SELECTION ON READTHROUGH PRODUCTS IN *SACCHAROMYCES*: IMPLICATIONS FOR PREADAPTATION AND CAPACITANCE

By

TAYLOR AUSTIN KESSINGER

A Thesis Submitted to The Honors College

In Partial Fulfillment of the Bachelor's degree
With Honors in

Ecology and Evolutionary Biology

THE UNIVERSITY OF ARIZONA

May 2009

Approved by:

Joanna Masel
Ecology and Evolutionary Biology

STATEMENT BY AUTHOR

I hereby grant to the University of Arizona Library the nonexclusive worldwide right to reproduce and distribute my thesis and abstract (herein, the "licensed materials"), in whole or in part, in any and all media of distribution and in any format in existence now or developed in the future. I represent and warrant to the University of Arizona that the licensed materials are my original work, that I am the sole owner of all rights in and to the licensed materials, and that none of the licensed materials infringe or violate the rights of others. I further represent that I have obtained all necessary rights to permit the University of Arizona Library to reproduce and distribute any nonpublic third party software necessary to access, display, run, or print my thesis. I acknowledge that University of Arizona Library may elect not to distribute my thesis in digital format if, in its reasonable judgment, it believes all such rights have not been secured.

SIGNED: ______________________________

Abstract

The misreading of stop codons as sense and subsequent translation of 3′ UTR content is described extensively in the molecular biology literature. However, little is known about the extent to which this permits natural selection to act on 3′ UTR information. We observe that, in *Saccharomyces cerevisiae*, in-frame "backup" stop codons, which provide a second chance for translation to terminate, frequently appear closer to the start of the 3′ UTR than predicted by random chance. Also, close backup stop codons tend to be favored by evolution and are conserved between *S. paradoxus* and *S. cerevisiae*, as well as between *Mus musculus* and *Rattus norvegicus*. Moreover, the ratio K_A/K_S is less than 1 in yeast 3′ UTR sequences. This suggests that low-level natural selection favors backup stops not only for metabolic reasons, but also due to purifying effects at the protein-coding level. This low-level selection on a form of mostly-cryptic variation is a sufficient condition for preadaptation to occur and implies that the yeast prion [PSI] plays a role in evolutionary capacitance (Masel 2006).

Preface

The author wishes to make clear that portions of this thesis encompass work performed by preceding students which are nonetheless essential to the current project. Specifically, the section on data acquisition represents work performed by Adam Hancock and Jason Slepicka, and the sections on sequence alignment, K_A/K_S analysis, statistical methods, and selection at the coding level in yeast 3′ UTRs represent work performed by Jason Slepicka.

Introduction

During translation, peptide synthesis is generally terminated upon the arrival of the ribosome at one of three stop codons: UAA, UAG, or UGA. However, misreading of these stop codons as sense occurs in eukaryotes as a result of competition between release factors and near-cognate tRNAs. In yeast, this occurs at a frequency of about 0.3% (Firoozan et al. 1991), but in some genes the frequency at which this misreading occurs is as high as 23% (Namy, Duchateau-Nguyen and Rousset 2002) depending partially on the identity of the stop, with UAA more reliably terminating translation than either UAG or UGA (Firoozan et al. 1991). The nucleotides upstream (Tork et al. 2004) and downstream (Namy, Hatin and Rousset 2001) of the stop also play a role in determining the readthrough rate.

When a stop codon is misread in this fashion, the nascent peptide is elongated as the ribosome reads codons in the mRNA molecule's 3′ UTR. When two ORFs are adjacent, this can lead to the production of hybrid proteins; however, in other cases this elongation process has a negative effect on the peptide. Elongating a peptide beyond its normal N-terminal site means spending additional energy, and if the peptide is rendered useless as a result, then all of the energy involved in producing the peptide is wasted. Moreover, the peptide may become toxic if improper 3′ UTR data is translated.

Natural selection favors the UAA stop codon which is less leaky, as well as other mechanisms to mitigate readthrough. It has been proposed that the presence of tandem or "backup" stop codons guards against readthrough by providing organisms with a failsafe in the event that a stop codon is misread (Nichols 1970). In prokaryotes, tandem or "backup" stop codons are frequently found at the +1 position past the primary stop. However, their preponderance may be traceable to readthrough-inhibiting features of the 3′ UTR such as the high frequency of U nucleotides after the stop (Major et al. 2002). In *S. cerevisiae*, backup stop codons exist in high frequency at the +1, +2, and +3 positions past the primary stop (Williams et al. 2004). Conservation of codons in the 3′ UTR also depends on whether or not the codon is a stop (Liang, Cavalcanti and Landweber 2005).

We propose that close backup stop codons are useful even if they are not immediately adjacent to the primary stop; this is because there is a negative correlation between the functionality of a peptide and the number of amino acids added to its N-terminus. Cells expend a significant portion of their energy on translation; consequently, faulty translation represents lost energy. Moreover, adding a greater number of amino acids to the N-terminus of a peptide increases the probability that a peptide's function will be compromised.

An additional feature of translation which backup stop codons serve to minimize is non-stop decay; if a ribosome reaches the end of an mRNA molecule, attached ribosomes are freed and the mRNA is brought to the lysosome for destruction. This process is costly to the cell.

We seek to determine the extent to which natural selection drives the evolution of 3′ UTRs. Several hypotheses are tested. First, if close backup stops mitigate the deleterious effects of readthrough, then backup stops should be closer than predicted by random chance. Second, these close backup stops should be conserved during evolution. Genes with conserved backup stops may be genes for which readthrough is especially harmful.

Conservation of UTRs leading to maintenance of close backup stop codons may be driven mainly by metabolic constraints or by constraints at the protein level. To determine if the latter is this case, the ratio K_A/K_S can be obtained for 3′ UTR sequences after aligning the UTRs of *Saccharomyces cerevisiae* and *Saccharomyces paradoxus*, a close relative; a K_A/K_S ratio less than one indicates purifying selection at the protein-coding level rather than mere selection for metabolic reasons.

Furthermore, it is predicted that peptide elongation may have more negative effects in frequently-produced peptides than in less frequently-produced ones. This has already been evidenced by the abundance of less leaky stop codons in more frequently translated genes (Williams et al. 2004), and frequent backup stops in corresponding 3′ UTRs are another signal of this. Thus, frequently translated genes are predicted to be more likely to have backup stops and in particular close, conserved backup stops.

Methods and materials

Data acquisition

Sequences for 5883 *S. cerevisiae* genes were obtained from the *Saccharomyces* Genome Database (SGD) (release date: Dec. 12, 2007), and 3′ UTR annotations came from mRNA tiling array data analyzed by David et al. (2006). From this, 3267 *S. cerevisiae* 3′ UTR annotations were acquired, but the annotations for 72 genes marked "excluded" and 179 genes marked "untranscribed" were omitted. Thus, there were 3016 *S. cerevisiae* genes with usable 3′ UTR annotations.

Sequences for 5485 *S. paradoxus* genes were pulled from SGD (release date: Jul 21, 2005). Annotations for 5127 orthologs between *S.* cerevisiae and *S. paradoxus* were obtained from Kellis et al. (2003). For the 3016 *S. cerevisiae* genes with usable 3′ UTR annotations, there were 2878 orthologous genes in *S. paradoxus*.

Sequence data for the 8745 genes from the brown rat, *Rattus norvegicus*, were acquired from the UCSC Genome Browser (release date: June 2003). 18048 genes from the common house mouse, *Mus musculus*, were obtained from the UCSC Genome Browser. Ortholog annotations were retrieved from the Mouse Genome Database; there were 5910 orthologous genes. 3′ UTR data was not available for these genomes; however, the average *S. cerevisiae* 3′ UTR length is 90 bp (David et al. 2006), so 90-bp suffixes were used to approximate the 3′ UTR of each mouse gene.

Sequence alignment

Sequences were aligned with MCALIGN2.0, a software package for aligning non-coding sequences developed by Wang et al. (Wang, Keightley and Johnson 2006). It is optimal for the alignment of non-coding regions because it does not place a significant penalty on insertions and deletions against which there is far less selective pressure in non-coding regions. MCALIGN2.0 requires the user to supply θ, the observed ratio of point mutations to indels, and the distribution of indel length frequencies. It uses the Kimura model (K80), which introduces the parameter κ, the ratio of transitions to transversions.

We added an anchor to each 3′ UTR to limit possible end effects, which are especially significant when indels are frequent. POA (Lee, Grasso and Sharlow 2002) was used to generate an initial alignment of coding regions and their 3′ UTRs; from this alignment, it was determined where the coding regions ended. Then, the beginning of each 3′ UTR was padded with 24 matching base-pairs from the corresponding coding region, and the end of each 3′ UTR was padded with a 90-base pair suffix.

With these padded 3′ UTRs, MCALIGN2.0 was run multiple times. The ratio of indels to substitutions θ, the transition-transversion ratio κ, and the indel length frequency distribution from the initial alignment were fed into MCALIGN2.0, which produced a second alignment. These parameters were re-estimated from the new alignment and fed back into a new run of MCALIGN2.0. The process was iterated 8 times; after this, the parameters converged. For yeast, we obtained $\kappa = 5.51218$, $\theta = 0.163327$, and a frequency distribution seen in figure 1A. For the mouse/rat comparion, we found a $\kappa = 4.16240$, $\theta = 0.190763$, and a frequency distribution seen in figure 1B.

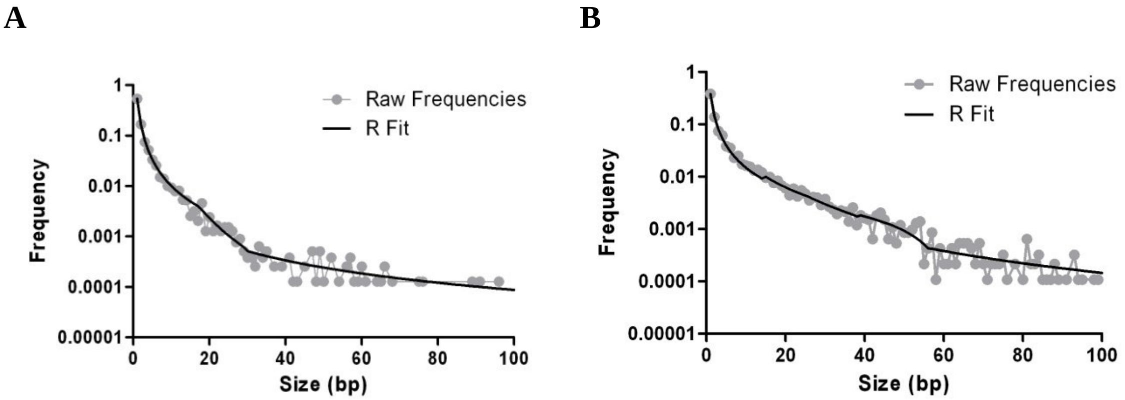

Fig. 1. Observed indel length frequency distributions for yeast (**A**) and mouse/rat (**B**), with the smoothed *R* fits shown.

To verify that additional coding regions were not present in the 3′ UTRs, a BLAST database of every known coding region in the relevant species was generated. 3′ UTRs where a 100% match was found were excluded from K_A/K_S analysis. This shortened the usable 3′ UTR sequence for many genes. Finally, the padding on the UTRs was removed before K_A/K_S analysis.

K_A/K_S analysis

Only sequences in the same frame could be considered, since sequences in different frames would be subject to different selective pressures. Thus, if an insertion or deletion set a 3′ UTR sequence in one species out of frame with its corresponding sequence in the other species, that portion of the sequence was omitted from analysis until another insertion or deletion placed the two sequences back in frame. Poorly aligned 3′ UTR sequences contained so many indels that they were effectively screened out by this frame criterion. A final set of 2588 usable 3′ UTRs was thus obtained for yeast. In rat/mouse, 4525 suffixes with usable in-frame data were acquired.

We used HyPhy, a phylogenetic package developed by Pond et al. (2005), to calculate K_A/K_S. HyPhy offered greater flexibility than other phylogenetic packages so that we could select which nucleotide and codon substitution models to use, as well as choose to include stop codons in our analysis. The Goldman-Yang model (GY94) was used. For π, the distribution of nucleotide frequencies, we used a 1x4 model in which only the global nucleotide frequency is used rather than separate frequencies for each frame. Some packages, such as PAML (Yang 1997) require stop codons to be removed. However, this was not appropriate for our analysis, and HyPhy was configured to treat stop codons just like any other set of synonymous codons.

We often simultaneously analyzed two categories or "partitions" of sequence, e.g. sequences with a uORF present versus sequences with no uORF present or the portions of sequence before and after a backup stop. HyPhy fitted different local values for the pertinent parameters of the different partitions.

We sought to remove uORFs from our analysis. However, even the two-codon sequence ATGTGA is technically a uORF. Removing all uORF data such as this would exclude many start codons from our analysis; since start codons do not have any synonymous in the genetic code, this would lead to an inflated value of K_A/K_S, as many nonsynonymous changes would vanish from our alignments. The 3′ UTR sequences were scanned and partitioned according to whether or not the sequences were within un-annotated ORFs, and the start codons of un-annotated ORFs within the UTRs were included as part of the rest of the 3′ UTR.

Finally, when K_A/K_S was analyzed in subsets of the data either with or without a conserved backup stop codon, the backup stop codon itself was removed from the K_A/K_S analysis, since its status as a synonymous or nonsynonymous substitution was predetermined.

Statistical methods

HyPhy uses maximum likelihood to obtain parameters which best fit the data; it can either fit all parameters or be constrained by some user-chosen parameters and fit the remainder. The parameters for the chosen substitution model were κ, the transition-transversion ratio, ω or K_A/K_S, and the frequencies of each nucleotide. The nucleotide frequencies were taken from the concatenated 3′ UTR set in question; when partitioned, the nucleotide frequencies for each partition were fixed at the global frequencies. Model comparisons were made by comparing two times the difference in the log likelihood to the χ^2 distribution. Selection was inferred by comparing a model with ω free to one with ω constrained to 1.

Results

In-frame yeast backup stop codons are slightly closer to the start of the 3′ UTR than out-of-frame backup stop codons

2920 *S. cerevisiae* 3′ UTR sequences, as annotated by David et al (2003), were parsed and scanned for backup stop codons in frame with the primary stop. As a control, the same scan was performed for backup stops in a +1-shifted frame; since most cellular readthrough involves the association of a stop codon with a near-cognate tRNA and not a frame-shift, little to no selection for stop codons should occur out of frame. On average, in-frame backup stops were 2.8 nucleotides closer (figure 2a) than out-of-frame stops, suggesting that there is selective pressure to force backup stop codons close to the primary stop ($p < 10^{-6}$; t-test).

Previous work (Williams et al. 2004) has demonstrated a marginal overabundance of stop codons at the +1, +2, and +3 codon positions; this was observed by comparing the frequency distribution of backup stops against a geometric distribution, which was obtained by calculating the probability that a backup stop will appear. However, one potential confounding factor in that study was the selection for stop codon contexts rich in U, G, and A nucleotides at those positions.

Our frequency distribution (figure 2b) of in-frame and out-of-frame stop codon distances confirms the previously observed overabundance of stops at the +2 and +3 position. However, excluding both in-frame and out-of-frame stop codons which were found to be closer than the +4 codon position reveals that the +3 and closer positions alone do not account for the position bias; in the +4 codon position and beyond, in-frame backup stops are 2.0 nucleotides closer than out-of-frame backup stops ($p < 10^{-3}$, t-test).

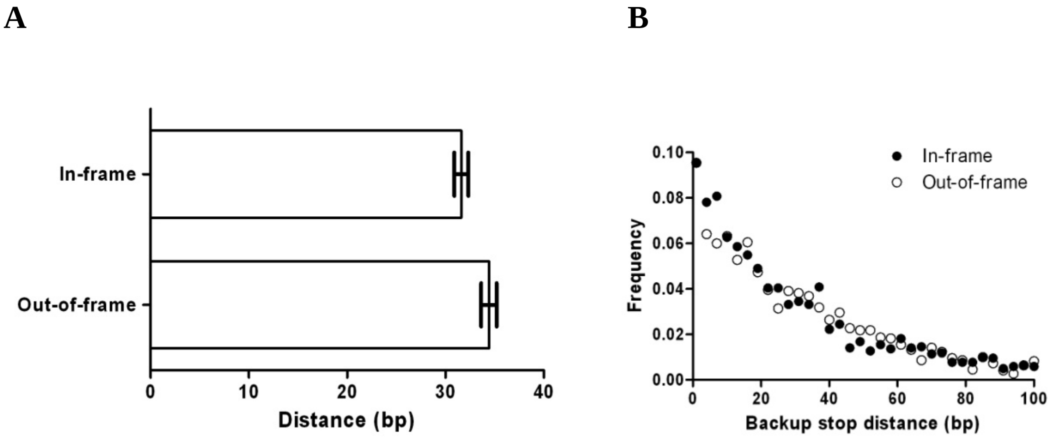

Fig. 2. Average distance to in-frame and out-of-frame backup stops (**A**) with S.E.M. error bars and observed distribution of backup stop codon distances (**B**).

The same method was used to observe if backup stop codons are closer than predicted by random chance in the genomes of human, mouse, chimpanzee, Drosophila, and *Arabidopsis*. No significant results were observed.

Conserved backup stop codons are much closer than nonconserved ones

The modest position bias of stop codons we observe may be due to strong effects at a subset of genes; moreover, evolutionary conservation of a feature can be an indication that the feature is favored by evolution. 1576 ORFs in *S. cerevisiae* were isolated whose backup stop locations are conserved in *S. paradoxus*, and 2781 ORFs were obtained whose backup stops were not conserved. Backup stops whose locations are conserved are 19 base pairs closer (figure 3a) than those whose location is not conserved ($p < 10^{-8}$; *t*-test). Conservation of codons at the +3 position has previously been found to depend on whether or not the codon is a stop (Liang et al. 2005); here we find strong results for all positions.

In a mouse/rat comparison, conserved backup stops are 25 nucleotides closer (figure 3b) than nonconserved ones ($p < 10^{-6}$, *t*-test), so this result is not unique to yeast.

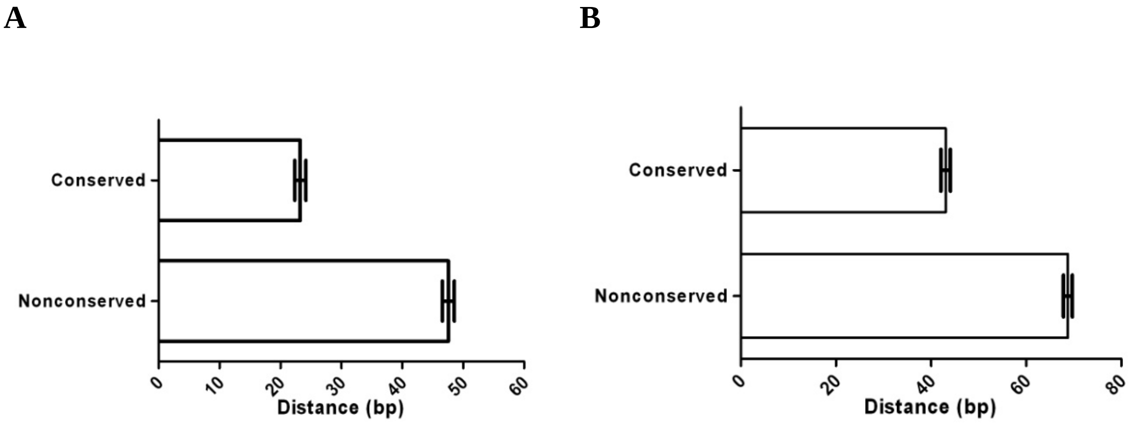

Fig. 3. Average distance to conserved and nonconserved in-frame backup stops between *S. cerevisiae* and *S. paradoxus* (**A**) and between *Mus musculus* and *Rattus norvegicus* (**B**) with S.E.M. error bars shown.

Highly expressed genes are under greater selection against readthrough

High ORF expression is a factor which suggests that readthrough of a stop codon is especially harmful. Yeast protein concentration levels from Ghaemmaghami et al. (2003) and mRNA concentration levels from Beyer et al. (2004) were obtained.

On average, genes with backup stop codons are more frequently expressed than genes without backup stops, and genes whose backup stops are conserved are more frequently expressed than genes whose backup stops are not conserved (figure 4).

However, no significant correlation is observed between distance to the backup stop and protein concentration (Kendall's τ = -.04, p > .05) or mRNA concentration (Kendall's τ = -.02, p > .05). No relationship was observed between the essentiality of a gene and the closeness or presence of associated backup stops.

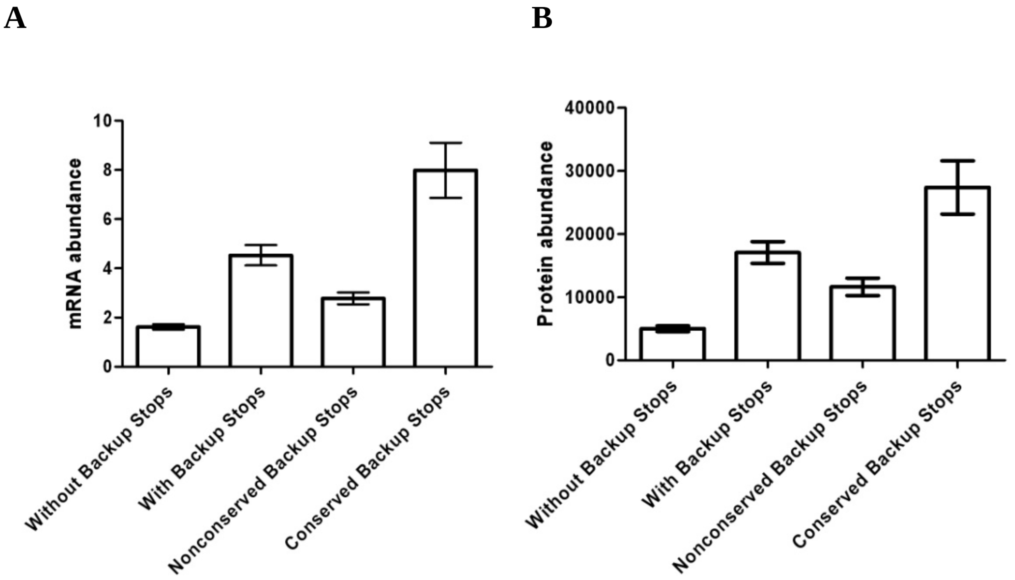

Fig. 4. Average expression level of genes with differing 3' UTR features. Expression level is shown in terms of mRNA abundance (**A**) and protein abundance (**B**), with S.E.M. error bars.

Selection occurs at the protein-coding level in yeast 3′ UTRs

Two possible hypotheses for the conservation of close backup stop codons are metabolic constraint and selection at the protein-coding level. Calculating K_A/K_S permits us to distinguish between these two hypotheses.

A K_A/K_S ratio of .875 was calculated for the sequences prior to conserved backup stop codons in the 3′ UTRs of *S. cerevisiae* and *S. paradoxus*, indicating ($p < .025$, χ^2 test) that purifying selection occurs at the protein-coding level in yeast 3′ UTRs. In contrast, K_A/K_S equaled .978 in a similar mouse-rat comparison and was not significantly different from 1, suggesting that the K_A/K_S value obtained for yeast is not an artifact of our protocol.

There is no reason to expect a non-unity K_A/K_S value other than selection at the protein-coding level. However, this selection need not be a result of readthrough. Genes can occasionally be excluded from annotation artificially owing to the heuristics used in characterizing them. Short ORFs are often not considered to correspond to genes and are excluded from K_A/K_S

analysis; however, since many genes have a K_A/K_S ratio around 0.1, it would not take many unannotated ORFs to wildly skew our K_A/K_S value.

Thus, to prevent short genes from influencing K_A/K_S, we separately analyzed all ORFs in the 3′ UTR. An analysis of yeast 3′ UTRs omitting all ORFs and including only the sequence prior to a conserved backup stop codon yielded a K_A/K_S value of 0.815 ($p < .0025$, χ^2 test).

Discussion

Preadapted variation corresponds to heritable phenotypic variation which is less likely to contain unconditionally deleterious variants. Preadaptation of cryptic variation can occur provided that the variation is not entirely cryptic, but only *mostly* cryptic. Small amounts of natural selection acting on this variation can serve to remove unconditionally deleterious variants, which is a sufficient condition for preadaptation (Masel 2006).

3′ UTR sequences clearly constitute a form of mostly but not entirely cryptic variation, since they are expressed in small amounts due to the readthrough of stop codons. Our work provides evidence that natural selection acts in small amounts on these sequences, which is a sufficient condition for them to be preadapted. Low levels of selection may be seen through the selective pressure for close backup stop codons and selection at the protein-coding level. If 3′ UTR sequences are later expressed in greater amounts, then either of these conditions renders it more likely to be adaptive and less likely to be unconditionally deleterious.

At least two mechanisms can increase the expression rate of 3′ UTR sequences. First, the yeast prion [PSI], an epigenetically inherited prion which increases the readthrough rate several fold, may appear. It has been previously suggested by Masel and Bergman (2003) that [PSI] may facilitate an increase in the evolvability of yeast via the expression of these UTR sequences.

Second, nonsynonymous substitutions at stop codons may lead to permanent in-frame assimilation of 3′ UTR data. This process may also be aided by [PSI]: Once preadapted cryptic variation is expressed by a revealing mechanism, natural selection may favor the removal of the variation’s dependence on the revealing mechanism (Masel 2005). Preadapting selection acting on 3′ UTRs is consistent with the fact that a disproportionately large number of in-frame 3′ UTR assimilation events have occurred in yeast (Giacomelli, Hancock and Masel 2007).

Acknowledgments

The author acknowledges Jimmy Matthews and Aron Talenfeld for their early work on backup stop codons, as well as Adam Hancock and Jason Slepicka for their foundational and related work on this project.. Funding was provided by NIH grant R01 GM076041 to Joanna Masel, UBRP grant R25 GM072733, and the BIO5 Institute. Joanna Masel is a Pew Scholar in the Biomedical Sciences.

References

Beyer, A., J. Hollunder, H. P. Nasheuer & T. Wilhelm (2004) Post-transcriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale. *Molecular & Cellular Proteomics,* 3**,** 1083-1092.

David, L., W. Huber, M. Granovskaia, J. Toedling, C. J. Palm, L. Bofkin, T. Jones, R. W. Davis & L. M. Steinmetz (2006) A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences of the United States of America,* 103**,** 5320-5325.

Firoozan, M., C. M. Grant, J. A. B. Duarte & M. F. Tuite (1991) Quantitation of readthrough of termination codons in yeast using a novel gene fusion assay. *Yeast,* 7**,** 173-183.

Ghaemmaghami, S., W. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea & J. S. Weissman (2003) Global analysis of protein expression in yeast. *Nature,* 425**,** 737-741.

Giacomelli, M. G., A. S. Hancock & J. Masel (2007) The conversion of 3 ' UTRs into coding regions. *Molecular Biology and Evolution,* 24**,** 457-464.

Kellis, M., N. Patterson, M. Endrizzi, B. Birren & E. S. Lander (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature,* 423**,** 241-254.

Lee, C., C. Grasso & M. F. Sharlow (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics,* 18**,** 452-464.

Liang, H., A. R. O. Cavalcanti & L. F. Landweber (2005) Conservation of tandem stop codons in yeasts. *Genome Biology,* 6**,** 8.

Major, L. L., T. D. Edgar, P. Y. Yip, L. A. Isaksson & W. P. Tate (2002) Tandem termination signals: myth or reality?, 84-89. Elsevier Science Bv.

Masel, J. (2005) Evolutionary capacitance may be favored by natural selection. *Genetics,* 170**,** 1359-1371.

Masel, J. (2006) Cryptic genetic variation is enriched for potential adaptations. *Genetics,* 172**,** 1985-1991.

Masel, J. & A. Bergman (2003) The evolution of the evolvability properties of the yeast prion [PSI+]. *Evolution,* 57**,** 1498-1512.

Namy, O., G. Duchateau-Nguyen & J. P. Rousset (2002) Translational readthrough of the PDE2 stop codon modulates cAMP levels in Saccharomyces cerevisiae. *Molecular Microbiology,* 43**,** 641-652.

Namy, O., I. Hatin & J. P. Rousset (2001) Impact of the six nucleotides downstream of the stop codon on translation termination. *Embo Reports,* 2**,** 787-793.

Nichols, J. L. (1970) Nucleotide Sequence from the Polypeptide Chain Termination Region of the Coat Protein Cistron in Bacteriophage R17 RNA. *Nature,* 225**,** 147-151.

Tork, S., I. Hatin, J. P. Rousset & C. Fabret (2004) The major 5 ' determinant in stop codon read-through involves two adjacent adenines. *Nucleic Acids Research,* 32**,** 415-421.

Wang, J., P. D. Keightley & T. Johnson (2006) MCALIGN2: Faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution. *Bmc Bioinformatics,* 7**,** 15.

Williams, I., J. Richardson, A. Starkey & I. Stansfield (2004) Genome-wide prediction of stop codon readthrough during translation in the yeast Saccharomyces cerevisiae. *Nucleic Acids Research,* 32**,** 6605-6616.