITs, but the difference is not extreme.

As shown in Fig. 7 (left three bars), the SIT did not improve expression of GFP. In fact, the gene with the SIT was expressed at a significantly reduced level—about 30% lower--than the gene with the FIT. It was surprising to us that the difference was this large—again, only the first 40 codons were synonymously recoded, and the protein sequences were identical.

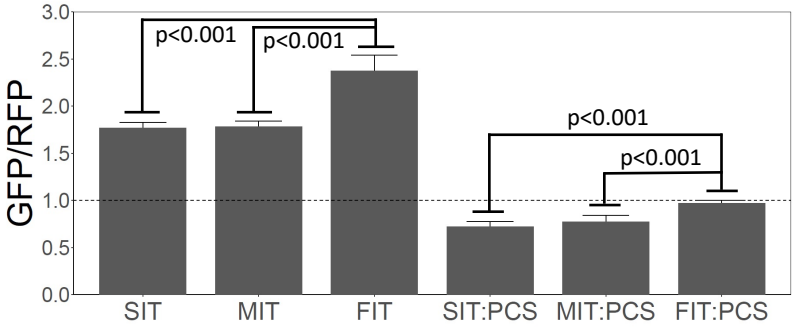

Figure 7. Slower initial translation does not increase gene expression. The first 40 codons of the GFP reporter were recoded to have slow (SIT), medium (MIT), or fast (FIT) translation. Within each set of three constructs, amino acid sequences were identical. The PCS constructs had a Putative Collision Site (PCS) (CGA-CGG) inside the gene. Each bar averages seven measurements (independent transformants).

Another possibility is that a SIT can protect against ribosome collisions when there is a site downstream that induces ribosome collisions. A site thought to induce ribosome collisions is the rare Arg-Arg codon pair CGA-CGG (Dao Duc & Song, 2018; Gamble, Brule, Dean, Fields, & Grayhack, 2016; Tesina et al., 2020). We therefore introduced this codon pair (replacing Asn-Asp, AAT-GAT) downstream of the first 40 amino acids, but still upstream of important GFP residues. Indeed, this single Arg-Arg pair, potentially inciting ribosome collisions, caused a large reduction (more than 50%) in expression of GFP (Fig. 7). However, this reduction was the same in the SIT, MIT, and FIT constructs. That is, the SIT did not at all protect against

these putative ribosome collisions. Instead, the fastest-translating gene once again gave the highest expression.

Discussion

Tuller et al. found that the 5' ends of genes are translated slowly, and posited that this was a selective advantage because it somehow increased efficiency of translation. However, this theory predicts positive correlations between slow initial translation and high gene expression, and slow initial translation and high ribosome density. In fact, by informatic analysis of existing data, and by direct experiments on gene expression, we find these predictions are not correct, and in fact the correlations are opposite to those predicted by the "Ramp" model.

The observation that 5' ends are unstable in evolution provides a completely different explanation for the region of slow initial translation. In this "N-terminal Instability" hypothesis, N-termini frequently change in evolution, gathering new 5' sequences *de novo.* These would contain all codons at similar frequencies—i.e., rare codons would not be especially rare. Although these rare codons would eventually be removed by selection, the fact that N-termini are relatively young means that this process might not be complete for all genes, and so some fraction of rare, slow codons still remain. These explain the initial region of slow translation. This hypothesis is consistent with the observed negative correlations between slow initial translation and high gene expression; and slow initial translation and high ribosome density, and with the results of gene expression experiments.

We have looked at the conservation of N-termini only in S. cerevisiae. However, our theory implies that a similar rapid turnover of 5' ends occurs in most eukaryotes, since most eukaryote genes have an initial region of slow translation. This in turn has implications for protein structure and evolution; for the interpretation of evolutionary sequence clocks; and for the rates of selection against rare codons.

It was surprising to us that recoding the first 40 codons of the GFP fusion protein from slow to fast increased the level of GFP expression by so much—25% to 30%. This is in the range seen for fully recoding a gene to preferred codons (Presnyak et al., 2015). Possibly, the increased expression achieved by fully recoding a gene is largely due to the recoding at the beginning of the gene. This could increase accessibility of the cap to the ribosome, for example.

Finally, in "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme", Gould and Lewontin (Gould & Lewontin, 1979) warned that not all biological phenomena are adaptive, or even a direct product of selection. We believe the slow initial translation of eukaryotic genes is an example of this.

Methods and Materials

Statistical calculations of relative initial rate of translation

Bioinformatics were performed on protein-coding open reading frames (ORFs) of *Saccharomyces cerevisiae* downloaded from the Saccharomyces Genome Database (SGD). All statistics were performed using The R Project for Statistical Computing. Only ORFs verified to produce protein were queried; dubious ORF were eliminated. Translation speed was measured using the ribosome residence time (RRT) which is a metric of the occupancy of ribosomes on each sense codon within the A-site (Table S1) (Gardin et al., 2014). Each sense codon has an RRT value that is a measurement of translation speed. For each gene, the relative initial rate of translation (RIT) was calculated across windows of the first 30, 40, 50, and 70 codons, which correspond with Tuller's window of reduced translational efficiency (see below).

$$RIT = \log 2 \frac{mean\ RRT\ (Codons\ 1{:}\ window)}{mean\ RRT\ (Codons\ (window + 1){:}\ end)}$$

We found that the first 40 codons had the greatest reduction in translation speed, and so we focused on these. All genes shorter than 132 nucleotides were omitted from analyses. Uncharacterized ORFs less than 300 nucleotides were omitted from analyses. For all RIT analyses, 180 out of 5917 ORFs were omitted leaving a total dataset size of 5737 ORFs.

Datasets

Ribosome density measurements were acquired from Arava et al. 2003 (Arava et al., 2003). mRNA transcript measurements, which we used as a proxy for gene expression, was acquired from Lipson et al. 2009 (Lipson et al., 2009).

Protein BLAST setup and diagnostics

S. cerevisiae proteins were downloaded from the SGD. The *Saccharomycotina* and *Saccharomyces* protein sequences were downloaded from NCBI using the weblinks https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=147537&lvl=3&lin=f&keep=1&srchmode=1&unlock

https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=4930&lvl=3&lin=f&keep=1&srchmode=1&unlock

We downloaded and compiled the source databases from DDBJ, EMBL, Genbank, RefSeq, and PIR. Duplicate sequences were deleted. To perform local BLAST, we downloaded the NCBI BLAST software version 2.12.0+ and used RStudio as a wrapper to operate the software; all default BLAST parameters were selected. Local protein BLAST of every *S. cerevisiae* protein was performed on *Saccharomycotina* omitting all species in the genus *Saccharomyces*; conserveness of proteins could be assessed between closely related, yet divergent yeast species without the *Saccharomyces* genus skewing observations. We eliminated submissions of duplicate species by limiting our database to the highest bitscores from sequence hits derived from each unique species. We were only interested in sequences that had high homology with queried *S. cerevisiae* genes, so all hits with bitscores below 50 were omitted. We used this approach to BLAST the full length of every *S. cerevisiae* query. Ultimately, we wanted to compare conservation at the beginning of proteins with conservation at the middle of those same proteins. Using the same method previously described, BLAST of the middle regions was performed by starting the query at the midpoint of each gene. For example, the length of Swi5 is 709 amino acids, and therefore BLAST of the middle region began at amino acid 355. In total, the BLAST of full-length *S. cerevisiae* proteins yielded 322978 high homology sequence matches across 419 unique species, whereas the blast of the middle proteins yielded 284664 high homology matches across 410 unique species.

Calculations of protein conservation scores and ratios

Conservation scores for the first 40 and middle 40 amino acids were calculated as the sum of weighted proportion of sequence matches within that window (Table S1). However, we only considered conservation scores from proteins that had homology with at least 40 unique species. This was because a high degree of homology among a few species is not equal to the same score across many species. We

also omitted proteins that were shorter than 200 amino acids, because BLAST at their midpoint might be too short to provide accurate conservation scores. As shown in Table 1, conservation scores ranged from 0 indicating no homology within the window of 40 amino acids, to 40 meaning that all orthologues had matches at the first amino acid. In total, our analyses on protein conservations focused on 3564 *S. cerevisiae* proteins with high homology hits at the beginnings and middle of proteins (Fig. 5).

We calculated the logarithmic-scaled ratios of the beginning and middle scores to reflect how conserved the N-termini of the protein is compared to the body of genes (Fig. 6). However, some proteins had N-termini or middle conservation scores of 0, and ln(0) gives an infinite value. To overcomes this problem, we calculated the mean of the 10% bottom (or worst) conserved N-termini scores and added this value to all beginning and middle conservation scores. After this approach, it was possible to calculate conservation ratio scores.

Design of the fluorescent reporter gene constructs

We created a reporter gene based on the original design of EKD1024 by Brule, Dean, and Grayhack 2016 (Brule, Dean, & Grayhack, 2016). Briefly, a bidirectional galactose reporter simultaneously regulates the expression of GFP and RFP in the presence of galactose. The reporter also has the *MET15* gene as a selected marker. Both *MET15* and RFP are in the reverse complement, whereas GFP is in the normal orientation. We separate the reporter construct into left, right, and middle pieces and then co-transform yeast with all pieces at the same time. We never change the left or right pieces, but we recode the middle piece (containing part of GFP) for the experimental. The middle piece was recoded to give the first 40 codons of GFP either a slow initial translation (SIT), medium initial translation (MIT), or fast initial translation (FIT) while preserving amino acid sequence. We also designed three more constructs (Fig. S3) with a SIT, MIT, or FIT but in the presence of one of the slowest and rarest codon pairs, CGA-CGG (replacing AAT-GAT, Asn-Asp), which is known to greatly attenuates gene expression in living yeast (Gamble, Brule, Dean, Fields, & Grayhack, 2016). Sequences are shown in Fig. S4. Other rare codon pairs (CGA-CGA and CGA-CCG) have been shown to promote ribosome stalling (Tesina et al., 2020), so in Fig. S3, right, CGA-CGG operates as a putative ribosome collision site (PCS). Instead of a PCS, the constructs in Fig. S3, left, had AAT-GAT (Asn Asp), a frequent codon pair with above average translation speed. The RIT score of SIT was 0.149; MIT was -0.059; FIT was -0.225; SIT+PCS was 0.133; MIT+PCS was-0.075; and FIT+PCS was -0.241.

Yeast Strains

The constructs were transformed in BY4741 (*MATa*; *his3Δ1*; *leu2Δ0*; *met15Δ0*; *ura3Δ0*). The reporter expresses *MET15* which permits successfully transformed cells to grow in the absence of methionine. Transformants were selected on HULA plates (0.075g/L Histidine; 0.075g/L Uracil; 0.25g/L Leucine; 0.075g/L Adenine; 20g/L D-Glucose; 5g/L Ammonium Sulfate; 1.7g/L Yeast Nitrogen Base). The reporter gene integrates into the *ADE2* locus, and thus successful transformants are *ade2-delete* which has a red phenotype when overgrown on YPD plates; this was used as a secondary biological marker to confirm successful transformants. A total of 7 transformants were selected corresponding with each GFP reporter construct.

Flow Cytometry analysis

The strains were inoculated in liquid HULA media, with 2% galactose, for 10 hours until they all had an OD600 around 0.3. The LSR Fortessa Flow Cytometer was used to measure GFP levels (P1 FITC-A Mean) and RFP levels (P1 PE-Texas Red-A Mean) across 75000 events. As a control, a strain lacking a fluorescent reporter was used, which had minimal P1 FITC-A Mean or P1 PE-Texas Red-A Mean readings. For all samples, GFP levels were normalized by RFP levels to reduce background fluorescent noise.

Statistical tests

None of our analyses made assumptions regarding the normality of our data. As such, we only performed nonparametric statistics. Wilcoxon signed-rank tests were done for relevant pairwise analyses; when necessary, p-values were corrected for multiple comparisons using the Holm–Bonferroni method (Fig. 1, Fig. 6, Fig.7). Spearman correlations were calculated for Fig. 2A, Fig. 2B, and Fig. S2. We used the kolmogorov smirnov goodness of fit test to confirm that the two distributions were significantly different ($p<0.001$) for Fig. 5.

Acknowledgements

We thank Steve Ketchum and Sangeet Honey for discussions that helped form the central idea of this manuscript.

Funding

This work was funded by NIH grant R01GM127542-04 to BF.

Competing Interests.

The authors declare no competing interests.

References

Arava, Y., Wang, Y., Storey, J. D., Liu, C. L., Brown, P. O., & Herschlag, D. (2003). Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A, 100*(7), 3889-3894. doi:10.1073/pnas.0635171100

Brule, C. E., Dean, K. M., & Grayhack, E. J. (2016). RNA-ID, a Powerful Tool for Identifying and Characterizing Regulatory Sequences. *Methods Enzymol, 572*, 237-253. doi:10.1016/bs.mie.2016.02.003

Dao Duc, K., & Song, Y. S. (2018). The impact of ribosomal interference, codon usage, and exit tunnel interactions on translation elongation rate variation. *PLoS Genet, 14*(1), e1007166. doi:10.1371/journal.pgen.1007166

Gamble, C. E., Brule, C. E., Dean, K. M., Fields, S., & Grayhack, E. J. (2016). Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast. *Cell, 166*(3), 679-690. doi:10.1016/j.cell.2016.05.070

Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S., & Futcher, B. (2014). Measurement of average decoding rates of the 61 sense codons in vivo. *Elife, 3*. doi:10.7554/eLife.03735

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc R Soc Lond B Biol Sci, 205*(1161), 581-598. doi:10.1098/rspb.1979.0086

Gritsenko, A. A., Hulsman, M., Reinders, M. J., & de Ridder, D. (2015). Unbiased Quantitative Models of Protein Translation Derived from Ribosome Profiling Data. *PLoS Comput Biol, 11*(8), e1004336. doi:10.1371/journal.pcbi.1004336

Lareau, L. F., Hite, D. H., Hogan, G. J., & Brown, P. O. (2014). Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *Elife, 3*, e01257. doi:10.7554/eLife.01257

Lipson, D., Raz, T., Kieu, A., Jones, D. R., Giladi, E., Thayer, E., . . . Causey, M. (2009). Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol, 27*(7), 652-658. doi:10.1038/nbt.1551

Presnyak, V., Alhusaini, N., Chen, Y. H., Martin, S., Morris, N., Kline, N., . . . Coller, J. (2015). Codon optimality is a major determinant of mRNA stability. *Cell, 160*(6), 1111-1124. doi:10.1016/j.cell.2015.02.029

Sharma, A. K., Sormanni, P., Ahmed, N., Ciryam, P., Friedrich, U. A., Kramer, G., & O'Brien, E. P. (2019). A chemical kinetic basis for measuring translation initiation and elongation rates from ribosome profiling data. *PLoS Comput Biol, 15*(5), e1007070. doi:10.1371/journal.pcbi.1007070

Tesina, P., Lessen, L. N., Buschauer, R., Cheng, J., Wu, C. C., Berninghausen, O., . . . Green, R. (2020). Molecular mechanism of translational stalling by inhibitory codon combinations and poly(A) tracts. *EMBO J, 39*(3), e103365. doi:10.15252/embj.2019103365

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., . . . Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell, 141*(2), 344-354. doi:10.1016/j.cell.2010.03.031

Tunney, R., McGlincy, N. J., Graham, M. E., Naddaf, N., Pachter, L., & Lareau, L. F. (2018). Accurate design of translational output by a neural network model of ribosome distribution. *Nat Struct Mol Biol, 25*(7), 577-582. doi:10.1038/s41594-018-0080-2

Wang, H., McManus, J., & Kingsford, C. (2017). Accurate Recovery of Ribosome Positions Reveals Slow Translation of Wobble-Pairing Codons in Yeast. *J Comput Biol, 24*(6), 486-500. doi:10.1089/cmb.2016.0147