PLOS GENETICS

RESEARCH ARTICLE

Transcriptome-wide investigation of stop codon readthrough in *Saccharomyces cerevisiae*

Kotchaphorn Mangkalaphiban, Feng He, Robin Ganesan, Chan Wu, Richard Baker, Allan Jacobson*

Department of Microbiology and Physiological Systems, University of Massachusetts Medical School, Worcester, Massachusetts, United States of America

* allan.jacobson@umassmed.edu

Abstract

Translation of mRNA into a polypeptide is terminated when the release factor eRF1 recognizes a UAA, UAG, or UGA stop codon in the ribosomal A site and stimulates nascent peptide release. However, stop codon readthrough can occur when a near-cognate tRNA outcompetes eRF1 in decoding the stop codon, resulting in the continuation of the elongation phase of protein synthesis. At the end of a conventional mRNA coding region, readthrough allows translation into the mRNA 3'-UTR. Previous studies with reporter systems have shown that the efficiency of termination or readthrough is modulated by *cis*-acting elements other than stop codon identity, including two nucleotides 5' of the stop codon, six nucleotides 3' of the stop codon in the ribosomal mRNA channel, and stem-loop structures in the mRNA 3'-UTR. It is unknown whether these elements are important at a genome-wide level and whether other mRNA features proximal to the stop codon significantly affect termination and readthrough efficiencies *in vivo*. Accordingly, we carried out ribosome profiling analyses of yeast cells expressing wild-type or temperature-sensitive eRF1 and developed bioinformatics strategies to calculate readthrough efficiency, and to identify mRNA and peptide features which influence that efficiency. We found that the stop codon (nt +1 to +3), the nucleotide after it (nt +4), the codon in the P site (nt -3 to -1), and 3'-UTR length are the most influential features in the control of readthrough efficiency, while nts +5 to +9 had milder effects. Additionally, we found low readthrough genes to have shorter 3'-UTRs compared to high readthrough genes in cells with thermally inactivated eRF1, while this trend was reversed in wild-type cells. Together, our results demonstrated the general roles of known regulatory elements in genome-wide regulation and identified several new mRNA or peptide features affecting the efficiency of translation termination and readthrough.

Author summary

Of the 64 codons that exist for translation of the genetic code into a polypeptide chain, only UAA, UAG, and UGA signify termination of continued protein synthesis. However,

OPEN ACCESS

Citation: Mangkalaphiban K, He F, Ganesan R, Wu C, Baker R, Jacobson A (2021) Transcriptome-wide investigation of stop codon readthrough in *Saccharomyces cerevisiae*. PLoS Genet 17(4): e1009538. https://doi.org/10.1371/journal.pgen.1009538

Editor: Anita K. Hopper, Ohio State University, UNITED STATES

Received: December 9, 2020

Accepted: April 6, 2021

Published: April 20, 2021

Data Availability Statement: All data files underlying figures and R codes to plot them are here: https://github.com/Jacobson-Lab/sup45-ts_readthrough The raw sequencing data are deposited and available at Gene Expression Omnibus (GEO) under accession number GSE162780.

Funding: This work was supported by grants to A. J. (5R01 GM27757-37 and 1R35GM122468-04) from the U.S. National Institutes of Health and from the Cystic Fibrosis Foundation (CFFKOROST20GO).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: A.J. is co-founder, director, and research consultant for PTC Therapeutics Inc.

the efficiency of translation termination is not the same for different mRNAs and is likely influenced by mRNA sequences proximal to these "stop" codons. Here, we sought to expand current understanding of termination efficiency by measuring termination failures, or "readthrough" on a genome-wide scale. Our analysis identified novel mRNA features that may regulate the efficiency of readthrough, including the identities of the stop codon, the penultimate codon, the proximal nucleotides, and the length of the mRNA 3'-untranslated region. As ~11% of human genetic diseases are caused by mutations that introduce premature stop codons, resulting in both accelerated mRNA decay and truncated protein products, our findings are valuable for understanding and developing therapeutics targeting premature termination of mRNA translation.

Introduction

Translation of an mRNA into a polypeptide is terminated when the release factor eRF1 interacts with a UAA, UAG, or UGA stop codon in the ribosomal A site and another release factor, eRF3, hydrolyzes GTP and stimulates the polypeptide release activity of eRF1 [1–3]. However, at low frequency, a near-cognate tRNA (nc-tRNA; a tRNA with one base pair mismatch in its anticodon) outcompetes eRF1 in decoding the stop codon, resulting in the continuation of translation elongation. When the stop codon is located at its normal site at the end of an open reading frame (ORF) elongation will continue into the mRNA 3'-untranslated region (3'-UTR), producing a C-terminally extended polypeptide product. Such events are termed stop codon readthrough or nonsense suppression [1,4–6].

The process of readthrough was first characterized primarily in viruses, which utilize this mechanism to increase the protein-coding capacity of their compact genomes [7,8]. It is only quite recently that C-terminally extended yet functional protein isoforms have been discovered in the proteomes of higher eukaryotes, and their expression is often attributable to readthrough levels that are 10 to 100-fold higher than basal readthrough frequencies [9–18]. In addition, ribosome profiling and phylogenetic investigation of flies, humans, and yeast cells revealed that readthrough is more prevalent than previously anticipated, with efficiencies varying by more than 100-fold for distinct mRNAs [12,19–21]. These observations indicate that the efficiency of translation termination can be subjected to transcript-specific regulation, and that a key to understanding this regulation may lie in specific *cis*-acting mRNA sequences.

Previous studies on termination using reporter systems mainly relied on the inverse relationship between termination and readthrough efficiencies, where the level of readthrough protein product was quantified to infer the efficiency of termination. These studies have shown that the extent of readthrough, and thus the efficiency of termination, is modulated by multiple *cis*-acting elements [4,5,22]. The three stop codons themselves differ in termination efficiency, with UGA resulting in the most readthrough, and UAA the least [12,23]. Nucleotides in the immediate vicinity of the stop codon, including two nucleotides 5' of the stop codon [24,25] and up to six nucleotides 3' of the stop codon in the ribosomal mRNA channel [8,16,26–28], have been shown to modulate readthrough efficiency. Additionally, stem-loop structures in the mRNA 3'-UTR are enriched in readthrough-prone transcripts [17,20,26,29]. It remains to be determined whether the effects of these elements are applicable at a genome-wide level for endogenous mRNAs *in vivo*. Transcriptome-wide ribosome profiling studies of flies, human, and yeast cells have detected readthrough of individual mRNAs [19,21,30–32], but detailed dissection of the relationships between readthrough efficiency and *cis*-acting mRNA sequences have only been minimally investigated.

While the stop codon and its immediate 3' flanking sequences have been studied quite extensively, with structural evidence showing certain nucleotide preferences for optimal interactions with eRF1 motifs and rRNAs [22,33,34], other proximal regions have not. Novel readthrough regulatory elements are likely to exist because readthrough has been observed in genes that do not have any of the known readthrough-promoting signals (although there are genes that have readthrough-promoting elements, but do not show detectable readthrough experimentally) [10,15,20]. Based on studies of other translational control events, such as ribosome stalling [1,35], several other mRNA features could affect termination and readthrough. The nascent peptide sequences in the ribosomal exit tunnel may affect termination or readthrough efficiency via their interference with the peptidyl transferase center. The proximity of the stop codon to the poly(A)-binding protein (PABP) is also thought to influence readthrough, as PABP is known to interact with eRF3 and enhance termination both *in vitro* and *in vivo* [36–38]. These possibilities have yet to be explored with regard to normal termination at a global scale.

Accordingly, we have carried out ribosome profiling of yeast cells expressing a wild-type or temperature-sensitive mutant allele of eRF1 (Sup45 in yeast) and developed bioinformatics strategies to measure readthrough efficiency of individual mRNAs on a genome-wide scale, and to identify mRNA and peptide features which influence that efficiency. Our results demonstrated the general roles of known regulatory elements, such as the identities of the stop codon and the surrounding nucleotides, in genome-wide regulation and identified several new mRNA features that appear to play a role in translation termination and readthrough, including the penultimate codon in the P site (when a stop codon is in the A site) and the length of the 3'-UTR.

Results

Inactivation of eRF1 promotes readthrough of both normal and premature stop codons and accumulation of ribosomes at those codons

We performed ribosome profiling and total RNA sequencing of isogenic wild-type (WT) and eRF1 mutant yeast cells. The latter harbor the *sup45-2* (subsequently referred to as *sup45-ts*) mutation that minimally affects eRF1 function when cells are grown at 25˚C, but renders eRF1 inactive within 30 minutes of growth at 37˚C [39]. Reads between replicates were reproducible, with Pearson's correlation coefficients (*r*) of 0.98 and 0.99 on average for ribosome profiling and RNA-Seq, respectively (S1 Fig). In addition to the profiling data obtained from these experiments, we also analyzed published ribosome profiling data obtained from yeast cells depleted of eRF1 via specific transcription shutoff [40] (S1 Table). For comparison to ribosome occupancies in the 3'-UTR that are not due to termination defects, we included analyses of ribosome profiling data obtained from cells that are depleted of the ribosome recycling factor, Rli1 [41] (S1 Table). In these cells, full-length newly synthesized protein is properly released, but the ribosomes fail to recycle, then reinitiate randomly downstream of the stop codon, and generally produce a peptide independent of the original ORF [41]. To evaluate 3'-UTR translation, ribosome footprints mapped to genes that have no experimental UTR annotations and genes that overlap by more than 18 base pairs on the same mRNA strand (where read assignment is ambiguous) were discarded from further analyses. With these filters, we were left with 2,693 genes, or ~41% of the yeast genome, for analyses. This set of genes is referred as the reference gene set.

To explore the consequence of eRF1 inactivation on translation, ribosome footprints were mapped relative to their respective start or stop codons based on their P-site locations and quantified (Fig 1A–1D). In all strains, footprints in the coding (CDS) region showed 3-nt periodicity, a profile expected of translating ribosomes. The absence of footprints at the penultimate codon of the open reading frames (ORFs) in *SUP45* cells at 25˚C indicated that, with the stop codon in the A site, termination was completed, and ribosomes were dissociated from the

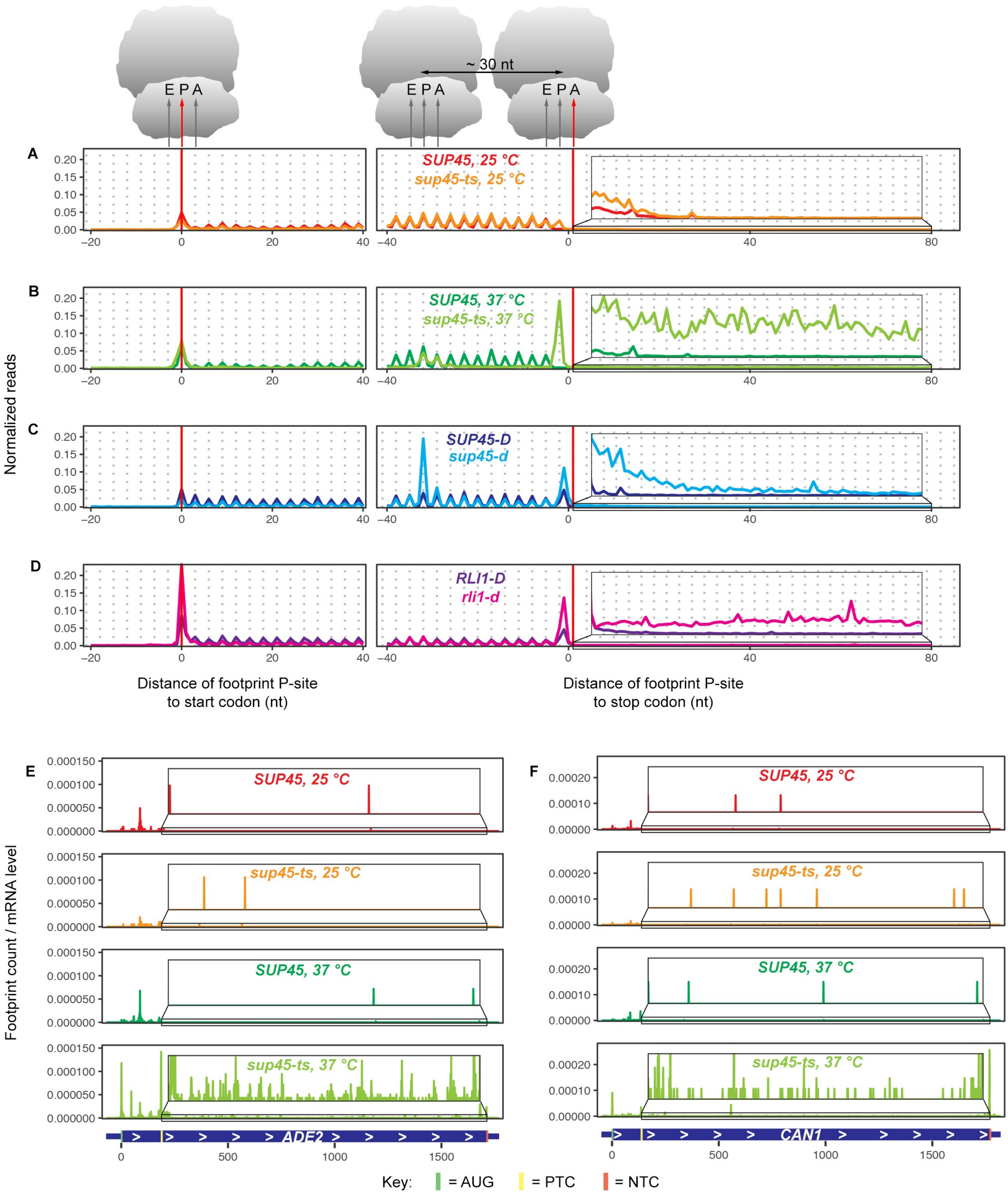
E P A
E P A
~ 30 nt
E P A
A
B
C
D
SUP45, 25 °C
sup45-ts, 25 °C
SUP45, 37 °C
sup45-ts, 37 °C
SUP45-D
sup45-d
RLI1-D
rli1-d
Normalized reads
0.00
0.05
0.10
0.15
0.20
–20
0
20
40
–40
0
40
80
Distance of footprint P-site to start codon (nt)
Distance of footprint P-site to stop codon (nt)
E
F
Footprint count / mRNA level
0.000000
0.000050
0.000100
0.000150
0.00000
0.00010
0.00020
SUP45, 25 °C
sup45-ts, 25 °C
SUP45, 37 °C
sup45-ts, 37 °C
ADE2
CAN1
0
500
1000
1500
Key: = AUG = PTC = NTC

Fig 1. Ribosome occupancy at and beyond the stop codon increases in yeast cells defective in translation termination or ribosome recycling. A-D. Ribosome footprints (2,693 genes with 3'-UTR annotations and without sequence overlapping) normalized to the total number of footprints in the indicated nt window (-20 to 40 nt around the start codon and -40 to 80 nt around the stop codon) aligned at their start or canonical stop codons. Footprints were plotted by the position of their P-sites. Each panel contains the data from WT and its respective mutant. The elevated number of footprints at the penultimate codon of the ORFs in mutant samples demonstrates ribosome stalling when the stop codon was in the A site. (The footprints at this codon and their shift to reading frame +1 in other yeast strains obtained from published data may be attributable to the differences in strain background or sequencing library preparation procedures.) Inset: Magnified view of the region in the box, showing increased 3'-UTR ribosome occupancy in *sup45-ts* cells at 37˚C, *sup45-d*, and *rli1-d* cells relative to their respective WT. **E-F** Ribosome footprints normalized to the respective mRNA levels mapped across the ORF of two PTC-containing alleles, *ade2-1* (E) and *can1-100* (F), in *SUP45* and *sup45-ts* strains at 25˚C and 37˚C.

https://doi.org/10.1371/journal.pgen.1009538.g001

mRNAs before they could be captured. The presence of footprints at this position in *sup45-ts* cells at 25˚C suggested that the mutation slowed down termination enough for the ribosomes to be captured. When eRF1 was inactivated in *sup45-ts* cells grown at 37˚C, or depleted in *sup45-d* cells, a high number of footprints was observed at the penultimate codon, indicating that ribosomes were stalling as they awaited successful termination. Similarly, the high number of footprints in *rli1-d* cells was also consistent with ribosome stalling, as the stop codon had entered the ribosomal A site, but a factor needed for successful recycling was lacking. Additionally, the peaks at approximately 30 nt upstream of the stop codon in *sup45-ts* cells at 37˚C and *sup45-d* samples identified ribosomes stacking against those stalled at the stop codon. As a consequence of failed termination or recycling, the mutant libraries, especially from the *sup45-ts* cells at 37˚C, had increased footprint reads in the 3'-UTR region compared to those of WT samples (Fig 1A–1D, insets), implying stop codon readthrough, frameshifting, or reinitiation events. Although inactivation (*sup45-ts* cells at 37˚C) and depletion (*sup45-d*) of eRF1 both caused loss of eRF1 function, these two different strategies likely resulted in loss of eRF1 function to different extent and yielded slightly different metagene profiles. While the *sup45-ts* strain lost normal eRF1 function within 30 minutes of the temperature shift, the *sup45-d* strain needed 9 hours of growth after transcription shutoff to gradually dilute the functional eRF1 protein level. In eRF1 depleted cells, although the level of functional eRF1 was significantly decreased, small amount of eRF1 still existed and was fully capable of promoting termination. The residual functional eRF1 in *sup45-d* cells likely resulted in a diminished quantity of 3'-UTR reads compared to those obtained with the *sup45-ts* strain. In addition to increased reads near the stop codons, all the mutant samples also showed increased footprint reads at the start codon and decreased footprint reads in the CDS compared to their WT counterparts. These observations are in agreement with the notion that termination and recycling steps are linked to translation initiation [3].

We also examined the effects of eRF1 inactivation on premature translation termination. Two alleles in our strain background, *ade2-1 and can1-100*, contain premature termination codons (PTCs). Their gene specific profiles revealed an accumulation of ribosomes 5' to the respective PTCs and increased ribosome density in the part of the CDS following the PTCs in *sup45-ts* cells grown at 37˚C (compared to *sup45-ts* at 25˚C or to WT cells at either temperature) (Fig 1E and 1F). This observation is further supported by computing the ratio of in-frame footprint counts downstream vs. upstream of PTCs, a measurement of readthrough efficiency that automatically takes into account the differences in mRNA levels between samples (S2A Fig). For the PTC in *ade2-1*, *sup45-ts* cells grown at 37˚C showed ~32-fold and 17-fold higher readthrough efficiencies relative to WT cells grown at 37˚C and *sup45-ts* cells grown at 25˚C, respectively. For *can1-100*, *sup45-ts* cells grown at 37˚C showed >100-fold and 26-fold higher readthrough efficiencies relative to WT cells grown at 37˚C and *sup45-ts* cells grown at 25˚C, respectively. Together, these data show that readthrough of PTCs also occurs in the absence of functional eRF1. Further evidence for PTC readthrough includes the gene-specific profiles for *RPL28* and *RPS0A* (S2B and S2C Fig). Intron-containing transcripts from both of these genes enter the cytoplasm and their translation stalls and triggers NMD (nonsense-mediated mRNA decay) when ribosomes encounter PTCs within the respective introns [42,43]. S2B and S2C Fig show that, for both transcripts, ribosome profiling reads that were generally absent from the respective intron regions in *sup45-ts* cells at 25˚C or in WT cells at either temperature were abundant in *sup45-ts* cells grown at 37˚C and accumulated mostly at in-frame stop codons (yellow rectangles).

To determine whether the increased ribosome density in regions following the stop codons in eRF1 mutants was due to stop codon readthrough rather than frameshifting or reinitiation events, we calculated the fraction of reads in each of the three reading frames in different mRNA regions: 5'-UTR, CDS, the 3'-UTR region between the canonical stop codon and the first downstream in-frame stop codon ("extension"), and the rest of the 3'-UTR ("distal 3'-UTR") (Fig 2A). If readthrough occurred, we expected i) the amount of reading frame 0 footprints, which represent in-frame translation in the CDS, to also be the dominant reading frame in the extension and ii) the amount of frame 0 footprints in the extension from *sup45-ts* cells grown at 37˚C to be higher than that of their WT counterparts. Indeed, we saw that frame 0, which was dominant in the CDS, was also dominant in the extension, but not in the distal 3'-UTR of most strains (Fig 2B). It is notable that the amount of reading frame 0 footprints in the extension region was less pronounced than that in the CDS, possibly because other recoding or recycling events, such as frameshifting or reinitiation, could also happen at the stop codon that contributed to the fraction of out-of-frame footprints. Nevertheless, compared to the reading frame fraction in the extension region of their WT counterparts, *sup45-ts* at 37˚C displayed notable enrichment of frame 0 reads while *sup45-d* and *sup45-ts* at 25˚C showed only slight increases of frame 0 reads relative to their WT counterparts (Fig 2B). On the other hand, *rli1-d* had a reading frame preference comparable to its WT counterpart (Fig 2B). These results indicate that, in cells with inactivated eRF1, a considerable proportion of footprints in the extension region of the mRNAs are caused by stop codon readthrough.

In summary, the increased in-frame footprints in the extension region of the 3'-UTR in eRF1 mutants indicated that a lack of functional eRF1 led to stop codon readthrough globally. Hence, we developed formal criteria for the quantitation of readthrough of individual mRNAs and a bioinformatic scheme to assess the genome-wide positive and negative *cis*-acting effects of neighboring sequences to the readthrough of stop codons.

Analysis pipeline development

To dissect the relationships between readthrough efficiency (Y variable) and mRNA or nascent peptide features (X variables), we used the random forest machine learning approach [44,45] to first narrow down the candidate features. We defined each of the variables as follows:

Defining readthrough efficiency (Y variable). Since termination efficiency is not directly measurable by ribosome profiling we calculated readthrough efficiency and utilized the generally inverse relationship of termination and readthrough (shown experimentally by Wu *et al.* (2020) [38]) to infer termination efficiency. Our subsequent interpretation of the results was based on this relationship, where high readthrough efficiency would point to low termination efficiency and vice versa in this experimental context. Readthrough efficiency in the context of ribosome profiling can be loosely defined as the ratio of ribosome density (footprint count normalized to the length of the region) in the 3'-UTR to ribosome density in the CDS; however, the nature of the mutants should be taken into account to ensure accuracy and minimize noise. First, due to ribosome stalling and queuing at the start and stop codons in mutant strains (Fig 1A), ribosome density in the CDS may be overestimated; therefore, we excluded footprints whose P-sites fall into the first 15 and the last 33 nucleotides of the CDS of each

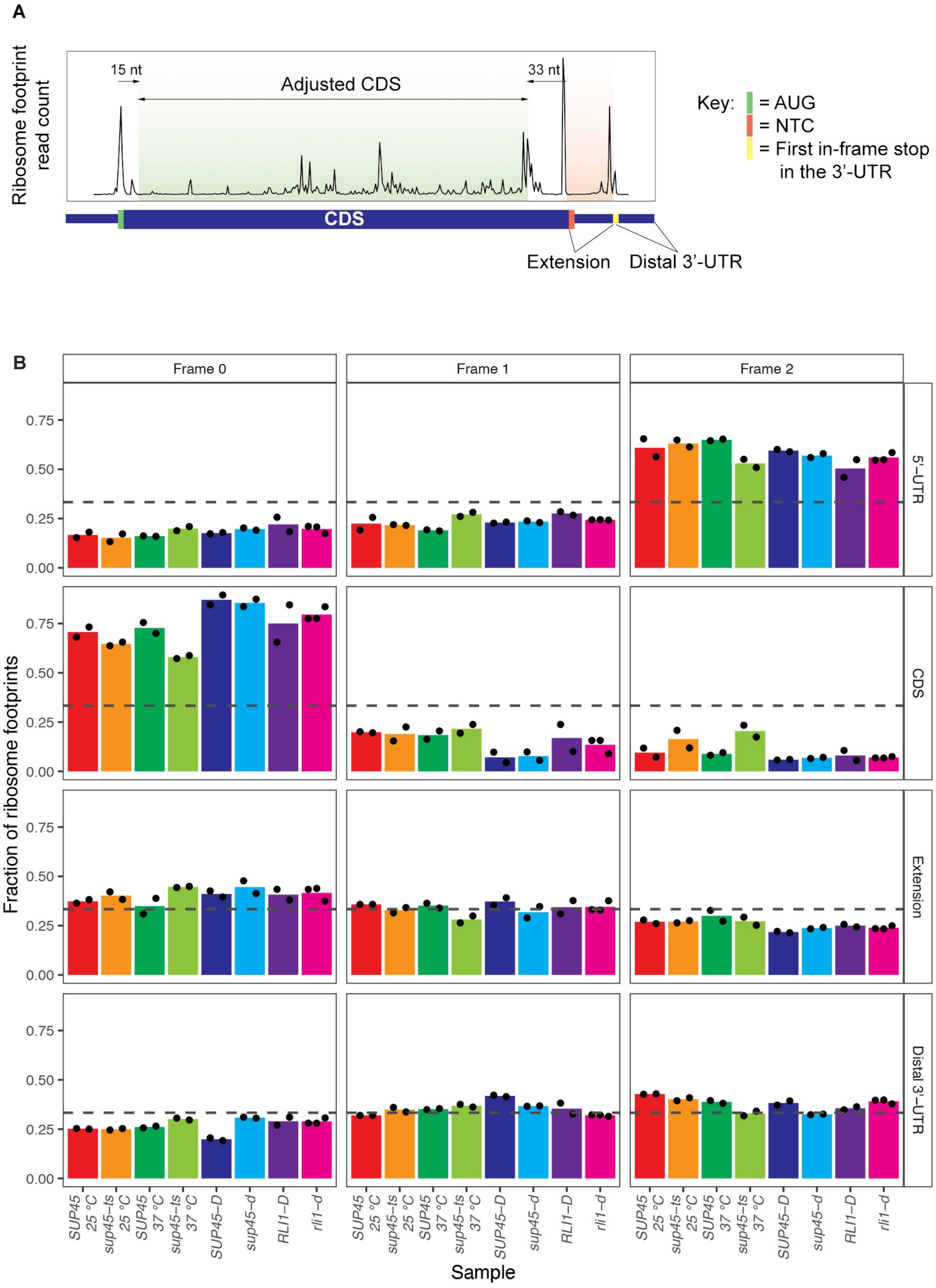

Fig 2. Readthrough events were characterized by maintenance of reading frame when ribosomes bypass the stop codon. A. Example of ribosome footprint count plot with labeled mRNA regions of interest. **B.** Fractions of reads in each of the three reading frames for different mRNA regions. The results of each sample were the average of 2 or 3 replicates, with black dots indicating individual data points. The dashed line drawn at 0.33 is the theoretical fraction at which all 3 reading frames are represented in equal amounts.

https://doi.org/10.1371/journal.pgen.1009538.g002

gene in all samples. Although the stalling of the ribosomes also occurred at the first downstream stop codon in the 3'-UTR, we could not use the same rules of footprint exclusion here because the lengths of many extension regions are smaller than 33 nucleotides. Thus, the apparent readthrough level could be somewhat overestimated. Second, as readthrough is by definition in-frame with the CDS (Fig 2B), only footprints that were in frame 0 were considered. The adjusted readthrough efficiency calculation, log-transformed, is shown in Fig 3A.

Additional sources of noise in the data include: 1) poorly-expressed genes that result in an exaggeration of the readthrough efficiency ratios and 2) genes lacking footprints in the 3'-UTR region, for which it was challenging to distinguish lack of readthrough from insufficient sequencing depth. Inclusion of these genes confounded previous analyses and led to conflicting results. For instance, genes belonging in the "leaky" set (genes that had 3'-UTR reads, i.e. readthrough) were highly expressed compared to genes in the "non-leaky set" (genes lacking 3'-UTR reads), but within the "leaky" set, readthrough rate was negatively correlated with gene expression [21]. Therefore, we discarded genes with both RPKM of CDS < 5 and RPKM of 3'-UTR < 0.5 [41] from further analyses.

Defining mRNA and nascent peptide features (X variables). To date, mRNA features affecting stop codon readthrough have frequently been identified 3' of the stop codon, but we also considered the possibility that mRNA sequence on the 5' side of the stop codon and the nascent peptide in the exit tunnel may contribute to variations in readthrough efficiencies (Fig 3B). These features can either be directly documented or inferred from the mRNA sequences (Table 1). Therefore, we used established algorithms to predict RNA secondary structure [46] and nascent peptide α-helix formation [47].

Other than the mRNA features listed in Table 1, many characteristics of the mRNAs (such as expression levels, codon optimality, translation efficiency, and gene length) were previously correlated with readthrough efficiency measured by ribosome profiling [21,31]. We performed similar analyses involving pairwise correlation between readthrough efficiency, gene expression level (determined by RSEM of RNA-Seq data), translation efficiency (ribosome density divided by gene expression level), length of transcript and mRNA regions, and codon optimality (tRNA adaptation index [48,49]) (S3 Fig). We found that, except for *sup45-ts* at 37˚C, readthrough efficiency was negatively correlated with gene expression and codon optimality, consistent with previous results [21,31]. This correlation may have an evolutionary, rather than a mechanistic, explanation; for example, the inverse relationship between gene expression and readthrough efficiency may result from more deleterious effects of readthrough from highly expressed genes than from poorly expressed genes, and thus highly expressed genes evolved to have lower readthrough efficiency [31]. Since gene expression level was positively correlated with codon optimality across the ORF, where mRNAs with optimal codons are more stable and hence more abundant than those with non-optimal codons [50], we could not make a distinction whether the observed negative correlation between readthrough efficiency and codon optimality was simply due to increased mRNA abundance or was instead mechanistically relevant to the readthrough process. We therefore did not include gene expression and codon optimality in further analyses. Moreover, readthrough measurements in general should be interpreted with caution and due consideration given the caveats described above.

Random forest analyses reveal stop codon identity, P-site amino acid identity, and 3'-UTR length as important factors in predicting readthrough efficiency

In order to identify without bias the mRNA and nascent peptide features affecting readthrough efficiency, we used two independent approaches involving the random forest algorithm

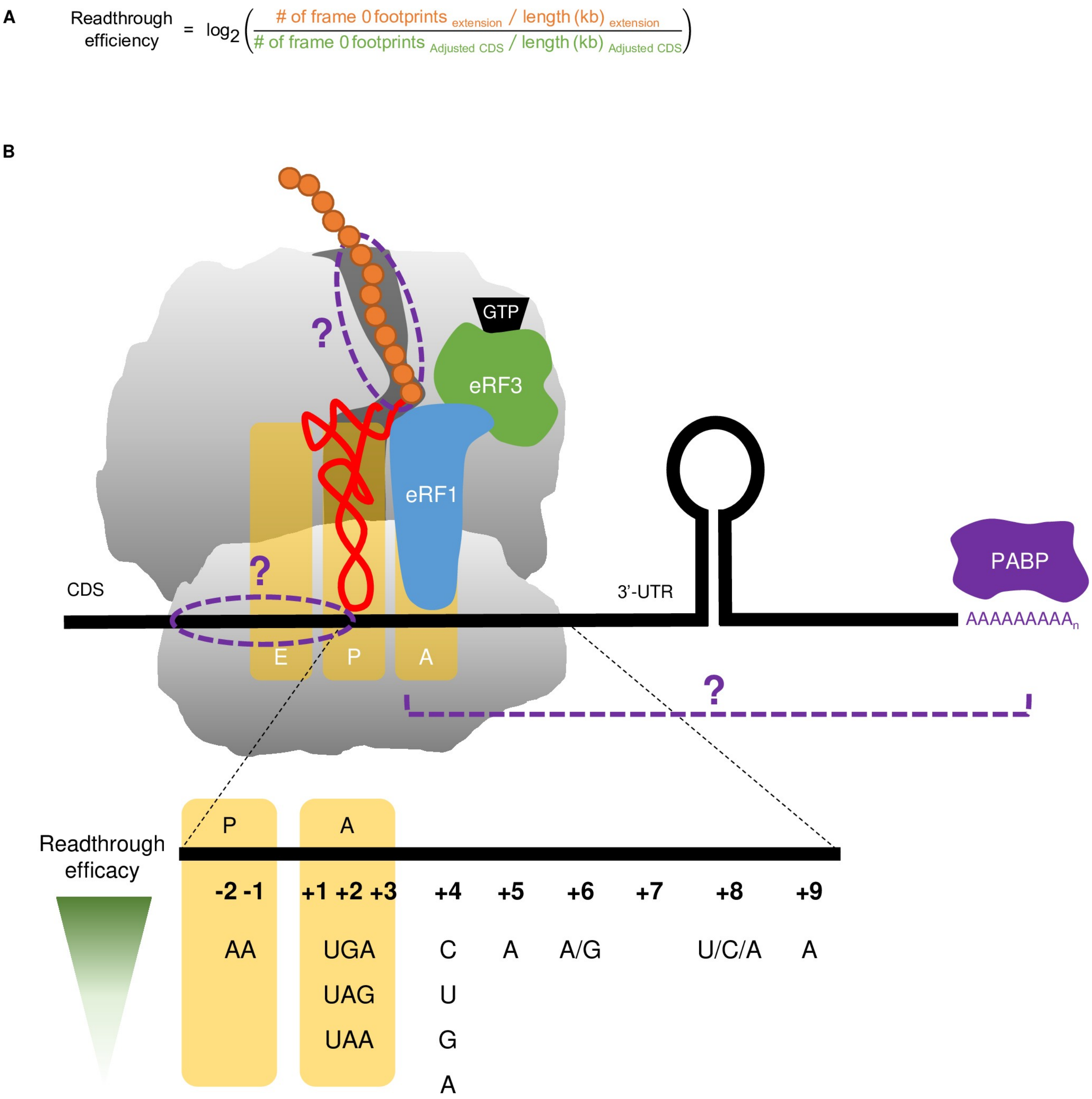
A
Readthrough efficiency = log2 (# of frame 0 footprints extension / length (kb) extension / # of frame 0 footprints Adjusted CDS / length (kb) Adjusted CDS)
B
GTP
eRF3
?
eRF1
?
CDS
E
P
A
3'-UTR
PABP
AAAAAAAAAn
?
Readthrough efficacy
P
A
-2 -1
+1 +2 +3
+4
+5
+6
+7
+8
+9
AA
UGA
UAG
UAA
C
U
G
A
A
A/G
U/C/A
A

Fig 3. Definition of X and Y Variables. A. Readthrough efficiency calculation formula, which takes into account the ribosome pile-up at NTCs and reading frame preference. **B.** mRNA and nascent peptide features of interest in the context of the ribosome and related translation termination components. Previously identified variables are colored and labeled. Hypothesized variables are in purple and marked with "?". Details of how each variable is measured are in Table 1. Information regarding the effects of particular nucleotides on readthrough efficacy is taken from a recent review, as is the general format of the schematic [5].

https://doi.org/10.1371/journal.pgen.1009538.g003

[44,45]: 1) random forest regression to predict readthrough efficiency based on the X variables, and 2) random forest classification to differentiate between "High" (top 15%) and "Low" (bottom 15%) readthrough genes based on the X variables. The number of genes in each type of analysis can be found in S2 Table. A regression model and a classification model with 5-fold

Table 1. Assessed X variables and their modes of measurement.

Variable		Measurement
Nucleotides in the mRNA channel	Stop codon (nt +1 to +3) in the A site	UAA, UAG, UGA
	nt -1 to -15 and nt +4 to +9	A, C, G, U
Nascent peptide in the ribosome exit tunnel	P-site and E-site amino acids (aa 1–2 from PTC)	A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
	Amino acids in the upper tunnel (aa 3–9 from PTC)	Charge: fraction of +, 0,—residues Polarity: fraction of polar, nonpolar residues Aromaticity: fraction of aromatic residues Hydrophobicity: fraction of hydrophilic, neutral, hydrophobic, very hydrophobic residues Tendency to form α-helix (only for aa 20–30): fraction of residues involved in alpha helix formation, predicted by Jpred [47]
	Amino acids at the constriction site (aa 10–12 from PTC)	
	Amino acids in the central tunnel (aa 13–19 from PTC)	
	Amino acids in the lower tunnel (aa 20–30 from PTC)	
3'-UTR secondary structure	Tendency for 3'-UTR sequence to form a secondary structure	The lowest minimum free energy (MFE) for a structure in 150 nt scanning window of entire 3'-UTR sequence, predicted by RNALFold function in ViennaRNA package [46]
	Distance of structure from stop codon	Number of nt from stop codon to the base of the structure
Proximity of PABP to stop codon	Distance of PABP (i.e., poly(A)-tail) to stop codon	3'-UTR length (nt)
Negative controls	Randomly assigned factor (categorical)	A, C, G, U
	Randomly assigned number (numeric)	1–100

https://doi.org/10.1371/journal.pgen.1009538.t001

cross-validation were created for each sample using the same tuning parameters. For the regression model, we used Normalized Root Mean Squared Error (NRMSE) as a measurement of predictive ability of X variables. We observed NRMSE of approximately 0.14 across all samples, indicating that the average difference between predicted and actual readthrough efficiency is 14% (Fig 4A). For the classification model, Area Under the Receiver Operating Characteristics (AUROC) was used as a measurement of model capability at distinguishing between the "High" and "Low" groups. AUROC values ranging from 0.72–0.82, 0.76 on average, were observed across all samples; therefore, the models had ~76% chance of classifying genes into the correct group (Fig 4B).

To determine which of the X variables were most responsible for the predictions, we permuted each X variable for the regression and classification models and assessed its importance by calculating either percent increase in mean squared error (%IncMSE) or Mean Decrease Accuracy (MDA). %IncMSE is the calculation of percent increase in prediction error, while MDA is the percent increase in misclassification, when the X variable was permuted. Thus, the higher the %IncMSE or MDA value is for an X variable, the more critical that variable is in predicting readthrough efficiency. We performed hierarchical clustering of these values across samples and found that in general, the eRF1 mutants were more similar to each other than to other strains in both regression and classification approaches (Fig 4C and 4D).

In order to determine which X variables exhibited meaningful %IncMSE or MDA values, we established their baselines (negative controls) by randomly assigning one numeric value between 1–100 (random number) and one of four categorical values (random factor) to each gene before model training. Because the assignment was random, these two variables should have no influence on the prediction. Indeed, random number and random factor were among the X variables with %IncMSE and MDA values close to zero (Fig 4C and 4D). Reassuringly, in both the regression and classification approaches, the variables previously known to affect readthrough efficiency, such as the identity of a stop codon, nucleotide (nt) +4 and +5 were

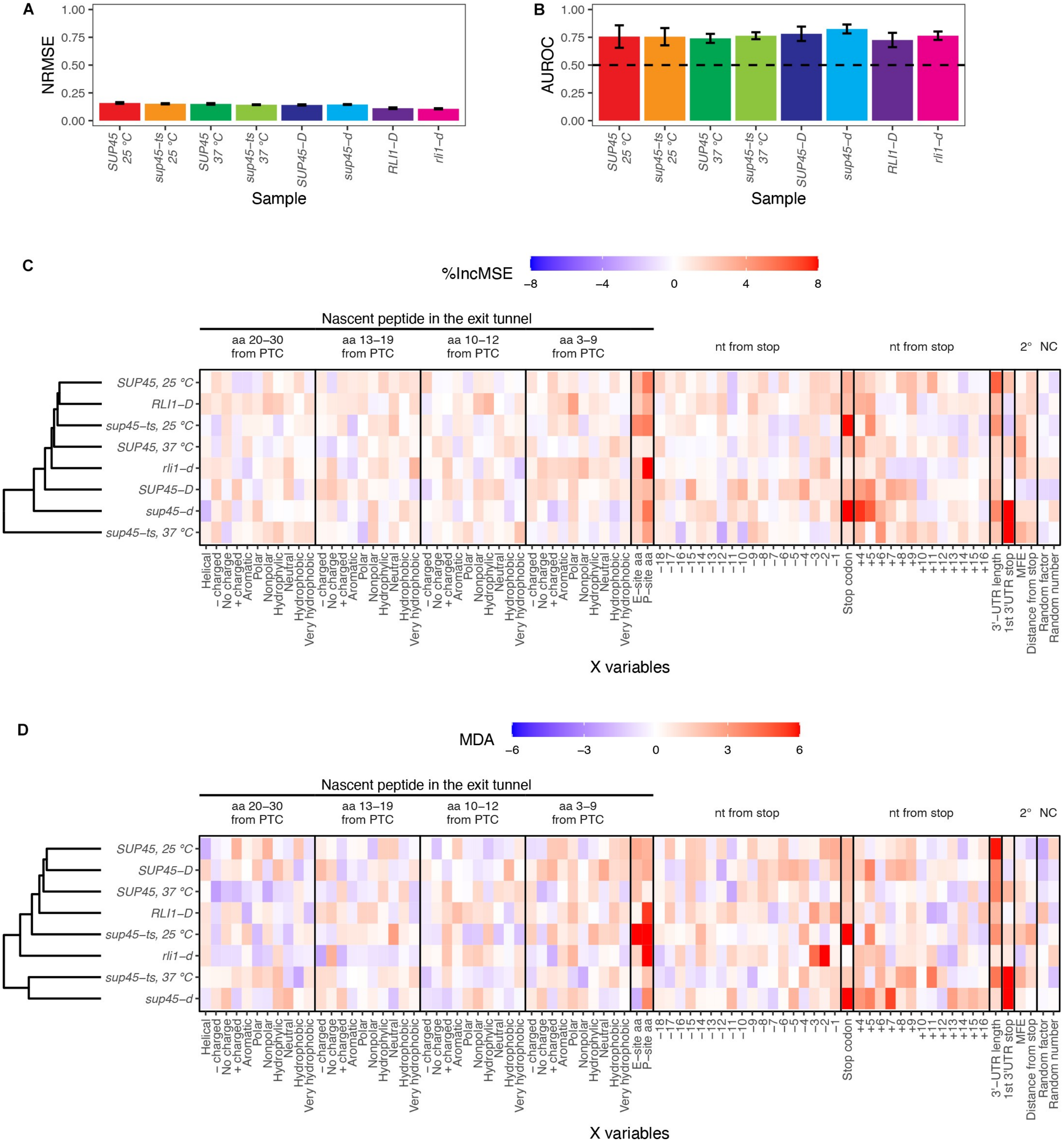

Fig 4. Random forest identified stop codon identity, P-site amino acid identity, and 3'-UTR length as factors critical in readthrough efficiency prediction. A. Performance metric for random forest regression displaying average Normalized Root Mean Squared Error (NRMSE) ± standard deviation across 5-fold cross-validation. **B.** Performance metric for random forest classification displaying average Area Under the Receiving Operator Curve (AUROC) ± standard deviation across 5-fold cross-validation. The value of 0.5 (dashed line) means the classification by the input X variables is no better than random chance. **C.** and **D.** Feature importance extracted from the random forest models. The relative importance of a feature is represented by percent increase in mean squared error (%IncMSE) for regression (C) and Mean Decrease Accuracy (MDA) for classification (D), which is percent increase in error that results from permuting the feature. The higher the %IncMSE or MDA is, the more crucial the feature is in predicting readthrough efficiency or correctly classifying genes into groups. Arbitrary continuous and discrete values randomly assigned to the genes (random number and random factor) are used as negative controls.

https://doi.org/10.1371/journal.pgen.1009538.g004

among the variables with %IncMSE and MDA values above the baseline in most samples, arguing that our pipeline was capable of distinguishing relevant mRNA features. In addition to the stop codon identities, 3'-UTR length and the P-site amino acid (or tRNA or codon) also had one of the highest %IncMSE and MDA values in most samples. The nt -3, and -2 in the P-site only had high MDA values in the ribosome recycling mutant, *rli1-d*, suggesting that they may play a role in recycling; however, this is only true in the classification model.

Interestingly, the first downstream stop codon in the 3'-UTR ("1st 3'-UTR stop" in Fig 4C and 4D) had high %IncMSE and MDA values only in *sup45-ts* at 37˚C and *sup45-d* samples. This observation is due to ribosome pile-up at those stop codons, as with canonical stop codons (Fig 1B and 1C). The identity of this stop codon affects the read count in the extension region–its high %IncMSE and MDA value reinforces the role of stop codon identity in readthrough but not mechanistic relevance to readthrough occurring specifically at the canonical stop codon.

In summary, random forest analyses reveal identities of the stop codon, P-site amino acid (or tRNA or nucleotides), and 3'-UTR length as mRNA features that control genome-wide readthrough efficiency.

Readthrough-promoting stop codons and nucleotides at positions +4 to +9 occur at a genome-wide level

Previous studies using synthetic dual reporters have shown that the UGA stop codon and certain nucleotides at positions -2, -1, +4 to +9 are the most permissive features for stop codon readthrough (Fig 3B) [5,8,16,22–28]. Since the random forest analyses also identified the stop codon identity and some of these nucleotides as features that, when permuted, led to higher prediction error than baseline, we wondered whether these same stop codon and nucleotide identities were also readthrough-permissive or readthrough-inhibiting at a genome-wide level. To answer this question, the genes were divided into 3 groups for analysis of stop codon identity or 4 groups for analyses of individual nucleotides at each position and Wilcoxon's rank sum test was used to compare the median of genes in each group to the overall sample median. The differences between group median and sample median were shown as a heatmap (Fig 5A). The significance values from Wilcoxon's rank sum test were represented by tile size, where a larger tile depicted p-value < 0.05. As expected, the distribution of random factor variables was not significantly different between the groups and sample median and showed no particular pattern of higher or lower readthrough efficiency across all samples.

The variables that showed the strongest pattern and significance were the stop codon and the nucleotide immediately downstream (nt +4). Consistent with the notion that UGA is the most readthrough-permissive stop codon and UAA is the strongest terminator [23], genes with UAA as stop codons had significantly lower readthrough efficiencies while genes with UGA had significantly higher readthrough efficiencies compared to the sample median in all samples except for *rli1-d* cells. For nt +4, genes with C at this position had significantly higher readthrough efficiencies, and those with G had slightly lower readthrough efficiencies. These observations are in line with previous data where C was reported to allow the most readthrough and G the least [5,23,26,27] and are supported by structural evidence showing preferential base stacking of purines at nt +4 with 18S rRNA, enhancing the stop codon recognition by eRF1 [33].

For nucleotides further into the mRNA channel 3' of the stop codon (nt +5 to +9), only nt +9 showed clear trends: genes with A had higher readthrough efficiencies (i.e., readthrough-permissive), consistent with previous reports, and genes with U had lower readthrough efficiencies (i.e., readthrough-inhibiting). At position +5, although the differences between

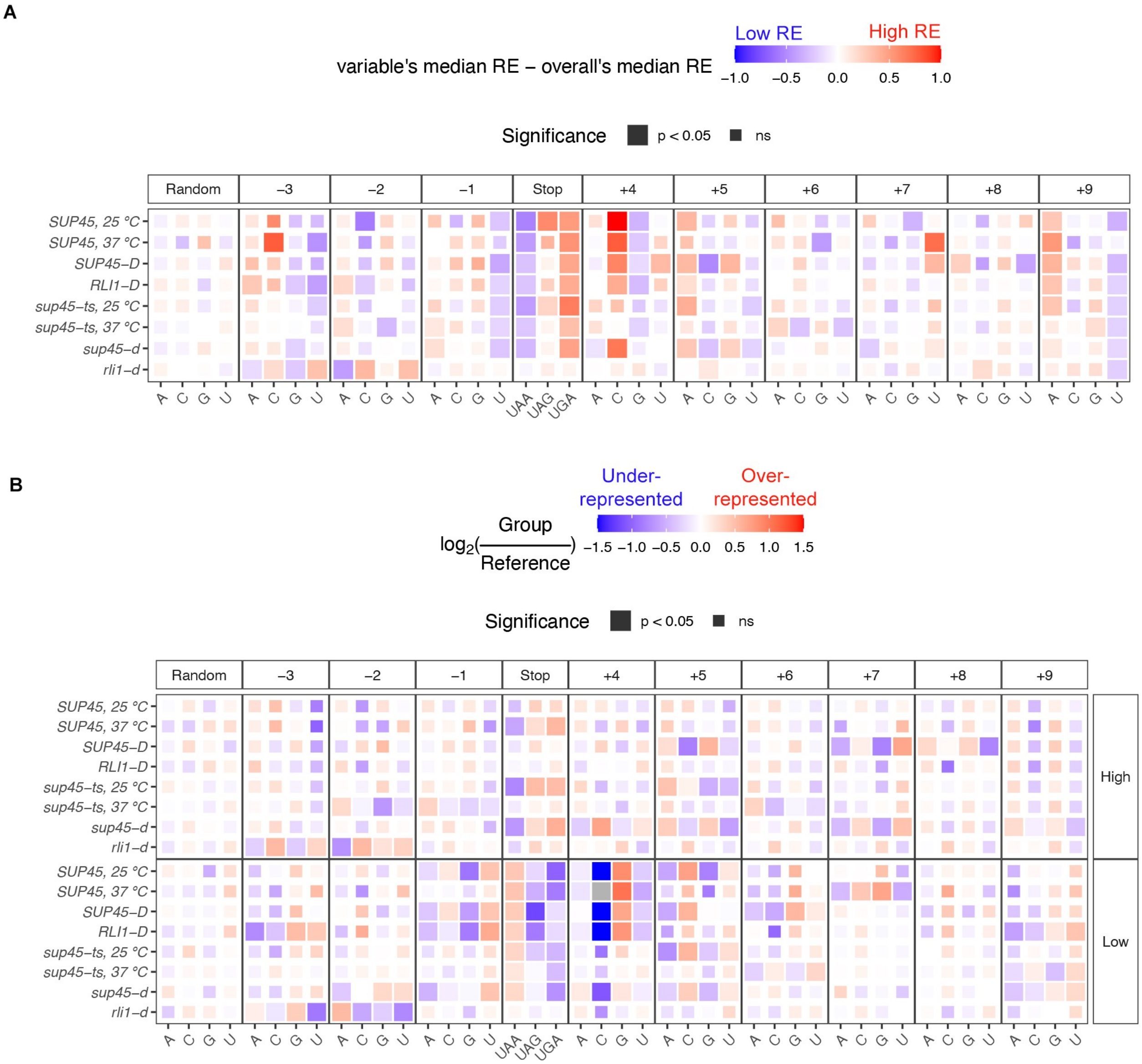

Fig 5. Known individual effects of stop codon and surrounding nucleotide identities on readthrough efficiency occur at a genome-wide level. **A.** Heatmap of median readthrough efficiency of genes containing particular stop codon or nucleotide relative to median readthrough efficiency of all genes in the sample. Positive value (red) indicates that the group of genes had higher readthrough efficiencies compared to the sample median, while negative value (blue) indicates lower readthrough efficiencies. Wilcoxon's rank sum test was used to determine whether the difference in group and sample media was significant. Significant difference was represented as a larger tile. **B.** Heatmap demonstrating over-representation (red) or under-representation (blue) of a stop codon or nucleotide in "High" or "Low" readthrough groups relative to the frequency in the reference gene set. χ^2 goodness of fit test with Bonferroni correction was performed to compare usage distribution between the "High" readthrough group, "Low" readthrough group, and the reference. Significant difference was represented as a larger tile. Grey tiles signify that there was zero observation of that nucleotide (essentially under-represented), and $\log_2$ calculation and statistical analysis could not be performed.

https://doi.org/10.1371/journal.pgen.1009538.g005

groups' readthrough efficiencies and overall readthrough efficiencies were not significant in every sample, genes containing adenine had a uniform pattern of higher readthrough efficiency compared to sample median (unlike random factor), indicating that A at nt +5 was relatively readthrough-permissive.

Adenine at nt -1 and -2 has been shown to be readthrough-permissive [24,25], but this was not reflected in our current analysis (Fig 5A). Instead, we observed that U at nt -1 was readthrough-inhibiting, and A and C but not G and U at nt -3 were readthrough-permissive. It is noteworthy that nt -3 was identified as having a significant role along with nt -2; this may mean that the amino acid they encode or its tRNA, rather than the nucleotides themselves, mechanistically affect the termination process. Thus, we also explored these three nucleotides as a codon (see below).

It is notable that most of the patterns for nt -3 and -2 in *rli1-d* sample did not match with other samples. In fact, some of them appeared to be the opposite. Since other samples, even the *sup45* mutants, had wild-type Rli1, the opposite patterns in *rli1-d* sample may suggest that nt -3 and -2 (or the amino acid they encode) played a role in a ribosome recycling step rather than the termination step. This also influenced the number of ribosomes translated into the 3'-UTR in *rli1-d*, as shown in the feature importance plot (Fig 4D).

To see whether we could extract more comprehensive information on these nucleotides, we used a different approach involving only the "High" and "Low" readthrough genes used for the classification models. χ^2 goodness of fit tests with Bonferroni correction were used to compare the frequencies of UGA, UAG, UGA, A, C, G, and U at each nucleotide position between the "High" readthrough group, "Low" readthrough group, and the reference gene set (2,693 genes). For data presentation purposes, the frequencies were converted into fractions and $\log_2$ ratios of "High" or "Low" groups to the reference were computed, as shown in the heatmap (Fig 5B). A positive $\log_2$ ratio (red) indicates an over-representation while a negative $\log_2$ ratio (blue) indicates an under-representation of X in the group compared to reference. The patterns, especially those of the "Low" readthrough group, corroborated our previous approach (Fig 5A). For instance, genes with C at the nt +4 position had higher readthrough efficiencies than the sample median (Fig 5A), and within the "Low" readthrough group (Fig 5B), C was severely under-represented (blue tile) or even absent (grey tile). Moreover, we were able to rank the nucleotides from readthrough-permissive to readthrough-inhibiting using this approach. Although the patterns of either groups were not significantly different from reference based on the χ^2 analysis, uniform patterns could be observed across all samples for most nucleotide positions (compared to random factor). For nt +5, the pattern allowed us to extrapolate that following readthrough-permissive G, were increasingly readthrough-inhibiting A, U, and C (which was most associated with termination). This information was difficult to obtain in the dual reporter assays because readthrough events were so low that they were indistinguishable from each other. Information compiled for other nucleotide positions is shown in Table 2.

Next, we asked how particular combinations of stop codon and nt +4 identities, the two strongest variables among all the nucleotides, affect readthrough efficiency together. Genes

Table 2. Comparison of nt -3 to +9 effects (based on Figs 5 and 6) to previous data [5]. The order of nucleotide/stop codon is from most to (">") least permissive to readthrough.

Position	Literature [5]	This study
-3	-	A/C > G > U
-2	A	A > G > U > C
-1	A	A > G > C > U
Stop codon (+1, +2, +3)	UGA > UAG > UAA	UGA > UAG > UAA
+4	C > U > G > A	C > U > A > G
+5	A	A > G > U > C
+6	A/G	A/C > G/U
+7	N	U
+8	U/C/G	G > A/U > C
+9	A	A > C/G > U

https://doi.org/10.1371/journal.pgen.1009538.t002

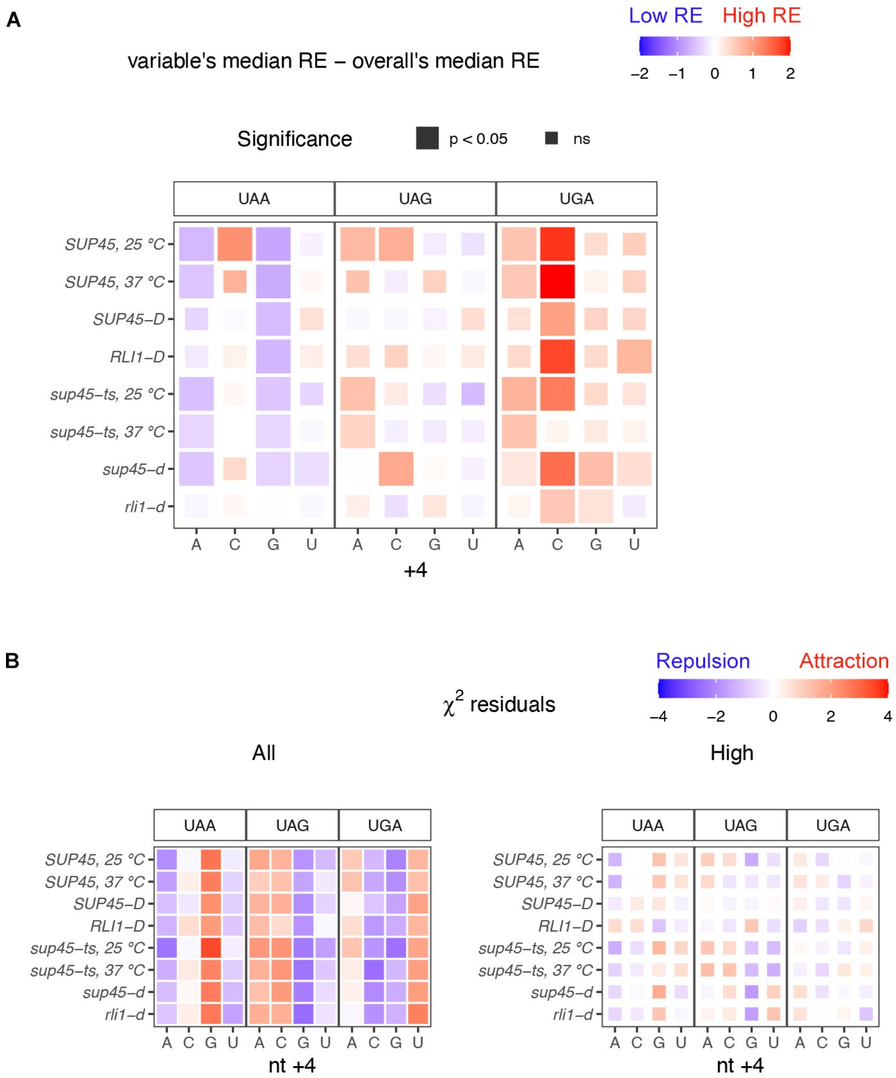

Fig 6. Readthrough-permissive combinations of stop codon and nt +4 identities have additive effects on readthrough efficiency and are avoided in the genome. A. Heatmap of median readthrough efficiency of genes containing particular combination of stop codon and nt +4 relative to median readthrough efficiency of all genes in the sample. Statistical analysis as in A. **B.** Heatmap showing residuals of χ^2 test of independence between stop codon and nt +4 identities. Significant values ($p < 0.05$), represented by larger tile size, indicate the significant association between stop codon and nt +4 identities. Positive (red) residuals specify positive association (attraction) and negative (blue) residuals negative association (repulsion) between variables.

https://doi.org/10.1371/journal.pgen.1009538.g006

were divided into groups based on their stop codon and nt +4 identities and the same analysis as described in Fig 5A was performed. We found that genes containing UGA, the most readthrough-permissive stop codon, followed by C, the most readthrough-permissive nt +4, had the highest readthrough efficiency relative to the overall median. UGA followed by G, the least readthrough-permissive nt +4, reduced readthrough efficiency but not to the level of overall median (Fig 6A). On the other hand, genes containing the most readthrough-inhibiting combination, UAAG, had the lowest readthrough efficiencies while UAAC brought readthrough up to a level that was equal to that of the overall median. With regard to UGAC association with the highest readthrough efficiency, we noticed that this trend was not true for the *sup45-ts* strain at 37˚C where, unlike in other strains, genes containing UGAC did not have higher readthrough efficiency than the median (Fig 6A). This result suggests that in the absence of functional eRF1, some combinations of stop codon and nt +4 identities had limited effects on readthrough efficiency variation.

To test whether particular stop codon and nt +4 identities tend to occur together at different readthrough rates, a χ^2 test of independence was performed using all genes ("All"), genes in the "High" readthrough group, or genes in the "Low" readthrough group (Fig 6B). We found that in general ("All"), the association between stop codon and nt +4 identities was significant in all samples. Readthrough-inhibiting features, UAA and G, appeared together more frequently than the expected frequency (positive residuals) while readthrough-promoting features, UGA and C, tended to repel each other (negative residuals) (Fig 6B, left panel). Similar but weaker associations were seen among genes with "Low" readthrough, but no association was observed among those with "High" readthrough. From an evolutionary standpoint, this analysis supported a previous proposal stating that genes have evolved to minimize stop codon readthrough [31], as low readthrough features tended to coexist but high readthrough features tended to exclude each other.

To summarize, we showed that readthrough-promoting stop codon and nucleotides at positions +4 to +9 previously determined in reporter assays also occur at a genome-wide level, with some combinations of stop codon and nt +4 identities selected for or against each other. Additionally, we identified readthrough-inhibiting nucleotides refractory to previous studies due to the detection limits of reporter assays.

Specific combinations of nucleotides and tRNA in the ribosomal P site may influence readthrough efficiency

Random forest identified the P-site amino acid to be a key factor in readthrough efficiency prediction, and we observed patterns of nucleotide usages at position -3, -2, and -1; however, it was unclear whether readthrough was influenced by the nucleotide themselves, the amino acid they encode, the tRNA that decoded them, or a combination of these three possibilities. Therefore, we performed analysis as described in Fig 5A, but as codons instead of individual nucleotide positions (Fig 7A) and were able to identify codons associated with genes having significant increases or decreases in readthrough efficiencies compared to the sample median.

From this analysis, two particular nucleotide combinations were likely responsible for low readthrough efficiencies, namely the C/G and U combinations and the UU and A/C combinations. In the first group, CGU and GCU were codons that appeared in genes with significantly lower readthrough efficiencies in most samples even though they code for different amino acids and were clearly distinct from their respective synonymous codons. Remarkably, genes containing CUG (leucine), which had exactly the same nucleotide composition, showed higher readthrough efficiencies compared to sample median. These opposite trends suggested that not only specific combinations of nucleotides, but also specific positional arrangements have to be optimal for efficient termination. In the second group genes containing UUA and UUC had lower readthrough efficiencies but only in WT conditions (WT strains and *sup45-ts* at 25˚C). Similar to the C/GU case, UUA and UUC had the same pattern despite coding for different amino acids and were distinct from their respective synonymous codons.

In both cases, the amino acid identity was ruled out as a possible mechanism. We wondered whether these two groups of codons share certain properties that may explain their occurrence in low readthrough genes and explored P-site codon usage frequency in the reference set, tRNA adaptation index (tAI) [48,49], tRNA anti-codon modifications [51], and wobble pairs (Fig 7B). None of these properties were shared exclusively among the four codons, suggesting the mechanisms by which UUA/UUC and CGU/GCU influence readthrough efficiency were different. Nevertheless, the role of tRNA properties could not be ruled out. It is likely that the

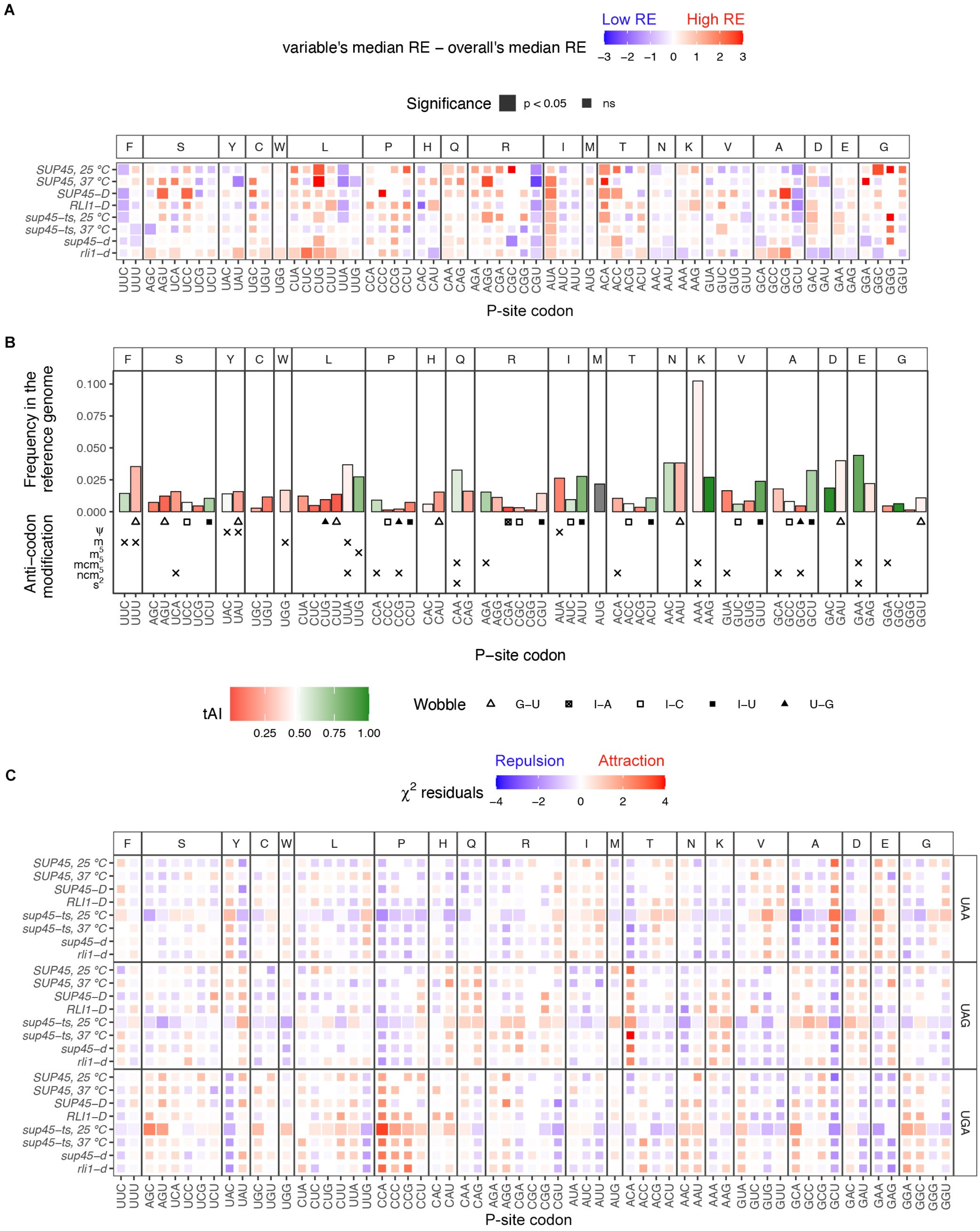
A
Low RE
High RE
variable's median RE – overall's median RE
−3 −2 −1 0 1 2 3
Significance
p < 0.05
ns
F S Y C W L P H Q R I M T N K V A D E G
SUP45, 25 °C
SUP45, 37 °C
SUP45–D
RLI1–D
sup45–ts, 25 °C
sup45–ts, 37 °C
sup45–d
rli1–d
UUC UUU AGC AGU UCA UCC UCG UCU UAC UAU UGC UGU UGG CUA CUC CUG CUU UUA UUG CCA CCC CCG CCU CAC CAU CAA CAG AGA AGG CGA CGC CGG CGU AUA AUC AUU AUG ACA ACC ACG ACU AAC AAU AAA AAG GUA GUC GUG GUU GCA GCC GCG GCU GAC GAU GAA GAG GGA GGC GGG GGU
P-site codon
B
Frequency in the reference genome
0.100
0.075
0.050
0.025
0.000
Anti–codon modification
ψ
m
m5
mcm5
ncm5
s2
P–site codon
tAI
0.25 0.50 0.75 1.00
Wobble
G–U
I–A
I–C
I–U
U–G
C
Repulsion
Attraction
χ2 residuals
−4 −2 0 2 4
UAA
UAG
UGA
P-site codon

Fig 7. Specific codons in the P site were associated with higher or lower readthrough efficiencies. A. Heatmap of median readthrough efficiency of genes containing a particular triplet codon in the ribosomal P site. Positive value (red) indicates that the group of genes had higher readthrough efficiencies compared to sample median, while negative value (blue) indicates lower readthrough efficiencies. Wilcoxon's rank sum test was used to determine whether the difference in group and sample media was significant. Significant difference was represented as a larger tile. **B.** General information regarding the P-site codons. The Y-axis shows frequency of codon usage in the P site among the reference gene set. Color of the bar represents tRNA adaptation index (tAI), a measurement for codon optimality where 0 is non-optimal and 1 is optimal [48,49]. Below the bar plots are wobble pair information and anti-codon modifications. **C.** Heatmap showing residuals of χ^2 test of independence between stop codon and P-site codon identities. Significant values ($p < 0.05$), represented by larger tile size, indicate the significant association between stop codon and P-site codon identities. Positive (red) residuals specify positive association (attraction) and negative (blue) residuals negative association (repulsion) between variables. Interpretation of the results should be done with caution as the analysis suffered from small sample size.

https://doi.org/10.1371/journal.pgen.1009538.g007

nucleotides and tRNAs together interact with the translation termination components in such a way that is optimal for a termination event to occur.

At the other end of the spectrum, codons that appeared in genes with significantly higher readthrough efficiencies include AUA, ACA, ACC, CUG, and GAC. Arguments similar to those made for the low readthrough set of codons applied here. First, the amino acid identity could be ruled out. Although ACA and ACC encode the same amino acid (threonine), it could be ruled out as a possible mechanism because other threonine codons showed different patterns. All five codons did not share any tRNA properties in common; however, individualized nucleotide-tRNA property combinations could not be ruled out as possible mechanisms.

Next, we performed a χ^2 test of independence to determine whether the P-site codons with prominent effects on readthrough efficiencies tended to appear in conjunction with particular stop codons. Although statistical analysis here suffered from small sample size and the results should be interpreted with caution, we could still observe a similar trend as seen with stop codon and nt +4 in that readthrough-inhibiting features attracted one another. We found the most readthrough-inhibiting P-site codons, CGU and GCU, coexisted more frequently with UAA than UAG or UGA stop codons (Fig 7C). On the other hand, ACA, which was found in genes having high readthrough efficiencies, coexisted more frequently with UAG than UAA.

As with nt -3 and -2, many P-site codons (e.g., GAC, CUA, CUC, UUA) in the *rli1-d* mutant had either the opposite effect on readthrough efficiencies or opposite level of significance compared to the eRF1 mutants and WT cells. Thus, it is possible that these codons affect the recycling step rather than termination step.

In summary, we identified four codons (CGU, GCU, UUA, and UUC) associated with lower readthrough efficiencies and five codons (AUA, ACA, ACC, CUG, and GAC) associated with higher readthrough efficiencies. It is unlikely that amino acid identity by itself is the reason for this observation; rather, specific combinations of nucleotide and tRNA properties likely influence termination and readthrough.

Readthrough efficiency increased with 3'-UTR lengths in the eRF1 mutants, but not in wild-type or recycling factor mutant cells

Poly(A)-binding protein (PABP, Pab1 in yeast) has been shown previously to interact with eRF3, enhancing termination efficiency *in vitro* [36]. Consistent with these observations, deletion of the *PAB1* gene in yeast, or extension of the mRNA 3'-UTR length downstream of a PTC, leads to inefficient termination and readthrough *in vivo* [38]. Accordingly, we asked whether readthrough efficiency and 3'-UTR length are positively correlated. A weak but significant positive correlation between readthrough efficiency and 3'-UTR length was observed in *sup45-ts* cells at 37°C and *sup45-d* cells while a negative correlation was observed in the respective WT strains (Fig 8). These results hinted that Pab1's expected role in aiding termination became prominent only when eRF1 activity was inefficient. Although this result may be prone to artifacts, as 3'-UTR length also correlates with other features of the mRNAs (S3 Fig), our

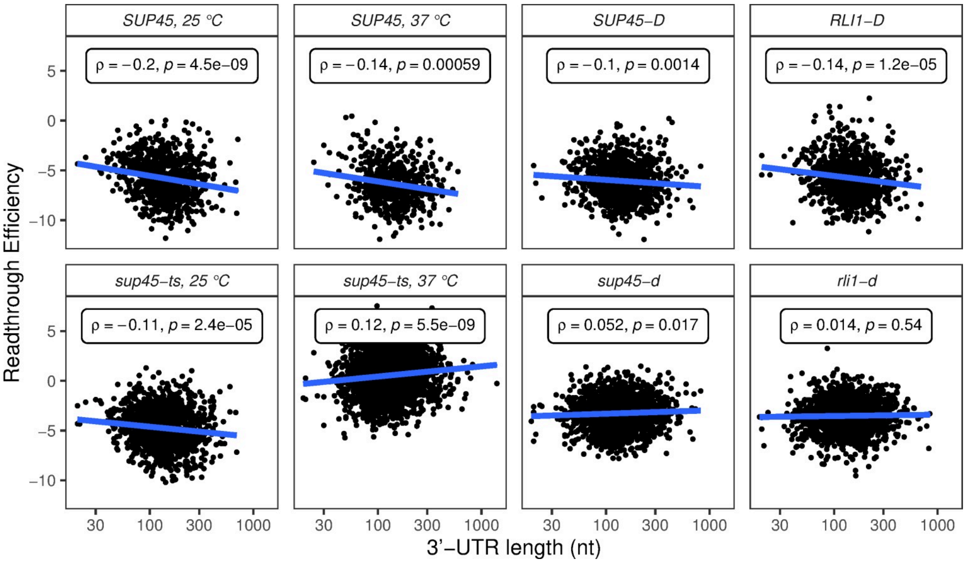

Fig 8. Readthrough efficiency increased with 3'-UTR lengths in eRF1 mutants, but not in wild-type or ribosome recycling factor mutant cells. Scatter plots of readthrough efficiency vs. 3'-UTR length. Spearman's correlation coefficient (ρ) and p-value were calculated for each sample.

https://doi.org/10.1371/journal.pgen.1009538.g008

finding was consistent with multiple studies linking the distance from the stop codon to Pab1 and termination efficiency, such as recent studies of context dependent translation termination in Ciliates [37].

Discussion

In this study, we analyzed ribosome profiling data of WT and mutant yeast strains defective in translation termination or ribosome recycling to identify mRNA or peptide features that influence readthrough efficiency at a genome-wide level. First, we characterized the phenotypes of eRF1 mutant cells, which included ribosomes stalling at start and stop codons and increased in-frame footprints in 3'-UTRs compared to wild-type cells (Figs 1 and 2B). We took these phenotypes into account when calculating readthrough efficiency of each transcript (Fig 3A) and excluded any transcripts that lack 3'-UTR annotation or sufficient footprints in the 3'-UTR from further analyses. Although these considerations left us with a smaller number of analyzable data points, we were able to reduce noise in the data that confounded a previous analysis attempt [21]. With our analysis strategies, we demonstrated that the readthrough-promoting stop codon and proximal nucleotides previously determined in reporter assays [16,26–28] are not reporter-specific, but also occur at a genome-wide level in yeast for endogenous mRNAs (Figs 5 and 6). Moreover, our analyses identify readthrough-inhibiting nucleotides that were refractory to previous studies using reporter gene assays because the low level readthrough was below the detection threshold.

Two novel mRNA features that we found to be major determinants of readthrough efficiency were the codon in the ribosomal P site (when a stop codon is in the A site) and 3'-UTR length. For the P-site codon (nt -3, -2, and -1), our results did not support previous conclusions of AA being the most readthrough-permissive nucleotides at position -2 and -1 [5,24,25] when we analyzed the data by nucleotide position (Figs 5 and 6 and Table 2). Moreover, when we analyzed nt -3, -2, and -1 together as a codon in the P site, we did not find codons NAA to be significantly associated with higher readthrough. Instead, we found that genes with AUA, ACA, ACC, CUG, and GAC as their penultimate codon had higher readthrough efficiencies compared to the sample median (Fig 7A), suggesting that there may be properties in the P site other than adenines that could influence readthrough. Previous studies ruled out properties of the last amino acid residue (encoded by the P-site codon) as features affecting readthrough, but reported conflicting results for tRNA in the P site [24,25]. We were also able to rule out amino acid properties, and although we could not pinpoint the importance of tRNA properties, we could conclude that no single tRNA property (such as specific wobble pair or modification) was responsible for high readthrough. Nevertheless, we could not rule out synergistic effects of both nucleotide and tRNA on readthrough efficiency. The same arguments applied to our discovery of readthrough-inhibiting candidates: CGU, GCU, UUA, and UUC. Further experiments are needed to deconstruct the roles of nucleotide and tRNA in mediating readthrough efficiency.

Although ribosome recycling was not the focus of our study, our use of profiling data from recycling factor-depleted cells [41] allowed us to observe that some nucleotides and codons in the P site have specific effects in *rli1-d* cells but not in wild-type or eRF1 mutant cells (Figs 5–7). Since wild-type and eRF1 mutant cells were both wild-type for recycling factors, these differences may be attributed to the recycling step. As failure to recycle leads to ribosomes reinitiating randomly within <10 nt downstream of the stop codon [41], there is ~33% chance that reinitiation stays in-frame with the ORF, and there was no way for us to distinguish these footprints from actual readthrough footprints. Thus, we speculated that the differences in "readthrough" efficiency between recycling factor-depleted cells and other samples could mean these nucleotides or codons affect the recycling step. If this is true, cautious use and interpretation of dual/bi-cistronic reporter assays in measuring readthrough efficiency may need to be exercised, as reinitiation that is in-frame with the original reading frame could still produce a functional readthrough reporter. This is evidenced by previous claims borne out of dual reporter experiments stating that functions of Rli1 and Hcr1 were to control termination and readthrough [52,53], while ribosome profiling data and subsequent experiments demonstrated their ribosome recycling properties [41,54]. Unless the readthrough product was defined by the combined size of both internal control and readthrough reporters, enzymatic or colorimetric assays that separately detect the internal control reporter and readthrough reporter may not be ideal methods for studying readthrough contexts.

Another mRNA feature that we found to be a significant determinant of readthrough efficiency was 3'-UTR length. Intriguingly, readthrough efficiency decreased with 3'-UTR length in WT cells but increased with 3'-UTR length in eRF1 mutant cells. The trend in eRF1 mutant cells is consistent with the hypothesis that proximity of a stop codon to PABP enhances termination efficiency [38]. The fact that only eRF1 mutant cells displayed such a trend implied that PABP's expected role in enhancing termination is principally observable when eRF1 function is not efficient. This result is consistent with recent work from our laboratory demonstrating that deletion of the yeast *PAB1* gene led to a reduced accumulation of premature termination products, increased readthrough at multiple PTCs, and increased readthrough of an otherwise "weak" PTC in parallel with 3'-UTR lengthening of the respective transcripts [38]. We surmise that the proximity of the stop codon to PABP, which is thought to enhance termination efficiency through interaction of PABP with eRF3 [36,55], is masked in the ribosome profiling data of WT cells where eRF1 is fully functional, but when this protein became less efficient or absent, PABP's role in recruiting eRF3 or other factors could be distinguished. The relationship between PABP's proximity to a stop codon and the efficiency of termination parallels that of PABP and mRNA stability in the "faux 3'-UTR" model for NMD, where increasing PABP's distance from a stop codon led to aberrant termination and destabilized the mRNAs [42,56].

Although normal and premature termination processes have been shown to differ in efficiency, they share the fundamental requirement for release factors and a stop codon [38,42,56]. While our study largely involved normal termination codons, we provided additional features of the mRNA sequences that influence readthrough, and hence termination efficiencies, which could be useful in designing or understanding therapeutic approaches for the large number of diseases caused by nonsense mutations [4].

Materials and methods

Yeast strains and culture growth conditions

Yeast strains used in this study are in the W303 background. The temperature-sensitive *sup45-2* strain (HFY1218) was derived from the wild-type strain (HFY114) [57] by the pop-in/pop-out gene replacement technique. Cells were grown in 1 L of YPD at 25˚C with shaking. When the OD_{600} of the culture reached 0.6–0.8, its temperature was shifted as follows: Cells were collected by centrifugation at 5,000 rpm for 5 min at room temperature (RT) using a JA10 rotor in a Beckman Coulter Avanti J-E centrifuge, resuspended in 400 ml of fresh YPD media, transferred to a new flask, and incubated in a 25˚C water bath with shaking. After 20 min of incubation, 400 ml of pre-warmed (57˚C) YPD were added to the flask, and cells were incubated in a 37˚C water bath for 30 minutes with shaking. As a control for the temperature shift procedure, the same protocol was carried out except that YPD was pre-warmed to 25˚C and the subsequent 30-minute incubation was at 25˚C.

Library preparation and sequencing

Ribosome profiling and RNA-seq libraries were prepared as described previously in our library preparation protocol for immunoprecipitated ribosomes [58]. Briefly, 1 L of cells of OD_{600} between 0.6 and 0.8 were harvested by rapid vacuum filtration and flash-frozen in liquid nitrogen in the presence of Footprinting Buffer (20 mM Tris-HCl pH 7.4, 150 mM NaCl, 5 mM $MgCl_2$) plus 1% TritonX-100, 0.5 mM DTT, 1 mM phenylmethylsulfonyl fluoride (PMSF) and 1X protease inhibitors. Cells were lysed in a Cryomill (Retsch) at 5 Hz for 2 min, then 10 Hz for 15 min. Lysates were clarified by ultracentrifugation in a Beckman Coulter Optima L-90K Ultracentrifuge at 18,000 rpm for 10 min at 4˚C, using a 50Ti rotor. Centrifugation was repeated for the supernatant at 18,000 rpm for 15 min at 4˚C.

For ribosome profiling [59], lysates were digested with RNaseI (Invitrogen, #AM2294), then layered onto a 1 M sucrose cushion in Footprinting Buffer plus 0.5 mM DTT and centrifuged in a Beckman Optima TLX Ultracentrifuge at 60,000 rpm for 1 hour at 4˚C using a TLA100.3 rotor to isolate 80S ribosomes. Ribosome-protected fragments were extracted using a miRNeasy kit (QIAgen, #217004) and depleted of rRNA using a Yeast Ribo-Zero Gold rRNA removal kit (Illumina, #MRZY1324) according to the manufacturer's protocol. Multiplexed cDNA libraries were prepared with the NEXTFlex Small RNA-Seq Kit v3 (BIOO Scientific, #NOVA-5132-06) according to the manufacturer's protocol and sequenced (single-end, 75 cycles) on a NextSeq500.

For RNA-Seq, total RNAs were extracted from lysates using a miRNeasy kit (QIAgen, #217004), depleted of genomic DNA contamination using Baseline-Zero DNase (Epicentre, #DB0715K), and depleted of rRNA using a Yeast Ribo-Zero Gold rRNA removal kit (Illumina, #MRZY1324) according to the manufacturer's protocol. Multiplexed cDNA libraries were prepared with a TruSeq Stranded mRNA sample prep kit (Illumina, #RS-122-2101) according to the manufacturer's protocol and sequenced (single-end, 75 cycles) on a NextSeq500.

Reads processing and alignment

The following software packages were used to process RNA- and ribo-seq library sequences: fastqc v0.10.1 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/); cutadapt v1.9 [60]; samtools v0.1.19 [61]; bowtie v1.0.0 [62], cufflinks v2.2.1 [63]; bam2fastq v1.1.0 (https://gsl.hudsonalpha.org/information/software/bam2fastq); RSEM v1.3.0 [64].

A customized yeast transcriptome was constructed from the S288C reference genome sequence and annotations downloaded from the Saccharomyces Genome Database (https://www.yeastgenome.org) on September 10, 2015 (S288C_reference_genome_R64-2-1_20150113). Annotations for the following classes of transcripts were parsed from the genomic gff3-formatted annotation file and processed separately: protein coding genes (6551); intron-containing genes (272); genes with 5'-UTR introns (24); frameshifted genes (47); blocked and pseudogenes (18), non-coding RNAs (ncRNAs) (16). For all but the last class, 5'- and 3'-UTR entries were added to the individual gff files employing the UTR annotations from Nagalakshmi et al. [65]. 5'- and 3'-UTR information was available for 2840 and 2849 genes, respectively. In cases where no UTR was annotated, we used the 75th percentile lengths of 97 and 173 nt, respectively. Unspliced versions (pre-mRNAs) of all intron-containing genes were defined as a continuous exon extending from the start of the 5'-UTR to end of the 3'-UTR. After editing the individual parsed gff files, they were concatenated, and a fasta file with sequences of all 6952 transcripts was generated using the cufflinks gffread program. Information included in the fasta header lines generated by gffread was used to construct a transcript-to-gene-map file for use with RSEM, which was used for gene- and isoform-level quantification of transcripts. The fasta file is available at https://github.com/Jacobson-Lab/yeast_transcriptome_v5. A modified transcriptome was used for riboWaltz (see Data Analyses section). The riboWaltz transcriptome contained only the spliced transcripts of protein coding genes (6551) and genes with 5'-UTR introns (24). In addition, the riboWaltz transcriptome considered the stop codon as a part of the 3'-UTR region rather than the CDS region, thus the length of CDS and 3'-UTR regions of each gene were adjusted accordingly.

For ribosome profiling libraries, de-multiplexed sequences were first processed by cutadapt to filter out low quality reads and read lengths shorter than 23 nt while trimming adapter sequences (cutadapt -a TGGAATTCTCGGGTGCCAAGG -q 10—trim-n -m 23). Reads aligning to non-protein coding RNAs (rRNA, tRNA, snoRNA, and other non-protein coding RNA sequences) were removed by aligning with bowtie (-m 100–3 4–5 4 -n 2 -l 15—un) and retaining only unaligned reads. The random 4-mers introduced on both sides of each RPF during library preparation were trimmed, concatenated, and stored as an identifier "barcode" used to remove PCR duplicates by aligning reads to the transcriptome (bowtie -m 10 -n 2 -l 15 -S—best—strata) and retaining only a single barcode (and associated read) at each position where multiple reads aligned. After duplicates removal, the SAM-formatted alignment file was converted either to BAM format using samtools or to fastq format using bam2fastq for use in downstream analyses. Transcript abundance measurements were determined by RSEM with settings—seed-length 15—bowtie-m 10. Reads processing and alignment statistics can be found in S3 Table.

The TruSeq stranded RNA-Seq libraries were processed using RSEM with no additional pre-processing.

Public ribosome profiling data sets

Public ribosome profiling data of *sup45-d* [40], *rli1-d* [41], and their WT counterparts were downloaded from NCBI's Gene Expression Omnibus (GEO) database. Yeast strain genotypes, cell growth conditions, accession numbers of the data series and specific sequencing runs are outlined in S1 Table. After download, reads were adapter-trimmed and length-filtered by cutadapt (cutadapt -a CTGTAGGCACCATCAAT -q 10—trim-n -m 15), after which reads aligning to non-protein coding RNAs were removed (bowtie -m 100 -n 2 -l 15—un) and remaining reads aligned to the transcriptome exactly as described above, with the omission of the PCR duplicate removal step. Reads processing and alignment statistics can be found in S3 Table.

Data analyses

Analyses of ribosome profiling data were performed in the R software environment (versions 3.5 and 3.6). RPFs of 20–23 nt and 27–32 nt in length were used for the analyses. P-site offsets of reads in ribosome profiling libraries were assigned using the R package riboWaltz [66], the results of which are provided in S4 Table. After assigning P-site location, reads from replicate libraries were combined. In addition to riboWaltz, the following packages were used for data analysis and visualizations: data.table, dplyr, reshape2, seqinr, ggplot2, and scales. Built-in base R functions, rcompanion, and ggpubr were used for statistical analyses. randomForest and caret were used to generate and analyze random forest models.

For analyses results that were visualized in a heatmap form, extreme results were made to fit the indicated scale in order to not jeopardize the visualization of most of the results. The actual results for each figure and associated R codes used to generate the figure are provided for reference at https://github.com/Jacobson-Lab/sup45-ts_readthrough.

Random forest models

Random forest analyses [44] were implemented in the R environment using R packages randomForest [45] and caret [67]. A classification (predicts readthrough groups) and a regression (predicts readthrough efficiency) model were created for each sample. Each model was trained with 100 trees (parameter ntree = 100) and 81 mRNA features. The number of features chosen to split at each tree node varied (parameter mtry ranging from 1 to 81, increasing by 10), with the number resulting in the highest accuracy automatically selected by the program. Model accuracy was evaluated by 5-fold cross-validation using receiver operating characteristic (ROC) as the performance metric for classification model and root mean squared error (RMSE) for regression model. To report performance metric for classification, area under the ROC curve (AUROC) for each fold was extracted, and the average ± standard deviation across 5 folds was calculated. For regression, RMSE was normalized by the range of the Y variable (readthrough efficiency) for each fold, and the average ± standard deviation across 5 folds was calculated.

Interpretation of each performance metric is as follows. AUROC is a common metric used to assess classification model. An ROC curve is generated by plotting true positive rates (TPR) against false positive rates (FPR) of multiple classification thresholds. A straight diagonal line indicates that TPR equals to FPR, resulting in area under the curve (AUROC) of 0.5, and means the chance that the model can classify genes into the correct groups is 50%–this is no better than random chance. Thus, the higher AUROC is than 0.5, the better the model performs, and AUROC of 1 indicates a perfect classification model [68]. Acceptable value for AUROC depends on the context of the study. For our study, we think the AUROC of ~0.7–0.8 is acceptable.

Root mean squared error (RMSE), one of the performance metrics for regression, is the average of the distance between the predicted values and the actual values. The unit of RMSE is the same as the Y variable. Whether the error is considered large or small depends on the range of the data. For instance, an average error of ± 0.5 may be considered small for the data ranging from 0–100, but considered large for the data ranging from 0–1. Therefore, in order to easily assess the errors across samples, we normalized the RMSE by the range of the data, where the resulting normalized RMSE (NRMSE) displays the average fraction of error relative to data range.

After the random forest models were trained, feature importance for each X variable was extracted with the importance() function. We selected Mean Decrease Accuracy (MDA) and % Increase in Mean Squared Error (%IncMSE) as measures of feature importance for classification and regression models, respectively.

Supporting information

S1 Fig. Reproducibility between replicates of *sup45* temperature-shift ribosome profiling and RNA-Seq datasets. Correlation between ribosome density (top) or mRNA abundance (bottom) in reads per kilobase per million (RPKM) of a pair of replicates for each yeast strain and growth temperature. Pearson's correlation coefficient (r) was calculated and reported for each pair of replicates.
(TIF)

S2 Fig. Readthrough efficiency of PTC-containing alleles and ribosome occupancy in the transcripts of two intron-containing genes. A. Average readthrough efficiency of PTC-readthrough in the *ade2-1* and *can1-100* alleles. PTC readthrough efficiency was calculated by dividing frame 0 footprint count downstream of the PTC to that upstream of the PTC. Average ± standard deviation of two replicates were plotted for each sample, with black dots indicating the individual data points. **B.** and **C.** Read coverage tracks from the Integrative Genomics Viewer (IGV) [69] showing coverage of ribosome profiling reads for the intron-containing genes, **B.** *RPL28* and **C.** *RPS0A*, in *SUP45* and *sup45-ts* strains at 25˚C and 37˚C. Yellow rectangles indicate the position of termination codons in frame with the respective initiation codons under conditions where the introns are translated. Full scale for A and B equals 50 reads.
(TIF)

S3 Fig. Correlation matrix of relationships between readthrough efficiency, gene expression, translation efficiency, codon optimality, and transcript length. Spearman's correlation coefficient (ρ) was calculated and reported for each pair of variables.
(TIF)

S1 Table. Ribosome profiling data and yeast strains.
(PDF)

S2 Table. Number of genes involved in statistical analyses.
(PDF)

S3 Table. Ribosome profiling reads processing and alignment statistics.
(PDF)

S4 Table. Offset from 5' and 3' ends of footprint to the first nucleotide in the P-site for each footprint length in each sample.
(PDF)

Acknowledgments

We thank Jill Moore, Zhiping Weng, Andrei Korostelev, Barry Cooperman, and Sean Ryder for helpful discussions.

Author Contributions

Conceptualization: Kotchaphorn Mangkalaphiban, Feng He, Allan Jacobson.

Data curation: Kotchaphorn Mangkalaphiban, Richard Baker.

Formal analysis: Kotchaphorn Mangkalaphiban, Feng He, Richard Baker, Allan Jacobson.

Funding acquisition: Allan Jacobson.

Investigation: Kotchaphorn Mangkalaphiban.

Methodology: Kotchaphorn Mangkalaphiban, Feng He, Robin Ganesan, Allan Jacobson.

Project administration: Robin Ganesan, Allan Jacobson.

Resources: Robin Ganesan, Chan Wu, Allan Jacobson.

Software: Kotchaphorn Mangkalaphiban, Richard Baker.

Supervision: Robin Ganesan, Chan Wu, Allan Jacobson.

Validation: Kotchaphorn Mangkalaphiban.

Visualization: Kotchaphorn Mangkalaphiban.

Writing – original draft: Kotchaphorn Mangkalaphiban, Allan Jacobson.

Writing – review & editing: Kotchaphorn Mangkalaphiban, Feng He, Robin Ganesan, Chan Wu, Richard Baker, Allan Jacobson.

References

1. Hellen CUT. Translation Termination and Ribosome Recycling in Eukaryotes. Cold Spring Harb Perspect Biol. 2018 Oct; 10(10):a032656. https://doi.org/10.1101/cshperspect.a032656 PMID: 29735640
2. Salas-Marco J, Bedwell DM. GTP Hydrolysis by eRF3 Facilitates Stop Codon Decoding during Eukaryotic Translation Termination. Mol Cell Biol. 2004 Sep 1; 24(17):7769–78. https://doi.org/10.1128/MCB.24.17.7769-7778.2004 PMID: 15314182
3. Schuller AP, Green R. Roadblocks and resolutions in eukaryotic translation. Nat Rev Mol Cell Biol. 2018 Aug; 19(8):526–41. https://doi.org/10.1038/s41580-018-0011-4 PMID: 29760421
4. Dabrowski M, Bukowy-Bieryllo Z, Zietkiewicz E. Translational readthrough potential of natural termination codons in eucaryotes–The impact of RNA sequence. RNA Biol. 2015 Sep 2; 12(9):950–8. https://doi.org/10.1080/15476286.2015.1068497 PMID: 26176195
5. Rodnina MV, Korniy N, Klimova M, Karki P, Peng B-Z, Senyushkina T, et al. Translational recoding: canonical translation mechanisms reinterpreted. Nucleic Acids Res. 2020 Feb 20; 48(3):1056–67. https://doi.org/10.1093/nar/gkz783 PMID: 31511883
6. Brenner S, Stretton AOW, Kaplan S. Genetic Code: The 'Nonsense' Triplets for Chain Termination and their Suppression. Nature. 1965 Jun; 206(4988):994–8. https://doi.org/10.1038/206994a0 PMID: 5320272
7. Schueren F, Thoms S. Functional Translational Readthrough: A Systems Biology Perspective. Brosius J, editor. PLOS Genet. 2016 Aug 4; 12(8):e1006196. https://doi.org/10.1371/journal.pgen.1006196 PMID: 27490485
8. Skuzeski JM, Nichols LM, Gesteland RF, Atkins JF. The signal for a leaky UAG stop codon in several plant viruses includes the two downstream codons. J Mol Biol. 1991 Mar 20; 218(2):365–73. https://doi.org/10.1016/0022-2836(91)90718-I PMID: 2010914
9. De Bellis M, Pisani F, Mola MG, Rosito S, Simone L, Buccoliero C, et al. Translational readthrough generates new astrocyte AQP4 isoforms that modulate supramolecular clustering, glial endfeet localization, and water transport: DE BELLIS et al. Glia. 2017 May; 65(5):790–803. https://doi.org/10.1002/glia.23126 PMID: 28206694
10. Eswarappa SM, Potdar AA, Koch WJ, Fan Y, Vasu K, Lindner D, et al. Programmed Translational Readthrough Generates Antiangiogenic VEGF-Ax. Cell. 2014 Jun; 157(7):1605–18. https://doi.org/10.1016/j.cell.2014.04.033 PMID: 24949972

11. Hofhuis J, Schueren F, Nötzel C, Lingner T, Gärtner J, Jahn O, et al. The functional readthrough extension of malate dehydrogenase reveals a modification of the genetic code. Open Biol. 2016 Nov; 6 (11):160246. https://doi.org/10.1098/rsob.160246 PMID: 27881739

12. Loughran G, Chou M-Y, Ivanov IP, Jungreis I, Kellis M, Kiran AM, et al. Evidence of efficient stop codon readthrough in four mammalian genes. Nucleic Acids Res. 2014 Aug 18; 42(14):8928–38. https://doi.org/10.1093/nar/gku608 PMID: 25013167

13. Loughran G, Jungreis I, Tzani I, Power M, Dmitriev RI, Ivanov IP, et al. Stop codon readthrough generates a C-terminally extended variant of the human vitamin D receptor with reduced calcitriol response. J Biol Chem. 2018 Mar 23; 293(12):4434–44. https://doi.org/10.1074/jbc.M117.818526 PMID: 29386352

14. Namy O, Duchateau-Nguyen G, Rousset J-P. Translational readthrough of the PDE2 stop codon modulates cAMP levels in Saccharomyces cerevisiae. Mol Microbiol. 2002; 43(3):641–52. https://doi.org/10.1046/j.1365-2958.2002.02770.x PMID: 11929521

15. Rajput B, Pruitt KD, Murphy TD. RefSeq curation and annotation of stop codon recoding in vertebrates. Nucleic Acids Res. 2019 Jan 25; 47(2):594–606. https://doi.org/10.1093/nar/gky1234 PMID: 30535227

16. Schueren F, Lingner T, George R, Hofhuis J, Dickel C, Gärtner J, et al. Peroxisomal lactate dehydrogenase is generated by translational readthrough in mammals. eLife. 2014 Sep 23; 3:e03640.

17. Steneberg P, Samakovlis C. A novel stop codon readthrough mechanism produces functional Headcase protein in Drosophila trachea. EMBO Rep. 2001 Jul 1; 2(7):593–7. https://doi.org/10.1093/embo-reports/kve128 PMID: 11463742

18. Yamaguchi Y, Baba H. Phylogenetically Conserved Sequences Around Myelin P0 Stop Codon are Essential for Translational Readthrough to Produce L-MPZ. Neurochem Res. 2018 Jan; 43(1):227–37. https://doi.org/10.1007/s11064-017-2423-5 PMID: 29081003

19. Dunn JG, Foo CK, Belletier NG, Gavis ER, Weissman JS. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. eLife. 2013 Dec 3; 2:e01179. https://doi.org/10.7554/eLife.01179 PMID: 24302569

20. Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, et al. Evidence of abundant stop codon readthrough in Drosophila and other metazoa. Genome Res. 2011 Dec 1; 21(12):2096–113. https://doi.org/10.1101/gr.119974.110 PMID: 21994247

21. Kleppe AS, Bornberg-Bauer E. Robustness by intrinsically disordered C-termini and translational readthrough. Nucleic Acids Res. 2018 Nov 2; 46(19):10184–94. https://doi.org/10.1093/nar/gky778 PMID: 30247639

22. Tate WP, Cridge AG, Brown CM. 'Stop' in protein synthesis is modulated with exquisite subtlety by an extended RNA translation signal. Biochem Soc Trans. 2018 Nov 12;BST20180190. https://doi.org/10.1042/BST20180190 PMID: 30420414

23. Bonetti B, Fu L, Moon J, Bedwell DM. The Efficiency of Translation Termination is Determined by a Synergistic Interplay Between Upstream and Downstream Sequences inSaccharomyces cerevisiae. J Mol Biol. 1995 Aug; 251(3):334–45. https://doi.org/10.1006/jmbi.1995.0438 PMID: 7650736

24. Mottagui-Tabar S, Tuite MF, Isaksson LA. The influence of 5' codon context on translation termination in Saccharomyces cerevisiae. Eur J Biochem. 1998 Oct 1; 257(1):249–54. https://doi.org/10.1046/j.1432-1327.1998.2570249.x PMID: 9799126

25. Tork S, Hatin I, Rousset J-P, Fabret C. The major 5′ determinant in stop codon read-through involves two adjacent adenines. Nucleic Acids Res. 2004; 32(2):415–21. https://doi.org/10.1093/nar/gkh201 PMID: 14736996

26. Anzalone AV, Zairis S, Lin AJ, Rabadan R, Cornish VW. Interrogation of Eukaryotic Stop Codon Readthrough Signals by *in Vitro* RNA Selection. Biochemistry. 2019 Feb 26; 58(8):1167–78. https://doi.org/10.1021/acs.biochem.8b01280 PMID: 30698415

27. Cridge AG, Crowe-McAuliffe C, Mathew SF, Tate WP. Eukaryotic translational termination efficiency is influenced by the 3′ nucleotides within the ribosomal mRNA channel. Nucleic Acids Res. 2018 Feb 28; 46(4):1927–44. https://doi.org/10.1093/nar/gkx1315 PMID: 29325104

28. Namy O, Hatin I, Rousset J-P. Impact of the six nucleotides downstream of the stop codon on translation termination. EMBO Rep. 2001 Sep 15; 2(9):787–93. https://doi.org/10.1093/embo-reports/kve176 PMID: 11520858

29. Firth AE, Wills NM, Gesteland RF, Atkins JF. Stimulation of stop codon readthrough: frequent presence of an extended 3′ RNA structural element. Nucleic Acids Res. 2011 Aug; 39(15):6679–91. https://doi.org/10.1093/nar/gkr224 PMID: 21525127

30. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science. 2009 Apr 10; 324(5924):218–23. https://doi.org/10.1126/science.1168978 PMID: 19213877

31. Li C, Zhang J. Stop-codon read-through arises largely from molecular errors and is generally nonadaptive. PLOS Genet. 2019 May 23; 15(5):e1008141. https://doi.org/10.1371/journal.pgen.1008141 PMID: 31120886

32. Wangen JR, Green R. Stop codon context influences genome-wide stimulation of termination codon readthrough by aminoglycosides. eLife. 2020 Jan 23; 9:e52611. https://doi.org/10.7554/eLife.52611 PMID: 31971508

33. Brown A, Shao S, Murray J, Hegde RS, Ramakrishnan V. Structural basis for stop codon recognition in eukaryotes. Nature. 2015 Aug; 524(7566):493–6. https://doi.org/10.1038/nature14896 PMID: 26245381

34. Shao S, Murray J, Brown A, Taunton J, Ramakrishnan V, Hegde RS. Decoding Mammalian Ribosome-mRNA States by Translational GTPase Complexes. Cell. 2016 Nov 17; 167(5):1229–1240.e15. https://doi.org/10.1016/j.cell.2016.10.046 PMID: 27863242

35. Wilson DN, Arenz S, Beckmann R. Translation regulation via nascent polypeptide-mediated ribosome stalling. Curr Opin Struct Biol. 2016 Apr 1; 37:123–33. https://doi.org/10.1016/j.sbi.2016.01.008 PMID: 26859868

36. Ivanov A, Mikhailova T, Eliseev B, Yeramala L, Sokolova E, Susorov D, et al. PABP enhances release factor recruitment and stop codon recognition during translation termination. Nucleic Acids Res. 2016 Sep 19; 44(16):7766–76. https://doi.org/10.1093/nar/gkw635 PMID: 27418677

37. Swart EC, Serra V, Petroni G, Nowacki M. Genetic Codes with No Dedicated Stop Codon: Context-Dependent Translation Termination. Cell. 2016 Jul 28; 166(3):691–702. https://doi.org/10.1016/j.cell.2016.06.020 PMID: 27426948

38. Wu C, Roy B, He F, Yan K, Jacobson A. Poly(A)-Binding Protein Regulates the Efficiency of Translation Termination. Cell Rep. 2020 Nov; 33(7):108399. https://doi.org/10.1016/j.celrep.2020.108399 PMID: 33207198

39. Stansfield I, Kushnirov VV, Jones KM, Tuite MF. A conditional-Lethal Translation Termination Defect in a sup45 Mutant of the Yeast Succhuromyces Cerevisiue. Eur J Biochem. 1997 May; 245(3):557–63. https://doi.org/10.1111/j.1432-1033.1997.00557.x PMID: 9182990

40. Wu CC-C, Zinshteyn B, Wehner KA, Green R. High-Resolution Ribosome Profiling Defines Discrete Ribosome Elongation States and Translational Regulation during Cellular Stress. Mol Cell. 2019 Mar; 73(5):959–970.e5. https://doi.org/10.1016/j.molcel.2018.12.009 PMID: 30686592

41. Young DJ, Guydosh NR, Zhang F, Hinnebusch AG, Green R. Rli1/ABCE1 Recycles Terminating Ribosomes and Controls Translation Reinitiation in 3′UTRs In Vivo. Cell. 2015 Aug; 162(4):872–84. https://doi.org/10.1016/j.cell.2015.07.041 PMID: 26276635

42. He F, Jacobson A. Nonsense-Mediated mRNA Decay: Degradation of Defective Transcripts Is Only Part of the Story. Annu Rev Genet. 2015 Nov 23; 49(1):339–66. https://doi.org/10.1146/annurev-genet-112414-054639 PMID: 26436458

43. Celik A, Baker R, He F, Jacobson A. High-resolution profiling of NMD targets in yeast reveals translational fidelity as a basis for substrate selection. RNA. 2017 May; 23(5):735–48. https://doi.org/10.1261/rna.060541.116 PMID: 28209632

44. Breiman L. Random Forests. Mach Learn. 2001 Oct 1; 45(1):5–32.

45. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002; 2(3):18–22.

46. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011 Nov 24; 6(1):26. https://doi.org/10.1186/1748-7188-6-26 PMID: 22115189

47. Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. 2015 Jul 1; 43(W1):W389–94. https://doi.org/10.1093/nar/gkv332 PMID: 25883141

48. Pechmann S, Frydman J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol. 2013 Feb; 20(2):237–43. https://doi.org/10.1038/nsmb.2466 PMID: 23262490

49. Reis M dos, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 2004; 32(17):5036–44. https://doi.org/10.1093/nar/gkh834 PMID: 15448185

50. Presnyak V, Alhusaini N, Chen Y-H, Martin S, Morris N, Kline N, et al. Codon Optimality Is a Major Determinant of mRNA Stability. Cell. 2015 Mar; 160(6):1111–24. https://doi.org/10.1016/j.cell.2015.02.029 PMID: 25768907

51. Johansson MJO, Esberg A, Huang B, Björk GR, Byström AS. Eukaryotic Wobble Uridine Modifications Promote a Functionally Redundant Decoding System. Mol Cell Biol. 2008 May 15; 28(10):3301–12. https://doi.org/10.1128/MCB.01542-07 PMID: 18332122

52. Beznosková P, Cuchalová L, Wagner S, Shoemaker CJ, Gunišová S, von der Haar T, et al. Translation Initiation Factors eIF3 and HCR1 Control Translation Termination and Stop Codon Read-Through in Yeast Cells. Hinnebusch AG, editor. PLoS Genet. 2013 Nov 21; 9(11):e1003962. https://doi.org/10.1371/journal.pgen.1003962 PMID: 24278036

53. Khoshnevis S, Gross T, Rotte C, Baierlein C, Ficner R, Krebber H. The iron–sulphur protein RNase L inhibitor functions in translation termination. EMBO Rep. 2010 Mar; 11(3):214–9. https://doi.org/10.1038/embor.2009.272 PMID: 20062004

54. Young DJ, Guydosh NR. Hcr1/eIF3j Is a 60S Ribosomal Subunit Recycling Accessory Factor In Vivo. Cell Rep. 2019 Jul 2; 28(1):39–50.e4. https://doi.org/10.1016/j.celrep.2019.05.111 PMID: 31269449

55. Roque S, Cerciat M, Gaugué I, Mora L, Floch AG, Zamaroczy M de, et al. Interaction between the poly (A)-binding protein Pab1 and the eukaryotic release factor eRF3 regulates translation termination but not mRNA decay in Saccharomyces cerevisiae. RNA. 2015 Jan 1; 21(1):124–34. https://doi.org/10.1261/rna.047282.114 PMID: 25411355

56. Amrani N, Ganesan R, Kervestin S, Mangus DA, Ghosh S, Jacobson A. A *faux* 3′-UTR promotes aberrant termination and triggers nonsense- mediated mRNA decay. Nature. 2004 Nov; 432(7013):112–8. https://doi.org/10.1038/nature03060 PMID: 15525991

57. He F, Li X, Spatrick P, Casillo R, Dong S, Jacobson A. Genome-Wide Analysis of mRNAs Regulated by the Nonsense-Mediated and 5′ to 3′ mRNA Decay Pathways in Yeast. Mol Cell. 2003 Dec; 12(6):1439–52. https://doi.org/10.1016/s1097-2765(03)00446-5 PMID: 14690598

58. Ganesan R, Leszyk J, Jacobson A. Selective profiling of ribosomes associated with yeast Upf proteins. Methods. 2019 Feb 15; 155:58–67. https://doi.org/10.1016/j.ymeth.2018.12.008 PMID: 30593864

59. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation *in vivo* by deep sequencing of ribosome-protected mRNA fragments. Nat Protoc. 2012 Aug; 7(8):1534–50. https://doi.org/10.1038/nprot.2012.086 PMID: 22836135

60. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011 May 2; 17(1):10.

61. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15; 25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

62. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10(3):R25. https://doi.org/10.1186/gb-2009-10-3-r25 PMID: 19261174

63. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May; 28(5):511–5. https://doi.org/10.1038/nbt.1621 PMID: 20436464

64. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011 Dec; 12(1):323.

65. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science. 2008 Jun 6; 320(5881):1344–9. https://doi.org/10.1126/science.1158441 PMID: 18451266

66. Lauria F, Tebaldi T, Bernabò P, Groen EJN, Gillingwater TH, Viero G. riboWaltz: Optimization of ribosome P-site positioning in ribosome profiling data. PLOS Comput Biol. 2018 Aug 13; 14(8):e1006169. https://doi.org/10.1371/journal.pcbi.1006169 PMID: 30102689

67. Kuhn M. Building Predictive Models in *R* Using the caret Package. J Stat Softw. 2008; 28(5).

68. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. New York, NY: Springer New York; 2013. (Springer Texts in Statistics; vol. 103).

69. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011 Jan; 29(1):24–6. https://doi.org/10.1038/nbt.1754 PMID: 21221095