UCSF
UC San Francisco Electronic Theses and Dissertations

Title
Stop Codon Readthrough is a Feature of Eukaryotic Translation

Permalink
https://escholarship.org/uc/item/9td666zh

Author
Dunn, Joshua Griffin

Publication Date
2015

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Stop Codon Readthrough is a Feature of Eukaryotic Translation

by

Joshua Griffin Dunn

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOCHEMISTRY & MOLECULAR BIOLOGY

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Acknowledgements

Many thanks are due to many people- quite a few more than I can name here. At the very least, I would like to thank the members of the Weissman lab from 2008-2015 for creating so open, collaborative, and generous a space. In particular, I thank Jonathan Weissman for his continued trust and support, and for the unusual clarity with which he sees the big pictures; Noam Stern-Ginossar, Calvin Jan, and Nick Ingolia for criticism and encouragement; Edwin Rodriguez, Alex Fields, Elizabeth Costa, Yi-Chang Liu, Silvi Rouskin, Chris Williams, Gloria Brar, & Michelle Chan for friendship; and Manny DeVera & Christopher Reiger for keeping everything running.

I thank John Atkins for introducing me to stop codon readthrough, and Cat Foo, Nicolette Belletier, Elizabeth Gavis, and Jonathan Weissman for pursuing that path with me for the 2013 eLife paper. For work not included here, I add special thanks to Alex Fields and John Hawkins for useful conversation and criticism concerning numerous topics in computational biology.

I thank my thesis committee, Carol Gross and Pat O'Farrell, for their critcism, judgement, thoughtfulness, and foresight. I also thank my previous scientific mentors Josh Trueheart, Peter Houston, & Maria Mayorga for showing me how to have fun in science; Etchell Cordero for her willingness to take risks on me; and Zuzana Storchova for patiently pushing the limits of my critical thinking abilities, and, no less significantly, teaching me genetics.

Finally, this work would have been impossible without my friends and family. Thank you to Doug & Jill Dunn, Brooks Dunn, Gillian & Chris Tucker, and their children Griffin, Jane, and Anderson. Thank you to Phoebe Kuo, Maria Pacana, & Vera Yin for adventure; to Emily Berry & Amanda Shareghi for mischief; to David Olem & Michael Torres for their pragmantic optimism; to Allison Master & Boting Zhang for trails of Post-It notes; to Theresa Berens, Anna Payne-Tobin Jost, Anna Reade, Karmela Ramos, & Kelly Nissen for their intuition and resolve; to Vida Ahyong & Natalie Petek for their inspirational grit; to Joel Street, James Kraemer, Bob Masys, Jason Damas, Jeff Farrell, Powen Shiah, Ben Schiller, Chris Mullendore, & Fabian Bündchen for showing up; to Gunta Kaza & David Arond for noting that the present is in fact actually happening right now; to Jessica McGrath, Roberto Mastroianni, Danielle Mishkin, Avril DePagter, Amanda Brenemen, Susan Brady, & Rachel Cunningham for keeping a space in Boston warm for me; to Nick Candau, Peter Cheeseman, & James Paulas for doing so in San Francisco; and to entire the Tetrad class of 2008 for seven and more years of shenanigans.

This work previously appeared as: Dunn, J. G., Foo, C. K., Belletier, N. G., Gavis, E. R., & Weissman, J. S. (2013). *Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster.* eLife, 2(0), E01179

Abstract

Ribosomes can read through stop codons in a regulated manner, elongating rather than terminating the nascent peptide. Stop codon readthrough is essential to diverse viruses, and phylogenetically predicted to occur in a few hundred genes in *Drosophila melanogaster,* but the importance of regulated readthrough in eukaryotes remains largely unexplored. Here, we present a ribosome profiling assay (deep sequencing of ribosome-protected mRNA fragments) for *Drosophila melanogaster,* and provide the first genome-wide experimental analysis of readthrough. Readthrough is far more pervasive than expected: the vast majority of readthrough events evolved within *D. melanogaster* and were not predicted phylogenetically. The resulting C-terminal protein extensions show evidence of selection, contain functional subcellular localization signals, and their readthrough is regulated, arguing for their importance. We further demonstrate that readthrough occurs in yeast and humans. Readthrough thus provides general mechanisms both to regulate gene expression and function, and to add plasticity to the proteome during evolution.

Contents

List of tables

List of figures

Stop codon readthrough is a feature of eukaryotic translation

Introduction

Upon encountering a stop codon, ribosomes can terminate translation with remarkable fidelity, yet they do not always do so. Stop codon readthrough, the decoding of a stop codon as a sense codon by the ribosome, plays important regulatory roles. Most immediately, readthrough diversifies the proteome by creating a pool of C-terminally extended proteins. In this capacity, it is essential to a variety of plant and animal viruses (Cimino et al., 2011; Li & Rice, 1989; Napthine et al., 2012; Skuzeski et al., 1991; Yoshinaka et al., 1985; reviewed in Beier & Grimm, 2001 and Firth & Brierley, 2012). In eukaryotic host genes, readthrough is functionally important insofar as it may suppress pathological phenotypes caused by premature stop codons (Fearon et al., 1994; Kopczynski et al., 1992), antagonize nonsense-mediated decay (Keeling et al., 2004), and, by changing the C-terminal sequence of a given protein, modulate its activity (Torabi & Kruglyak, 2012), stability (Namy et al., 2002), and/or localization (Freitag et al., 2012). In yeast, the efficiency of translation termination is modulated by *[PSI$^+$]*, an epigenetic state resulting from prion-like aggregates of Sup35p, the yeast homologue of the translation termination factor eRF3 (reviewed in Tuite & Cox, 2007). Various yeast strains exhibit *[PSI$^+$]*-dependent fitness advantages, implying that increased readthrough activates useful genetic diversity that is ordinarily masked by stop codons (Halfmann et al., 2012; True & Lindquist, 2000). In addition, a small baseline level of readthrough appears to be beneficial in wild *[psi*$^{-}$*]* yeast strains, as alleles of various factors controlling termination efficiency are under balancing selection (Torabi & Kruglyak, 2011).

However, a broad understanding of the biological roles of readthrough in eukaryotes remains elusive due to a lack of experimental data. To date, only a handful of eukaryotic host genes have been experimentally demonstrated to undergo readthrough in wild-type or prion-free organisms (Freitag et al., 2012; Geller & Rich, 1980; Jungreis et al., 2011; Klagges et al., 1996; Namy et al., 2002; Steneberg et al., 1998; Torabi & Kruglyak, 2012; Xue & Cooley, 1993; Yamaguchi et al., 2012). Compelling evidence that readthrough is broadly important in eukaryotes came with the development of algorithms (CSF and PhyloCSF) that use orthologous nucleotide sequences from related organisms to identify protein-coding regions of a reference genome based upon signatures of amino acid conservation (Lin et al., 2007; Lin et al., 2011). Using this approach, Lin and colleagues predicted 283 readthrough events in *Drosophila melanogaster*, six of which they confirmed experimentally (Jungreis et al., 2011; Lin et al., 2007). While these algorithms provide a powerful means to identify ancient and phylogenetically conserved readthrough events, they are limited in their ability to detect evolutionarily recent events. Nor can bioinformatic approaches identify a priori the tissues or cell types in which readthrough occurs, measure the fraction of ribosomes that read through a given stop codon, or determine whether any of these processes are regulated: such questions demand experimental approaches.

To this end, we present a modified ribosome profiling protocol — based on the deep sequencing of ribosome-protected footprint fragments (Ingolia et al., 2009) — that enables analysis of translation at a genome-wide level in *Drosophila melanogaster*. Application of the *Drosophila* ribosome profiling strategy allows annotation of the *Drosophila* proteome using empirical data. By examining the physical locations of ribosomes along mRNAs, we discover that readthrough is far more pervasive than expected: we identify more than three hundred readthrough events not predicted by phylogenetic approaches. We provide evidence that these novel extensions are of recent evolutionary origin, and show using specific examples that both the novel and conserved extensions can produce stable protein products, be produced in a regulated manner, and contain functional subcellular localization signals. We further demonstrate that readthrough occurs at many loci in *[psi⁻]* yeast and in primary human foreskin fibroblasts, arguing that readthrough is both a ubiquitous feature of eukaryotic translation and a novel mechanism to regulate gene expression. Stop codon readthrough thus adds plasticity to the proteome during development, and provides an evolutionary mechanism for extant genes to acquire new functions.

Results

Development of a ribosome profiling assay for cultured Drosophila cells

In order to study translation and, more specifically, stop codon readthrough in *Drosophila* melanogaster, we sought to develop a robust ribosome profiling assay for this organism. We initially developed our protocol in S2 cells, a macrophage-like lineage derived from late-stage *Drosophila* embryos.

In previous studies, ribosome-protected fragments or *footprints* were generated by digesting eukaryotic polysome lysates with RNase I (Ingolia et al., 2009; Ingolia et al., 2011). In contrast to yeast and mammalian cell lines, we found that *Drosophila* ribosomes are highly sensitive to RNase I, potentially due to their unusual rRNA sequences and structures (*Fig. 1*, supplement *1A*; Hancock et al., 1988; Jordan, 1975; Jordan et al., 1976; Pavlakis et al., 1979). By contrast, we found that *Drosophila* ribosomes tolerate micrococcal nuclease (MNase) over a wide range of concentrations (*Fig. 1*, supplement *1B–D*). In contrast to RNase I, MNase has a strong 3′ A/T bias. This gives rise to a small amount of positional uncertainty with P-site mapping in MNase datasets, and prevents us from achieving the sort of sub-codon resolution seen in ribosome profiling datasets generated with RNase I.

Nonetheless, replicate experiments established that our measure of translation rate (the ribosome footprint density, defined as the number of ribosome-protected fragments per kilobase

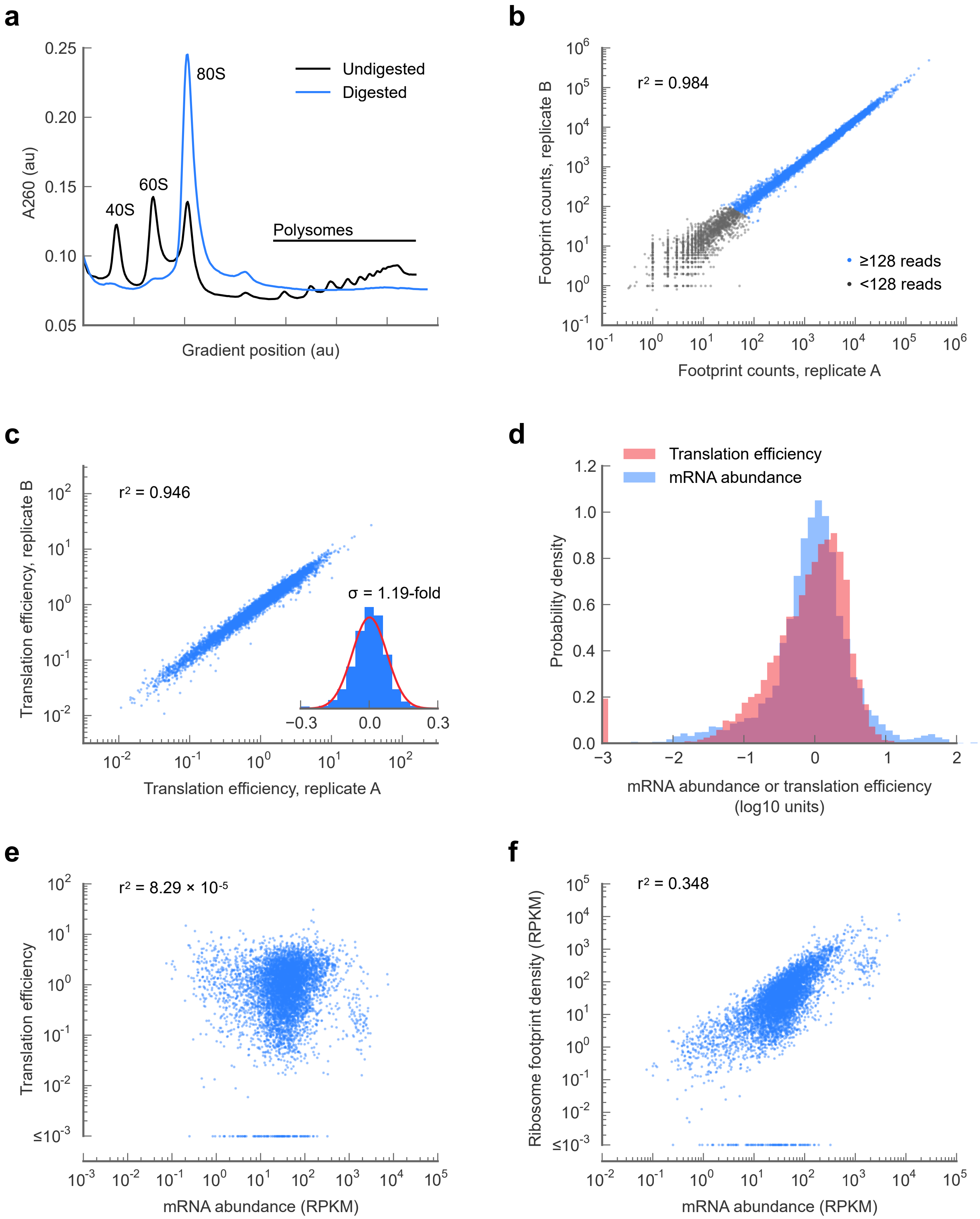


Figure 1: Development and validation of a ribosome profiling assay for *Drosophila melanogaster*

a. Aliquots of polysome lysate from 0–2 hr embryos were fractionated on 10–50% sucrose gradients with or without prior micrococcal nuclease digestion. Digestion of exposed mRNA between ribosomes collapses the polysome peaks into the monosomal (80S) peak. The of coding region per million aligning reads in the dataset; RPKM), is highly reproducible and insensitive to changes in buffer conditions (*Fig. 1*, supplement 1E; *Fig. 1*, supplement *2A & B*; full data in Supplemental table 1 at Dryad: Dunn et al., 2013). Focusing on coding regions that had a minimum of 128 reads, we observed strong correlation between replicates (r^2 = 0.998; *Fig. 1*, supplement *2*) and an inter-replicate standard deviation of 1.07-fold, comparable to our protocols in yeast and mammalian cells. Furthermore, our measurements are robust to the number of isoforms per gene, the fraction of sequence-degenerate positions in a gene, gene length, A/T content, and distribution of ribosome density within a gene (*Fig. 1*, supplement *3*).

Development of a ribosome profiling protocol for Drosophila embryos

In early (0–2 hour) *Drosophila* embryos, the vast majority of transcripts are maternally supplied and therefore regulated by post transcriptional processes, such as poly- or deadenylation, capping or de-capping, localization, degradation, and control of translation initiation. The early *Drosophila* embryo has thus been an important system for the study of post-transcriptional and specifically translational regulation (reviewed in Lasko, 2011).

To enable the broad analysis of these processes, we developed a sample harvesting strategy that captures the translational state of early embryos with minimal perturbation. Specifically, we developed a cryolysis protocol in which embryos are collected directly from egg-laying dishes into liquid nitrogen, homogenized while frozen, and thawed in the presence of translation inhibitors to prevent post-lysis translation. Notably, we omit dechorionation and rinsing, steps which could induce cold shock, anoxia, and related translational artifacts.

area under the monosome peak in the digested sample is 1.04-fold the combined area under the monosome and polysome peaks in the undigested sample, indicating quantitative recovery. **b and c**. Measurements of translation are reproducible between replicates samples of 0–2 hr embryos. Pearson correlation coefficients (r^2) are shown for total ribosome-protected footprint counts in coding regions for all genes sharing at least 128 summed footprint counts between replicates (b), or translation efficiency measurements for all genes sharing 128 summed mRNA fragment counts between replicates (c). Histogram of log10 fold-changes in translational efficiency for each gene between two embryo replicates, along with normal error curve (c, inset). **d, e, & f**. Pooled data for genes containing at least 128 summed mRNA counts between both embryo replicates. Median-centered histograms of translation efficiency (pink) and mRNA abundance (blue) (d). Translational efficiency versus mRNA abundance for each gene (e). Ribosome density versus mRNA abundance for each gene (f). *Source data in supplementary table 1*

We collected replicate samples of 0–2 hour embryos, and subjected them to ribosome profiling and RNA-seq of poly(A)-selected mRNA. A subset of ribosomes partition into heavy polysomes (*Fig. 1A*), consistent with reports that a distinct subset of messages is well-translated at this stage (Qin et al., 2007). Ribosome density measurements from replicate embryo collections are correlated nearly as well ($r^2 = 0.984$; *Fig. 1B*; Supplemental table 1 at Dryad: Dunn et al., 2013) as measurements from technical replicates from a single culture of S2 cells ($r^2 = 0.998$; *Fig. 1*, supplement *1E*). The *Drosophila* embryo thus provides a system in which experimental noise approaches the precision of our measurements, a fact that will facilitate detection of even small expression differences between wild-type and mutant fly strains.

Translational control is measured by a gene's translation efficiency, estimated as the ratio of ribosome footprint density (from ribosome profiling) to mRNA abundance (from mRNA-seq) for each gene. Translation efficiency measurements between replicate embryo collections are highly reproducible ($r^2 = 0.946$; *Fig. 1C*) and consistent with prior measurements made by semiquantitative methods (*Fig. 1*, supplement *4*). The standard deviation of fold-changes between biological replicates is 1.19-fold (*Fig. 1C*, inset), allowing detection of even modest changes in translation efficiency.

Remarkably, we find that the range of translation efficiencies for different messages spans four orders of magnitude, a range comparable to that observed for mRNA abundance of well-counted genes (*Fig. 1D*). Moreover, translation efficiency is uncorrelated with mRNA abundance ($r^2 = 8.29 \times 10^{-5}$; *Fig. 1E*) and mRNA abundance predicts only one third of the variance in the rate of protein production as measured by ribosome footprint density (*Fig. 1F*). Translational regulation is therefore a major determinant of gene expression in the early embryo (Supplemental table 1 at Dryad: Dunn et al., 2013), and ribosome profiling provides a quantitative and robust means to monitor translational regulation during development.

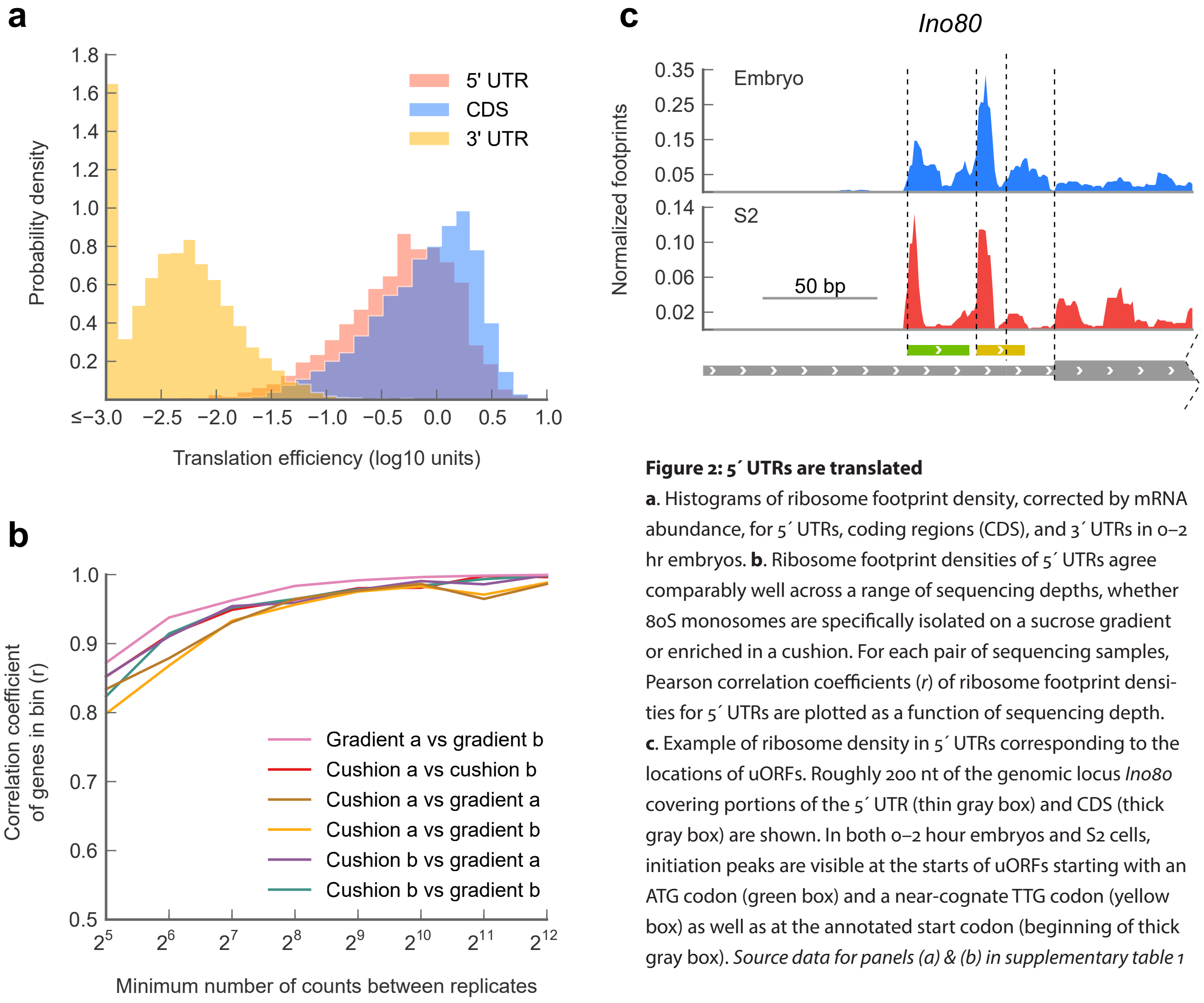
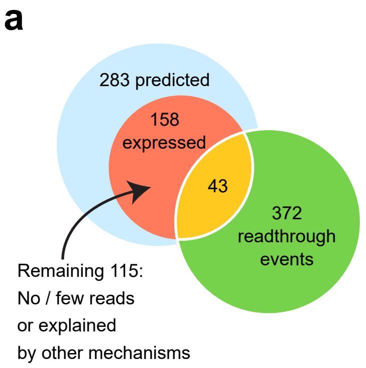


Figure 2: 5´ UTRs are translated
a. Histograms of ribosome footprint density, corrected by mRNA abundance, for 5´ UTRs, coding regions (CDS), and 3´ UTRs in 0–2 hr embryos. **b**. Ribosome footprint densities of 5´ UTRs agree comparably well across a range of sequencing depths, whether 80S monosomes are specifically isolated on a sucrose gradient or enriched in a cushion. For each pair of sequencing samples, Pearson correlation coefficients (*r*) of ribosome footprint densities for 5´ UTRs are plotted as a function of sequencing depth. **c**. Example of ribosome density in 5´ UTRs corresponding to the locations of uORFs. Roughly 200 nt of the genomic locus *Ino80* covering portions of the 5´ UTR (thin gray box) and CDS (thick gray box) are shown. In both 0–2 hour embryos and S2 cells, initiation peaks are visible at the starts of uORFs starting with an ATG codon (green box) and a near-cognate TTG codon (yellow box) as well as at the annotated start codon (beginning of thick gray box). *Source data for panels (a) & (b) in supplementary table 1*

Ribosome density on 5´ UTRs is similar to that of coding regions

In addition to measuring gene expression, ribosome profiling maps the physical positions of ribosomes on each transcript, and thus provides a powerful tool to annotate which portions of mRNAs are translated. Consistent with our previous work in mammals (Ingolia et al., 2011) and yeast (Brar et al., 2012; Ingolia et al., 2009), many 5´ UTRs in *Drosophila* contain substantial footprint density (*Fig. 2A*, *Fig. 2*, supplement *1*; Supplemental file *1A*) covering sequences that appear to be upstream open reading frames (uORFs; *Fig. 2C*).

We attribute this density to translating 80S ribosomes rather than 48S preinitiation complexes for three reasons: first, the length distribution of protected fragments in 5´ UTRs



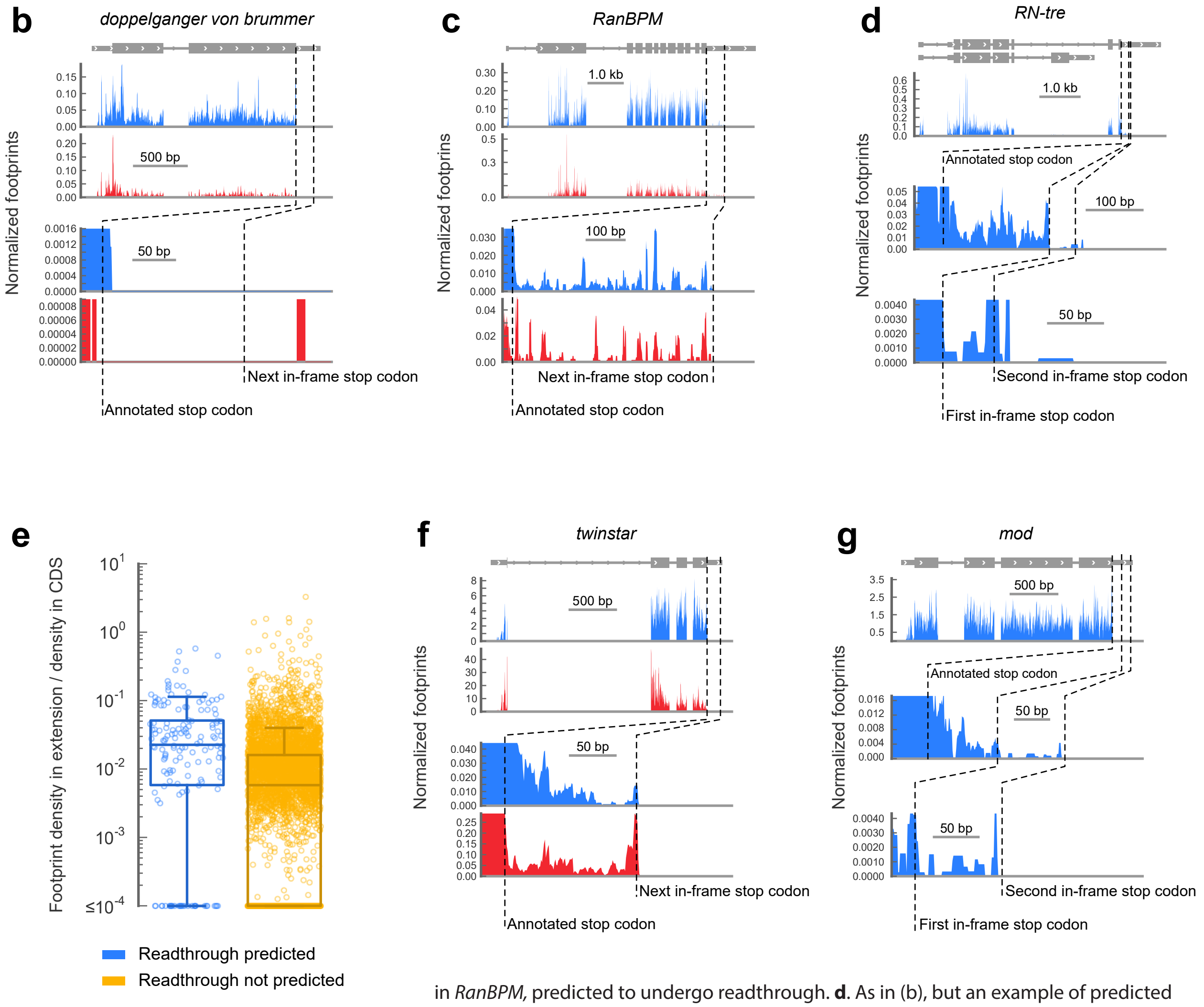


Figure 3: A subset of genes exhibit apparent stop codon readthrough
a. Venn diagram summarizing readthrough events. Of 283 predicted extensions, 256 were consistent with FlyBase annotation r5.43. For 158 of these, the corresponding coding regions were expressed in 0–2 hour embryos. Of these, 43 exhibited clear signs of readthrough. Others were ambiguous, untranslated, or could be explained by other mechanisms (see Fig. 3, supplement 1). We identified 307 additional examples of readthrough that were not not phylogenetically predicted. **b**. A gene that does not exhibit readthrough. *Top:* genomic locus with UTRs (thin boxes), introns (line), and coding regions (thick boxes). *Middle:* normalized footprint density covering the locus in 0–2h embryos (blue) and S2 cells (red) in reads per million. *Bottom:* magnification of region where a putative C-terminal extension would be found. *Dashed lines*: annotated and next in-frame stop codons. **c**. As in (b), except stop codon readthrough creates a C-terminal protein extension in *RanBPM,* predicted to undergo readthrough. **d**. As in (b), but an example of predicted double-readthrough. **e**. Ratios of the ribosome footprint density in putative extensions to corresponding coding regions. *Blue:* extensions predicted to undergo readthrough. *Yellow:* all other possible extensions. Extensions that overlapped any annotated CDS, snoRNA, or snRNA were excluded. *Boxes:* IQR. *Whiskers:* 1.5 × IQR. **f**. As in (c), except this transcript was not predicted to undergo readthrough. **g**. As in (d), except not predicted to undergo readthrough. *Source data in supplementary table 2 (at Dryad: Dunn et al., 2013).*

(25–35 nt) is indistinguishable from the length distribution of ribosome-protected fragments in coding regions (*Fig. 2*, supplement *2*), while the protected footprint of a preinitiation complex is reported to be larger (40–70 nt; Lazarowitz & Robertson, 1977; Pisarev et al., 2008). Second, our measurements of 5′ UTR density are indistinguishable whether we enrich digested monosomes by sedimentation through a sucrose cushion (which collects all heavy particles) or specifically separate them from preinitiation complexes by fractionation of a sucrose gradient (*Fig. 2B* and *Fig. 2*, supplement *3*). Thus, the dominant signal contributing to our measurement of footprint density in 5′ UTRs is derived from fragments protected by 80S ribosomes. Third, because initiation and termination of translation are slow compared to elongation, initiation and termination events produce peaks of ribosome density (*Fig. 2*, supplement *1*; Ingolia et al., 2011). Such peaks are frequently visible at the boundaries of predicted uORF sequences (example in *Fig. 2C*), again arguing that reads aligning to 5′ UTRs represent translation events. Given the known roles of uORFs in regulating both the translation and the stability of mRNAs (reviewed in Meijer & Thomas, 2002) we anticipate that our methods will facilitate future analyses of the contributions of uORFs to control of gene expression throughout fly development.

A subset of genes exhibit stop codon readthrough, resulting in C-terminal protein extensions

Comparative analysis of the genomes of twelve sequenced *Drosophila* species has provided a powerful strategy for annotating protein-coding regions in *Drosophila* genomes (Lin et al., 2007; Lin et al., 2011). Using this approach, 283 transcripts in *Drosophila melanogaster* were demonstrated to contain clear phylogenetic signatures of amino acid conservation in the region between the annotated and next in-frame stop codons. It was therefore concluded that these regions encode C-terminal protein extensions (hereon called *predicted extensions*), produced by stop codon readthrough events (Jungreis et al., 2011; Lin et al., 2007).

In our data, the density of ribosomes on 3′ UTRs is several orders of magnitude lower than in coding regions and 5′ UTRs (*Fig. 2A*, Supplemental file *1A*), and many genes show highly efficient termination (example in *Fig. 3B*). However, a subset of transcripts exhibit high footprint density within the predicted extensions. To determine whether the footprint density was consistent with stop codon readthrough (as opposed to alternate explanations, like frameshift), we manually scored each predicted extension whose corresponding structural gene was sufficiently expressed in our embryo sample (158 in total). An extension was scored positively if there existed ribosome density in the extension, ribosome density vanished or unambiguously decreased following the first in-frame stop codon, and positions occupied by ribosomes in the putative extension evenly covered the majority of the extension's length (see methods for further details). By these criteria, 43 of the 283 transcripts predicted to undergo stop codon readthrough contained ribosome density consistent with a readthrough event (example *Fig. 3C*, full data in Supplemental table 2 at Dryad: Dunn et al., 2013), including one example of double readthrough (*Fig. 3D*). We expect that the many of the remaining 240 transcripts also undergo readthrough, either at levels too low to detect at our sequencing depth, or at other developmental stages (discussed further below).

Surprisingly, we observed that a distinct set of transcripts not predicted to undergo readthrough also exhibits substantial footprint density between the annotated and next in-frame stop codons (*Fig. 3E*). We therefore searched for C-terminal extensions among all transcripts that met the following criteria: (a) a minimum of 128 footprint in the corresponding CDS, (b) a minimum footprint read density of 0.2 RPKM in the extension, (c) a minimum readthrough rate of 0.001, and (d) a lack of methionine codons in the first three codons of the extension, as this latter group could be explained by initiation within the extension rather than readthrough of the upstream stop codon. We additionally excluded extensions whose translation could be explained by alternately spliced transcript isoforms that omit the stop codon. Scoring this group by the same criteria used for the predicted extensions, we identified 307 additional examples of stop codon readthrough (hereon referred to as *novel extensions;* see example in *Fig. 3F*), including another example of double readthrough (*Fig. 3G*). In addition, we identified several transcripts that contained 3´ UTR footprint density more consistent with ribosomal frameshift (*Fig. 3*, supplement *1A, B*), or the presence of additional downstream cistrons, RNA structure, or protein binding (*Fig.*

3, supplement *1C, D*). These were excluded from further analysis.

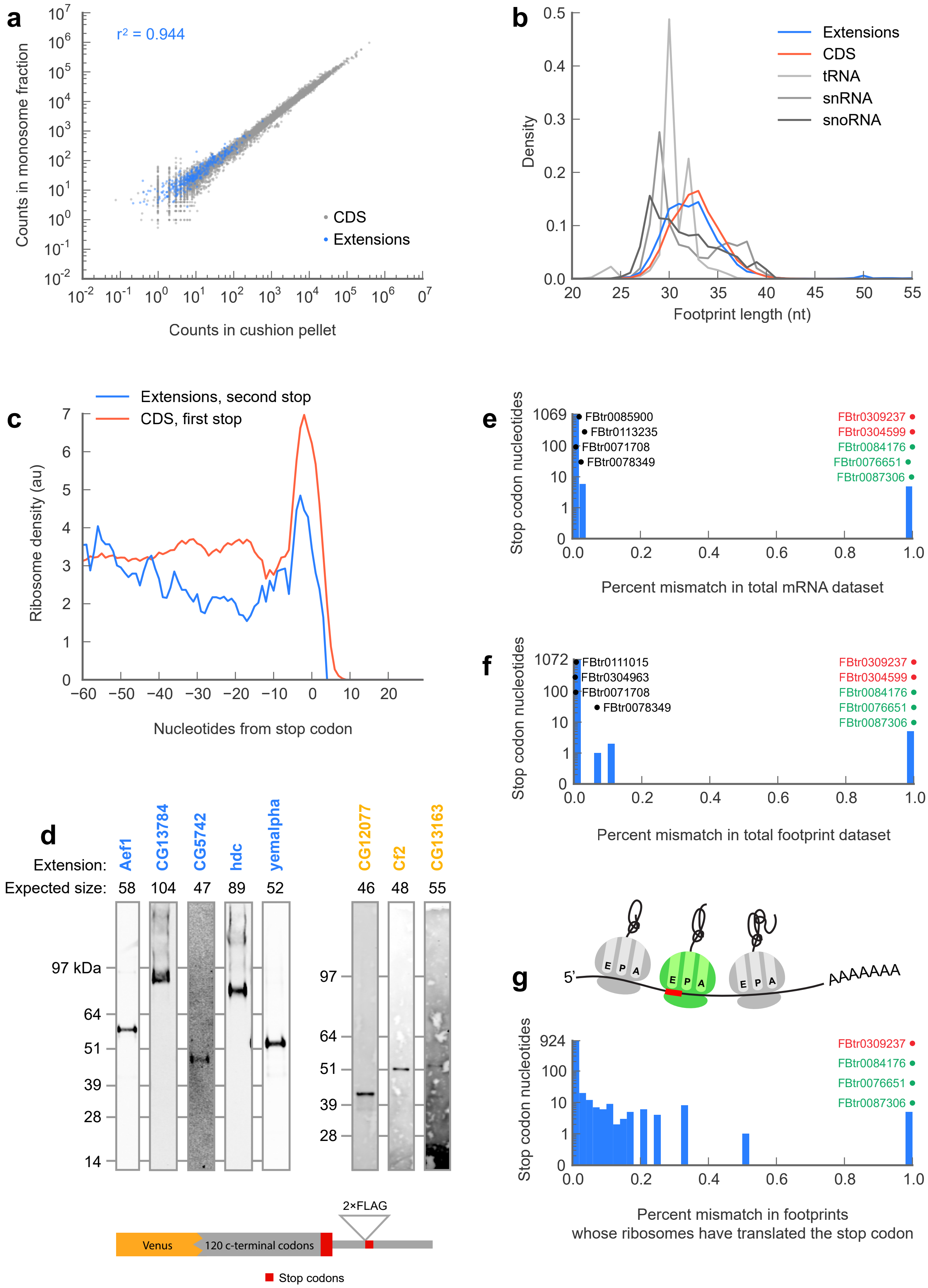
Ribosome-protected footprints in C-terminal extensions show signatures of translation

Because footprint density generally is far lower in 3′ UTRs than in 5′ UTRs or coding regions (*Fig. 2A*), it is possible that various sources of noise (e.g. regions of mRNA protected by RNA structures or by RNA-binding proteins) might contribute more substantially to this density than to the density in coding regions. We therefore asked whether footprints in 3′ UTRs exhibited behaviors specific to footprints protected by 80S ribosomes.

In order to distinguish whether reads mapping to extensions were either protected by ribosomes or derived from alternate sources, we compared the total number of reads aligning to extensions in samples prepared from sucrose cushions, which collect all heavy macromolecular complexes, to those in which we specifically isolated 80S ribosomes on sucrose gradients. Footprint count measurements for each extension are highly correlated between libraries made using these two sample preparation methods, indicating that these footprints are either protected by 80S ribosomes, or by another RNA binding protein that co-sediments with 80S ribosomes (*Fig. 4A*; $r^2 = 0.945$).

Because various ribosome-binding proteins protect nucleotide fragments of distinct lengths, the size distribution of protected mRNA fragments provides a powerful approach for distinguishing 80S footprints from other sources (Ingolia et al., manuscript in preparation). Footprints in C-terminal extensions exhibit a length distribution very similar to footprints in coding regions, while those derived from non-coding sources, such as snoRNAs and tRNAs, show dramatically different length distributions (*Fig. 4B*). Thus, footprints aligning to extensions appear to be protected by 80S ribosomes.

Finally, we sought to determine whether the ribosomes that appear to translate extensions are engaged in active translation, as opposed to some aberrant process of stalling or slippage (e.g. as described in Skabkin et al., 2013). Because terminating ribosomes produce a characteristic peak of ribosome density over annotated stop codons (*Fig. 2*, supplement *1*; Ingolia et al., 2009),


a
r² = 0.944
Counts in monosome fraction
Counts in cushion pellet
10^{7} 10^{6} 10^{5} 10^{4} 10^{3} 10^{2} 10^{1} 10^{0} 10^{-1} 10^{-2}
10^{-2} 10^{-1} 10^{0} 10^{1} 10^{2} 10^{3} 10^{4} 10^{5} 10^{6} 0^{7}
• CDS
• Extensions
b
Density
0.5 0.4 0.3 0.2 0.1 0.0
20 25 30 35 40 45 50 55
Footprint length (nt)
Extensions
CDS
tRNA
snRNA
snoRNA
c
Extensions, second stop
CDS, first stop
Ribosome density (au)
7 6 5 4 3 2 1 0
−60 −50 −40 −30 −20 −10 0 10 20
Nucleotides from stop codon
d
Extension:
Expected size:
Aef1 58
CG13784 104
CG5742 47
hdc 89
yemalpha 52
CG12077 46
Cf2 48
CG13163 55
97 kDa
64
51
39
28
14
97
64
51
39
28
Venus
120 c-terminal codons
2×FLAG
Stop codons
e
Stop codon nucleotides
1069 100 10 1 0
FBtr0085900
FBtr0113235
FBtr0071708
FBtr0078349
FBtr0309237
FBtr0304599
FBtr0084176
FBtr0076651
FBtr0087306
0.0 0.2 0.4 0.6 0.8 1.0
Percent mismatch in total mRNA dataset
f
Stop codon nucleotides
1072 100 10 1 0
FBtr0111015
FBtr0304963
FBtr0071708
FBtr0078349
FBtr0309237
FBtr0304599
FBtr0084176
FBtr0076651
FBtr0087306
0.0 0.2 0.4 0.6 0.8 1.0
Percent mismatch in total footprint dataset
g
5’
E P A
E P A
E P A
AAAAAAA
Stop codon nucleotides
924 100 10 1 0
FBtr0309237
FBtr0084176
FBtr0076651
FBtr0087306
0.0 0.2 0.4 0.6 0.8 1.0
Percent mismatch in footprints
whose ribosomes have translated the stop codon

we asked whether the stop codons that terminate the C-terminal extensions also showed this behavior. Indeed, C-terminal extensions exhibit peaks at their stop codons, clearly arguing that footprint density in C-terminal extensions is attributable to actively-translating ribosomes (*Fig. 4C*).

Because this meta-gene analysis represents a group average, we also compiled individual statistics on ribosome release in a manner similar to the RRS score described by Guttman and colleagues (Guttman et al., 2013). Briefly, we tabulated the ratio of the total number of reads aligning within a five codon window immediately downstream of a stop codon to the number of reads aligning to the five codon window immediately upstream of that codon, with the expectation that if ribosomes terminate at a given stop codon, the score for that codon should approach zero. We performed this calculation separately for: (1) stop codons that terminate annotated coding regions, (2) stop codons that terminate C-terminal extensions, and (3) as a negative control, randomly selected codons internal to annotated coding regions. We find that the scores of stop codons that terminate C-terminal extensions fall within the distribution of scores for stop codons that terminate annotated coding regions (*Fig. 4*, supplement *1*), again arguing that the

Figure 4: Translation downstream of the stop codon is due to stop codon readthrough
a. Ribosome footprint counts for each C-terminal extension are well correlated between samples prepared by sedimentation through sucrose cushions or by fractionation on sucrose gradients (blue). For comparison, footprint counts for annotated coding regions in each sample type are plotted (gray). The Pearson correlation coefficient (r^2) for C-terminal extensions is shown. **b**. Distributions of read lengths for footprints aligning to annotated coding regions (CDS, red) and to C-terminal extensions (blue) are similar, while lengths of footprints aligning to tRNAs, snRNAs, and snoRNAs are quite different. **c**. Metagene average of ribosome density at the annotated stop codons of coding regions (red), or at the stop codons that terminate extensions (blue). Both averages show characteristic peaks of ribosome density above the stop codon, characteristic of translation termination. **d**. Readthrough produces detectable protein products. *Bottom:* schema of reporters. Reporters containing the GFP variant Venus fused to the 120 C-terminal codons and entire endogenous 3´ UTR of a gene of interested were transfected into S2 cells. To facilitate detection of readthrough products, a double-FLAG epitope was inserted upstream of the stop codon (red) that terminates the putative extension. *Top:* Reporters were immunoprecipitated with anti-GFP antibodies. Immunoprecipitates were then resolved by SDS-PAGE and western blotted with anti-FLAG antibodies to detect protein products of readthrough. *Blue:* names of genes containing extensions predicted to undergo readthrough. *Yellow:* names of genes containing novel extensions. **e**. For each nucleotide in each stop codon that undergoes readthrough, we counted the fraction of reads containing nucleotide mismatches and present the data as a histogram. Transcripts containing stop codon nucleotides with significantly elevated mismatch rates are explicitly noted. *Green:* transcripts containing genomic polymorphisms that mutate one stop codon to another. *Red:* transcripts containing genomic polymorphisms that convert stop codons to sense codons. Black: other transcripts containing significantly elevated proportions of mismatches. **f**. As in (e), but for ribosome-protected footprint data. **g**. As in (f), but the analysis was restricted to the subset of footprints that both include the sequence of the stop codon and derive from ribosomes that have already translated the stop codon (top, green ribosome in cartoon).

read density covering putative C-terminal extensions are in fact produced by ribosomes that have undergone stop codon readthrough rather than other processes.

Readthrough produces detectable translation products

It is possible that the population of ribosomes that read through stop codons is engaged in a pathological translation process that might not produce detectable protein products. We therefore asked whether we could detect translation products by immunoprecipitation (IP) and western blotting. We created reporter constructs for a panel of transcripts including five predicted and ten novel extensions that exhibited readthrough in both 0–2 hour embryos and S2 cells. In each construct, we fused Venus (a GFP variant) upstream of a portion of each transcript containing the C-terminal 120 codons of the annotated coding sequence and the entire endogenous 3′ UTR. To visualize readthrough, we fused a double FLAG epitope to the C-terminus of the putative C-terminal extension. We transfected these constructs into S2 cells, immunoprecipitated the reporter at the N-terminus using anti-GFP beads, and detected the extensions by western blotting using an anti-FLAG antibody. We detected readthrough products of the correct size for eight of the reporters, arguing that at least this subset of extensions yields C-terminally extended proteins *in vivo* (*Fig. 4D*).

While we did not seek to detect C-terminally extended proteins generated by endogenous transcripts (e.g. through mass spectrometry), we do believe our reporter constructs to be at least as faithful as those used in earlier literature, as we included substantially more nucleotide context (120 codons upstream of stop plus the entire endogenous 3′ UTR) than other groups screening through candidate genes to find readthrough signals (2–8 codons upstream and 3–15 codons downstream of the stop codon; Fearon et al., 1994; Harrell et al., 2002; Namy et al., 2002; Namy et al., 2003).

Extensions are not products of selenocysteine insertion, genomic polymorphisms, or mRNA editing

The appearance of stop codon readthrough, both in ribosome profiling data and in IP-westerns, could result from several other processes, such as selenocysteine insertion, genomic mutation of stop codons to sense codons, or the editing of stop codons in mRNAs. We consider each of these in turn.

UGA stop codons may be decoded by specialized translation machinery as the unconventional amino acid selenocysteine if the 3′ UTR contains a selenocysteine insertion (SECIS) element. However, UGA stop codons represent only 25% of the readthrough events we report, and none of these are annotated as selenoproteins in either FlyBase (McQuilton et al., 2012) or SelenoDB (Castellano et al., 2008). Furthermore, we were unable to detect SECIS elements in any of their 3′ UTRs using SeciSearch 2.19 (Kryukov et al., 2003). Thus, at most, even unannotated selenocysteine insertion events could only account for a small fraction of the readthrough events we report.

We also exclude the possibility that readthrough might result from genomic polymorphisms or RNA editing at the stop codon. Because both types of events would be represented in our data as mismatches between read alignments and the reference transcript sequence, we counted the total number of matching and mismatching reads covering each nucleotide position in each stop codon in our mRNA-seq and ribosome footprint datasets. For each dataset, we calculated a global average proportion of mismatching reads, and used a binomial test to identify stop codon nucleotides whose individual proportion of mismatches significantly deviated from the corresponding global average.

Together, the mRNA and footprint datasets identified a total of ten nucleotide positions whose mismatch rates significantly exceeded the average (*Figs. 4E & F*). Three positions contained genomic polymorphisms that changed one stop codon to another stop codon (*Figs. 4E & F*, green). Two contained genomic polymorphisms that converted the stop codon to a sense codon (*Figs. 4E & F,* red). These two transcripts were therefore excluded from further study. The remaining five positions contained a variety of mismatches each occurring at low frequency. These observations are inconsistent with the presence of a genomic polymorphism at those positions, which should cause a 50% or 100% frequency of a single mismatch, depending on whether the polymorphism is hetero- or homozygous (*Figs. 4E* & F, black).

An alternate explanation for a low but elevated proportion of mismatches is RNA editing, the conversion of one nucleotide to another in an mRNA. In *Drosophila*, the only mechanism known to edit mRNA is the deamination of adenine to inosine, which is converted to guanine by reverse transcriptase (Ramaswami et al., 2013). A-to-I editing thus appears in sequencing data as a preference for A-to-G transitions among mismatches. Of the five mismatching positions we could not ascribe to genomic polymorphisms, four contain thymine or guanine rather than adenine residues in the reference sequence, and therefore cannot be edited by this pathway. We therefore attribute these mismatches to sequencing error. The majority of mismatches at the single remaining position are transversions from adenine to thymine, similarly arguing that these mismatches are more likely due to sequencing error than to A-to-I editing.

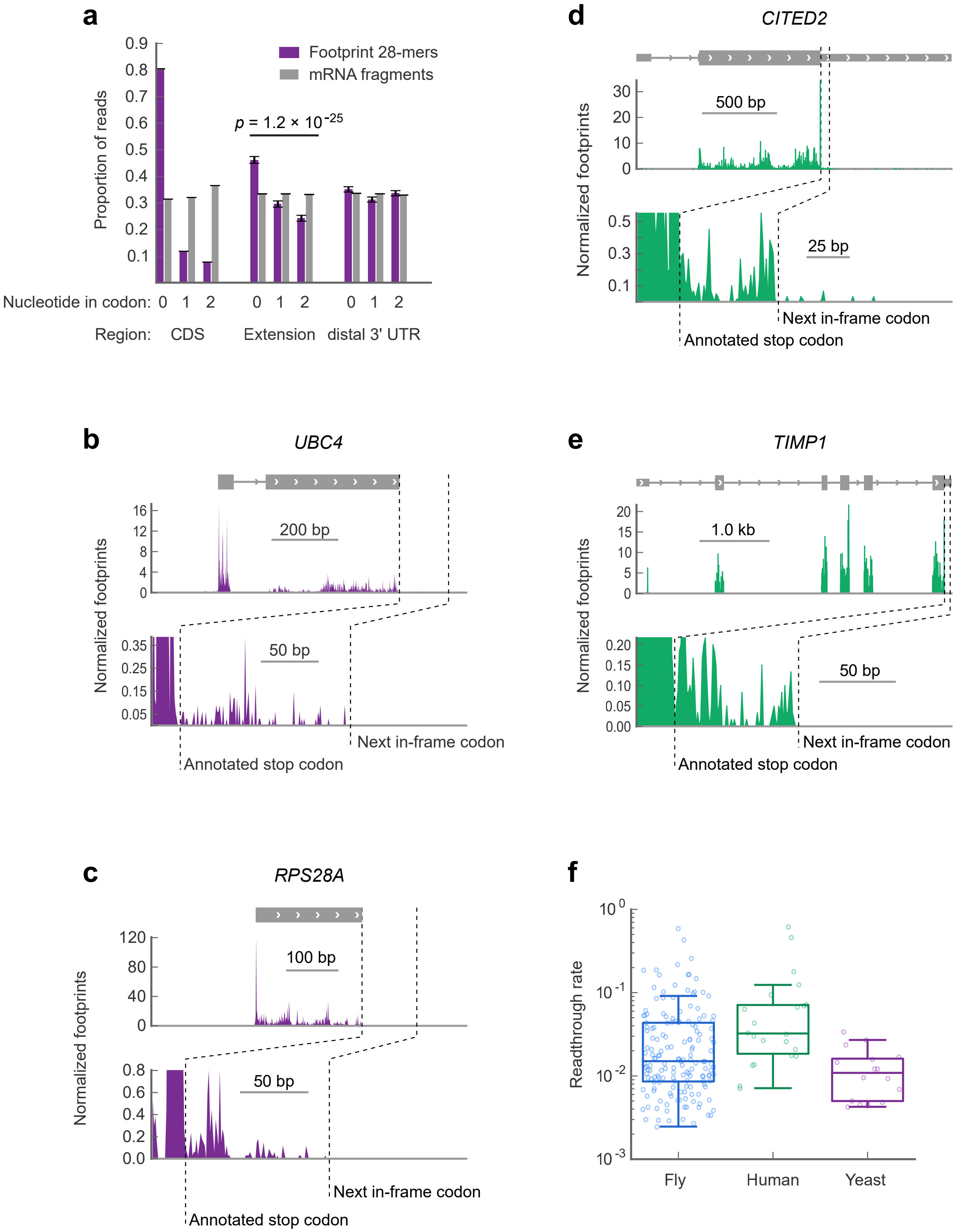
Formally, it is possible that a minor fraction of transcripts are edited, but that this fraction, even if small as measured in the RNA-seq or total footprint data, might account for all of the stop codon readthrough we observe. Analysis of the ribosome footprint data allows us to explore this possibility directly. Specifically, were this the case, the sequences of all the footprints deriving from ribosomes that have undergone readthrough — namely, those whose A-sites have already translated the stop codon — should contain evidence of editing (*Fig. 4G*, top). We therefore separately analyzed the footprints deriving from this specific pool of ribosomes. Our dataset provided sufficient coverage to test 419 of 450 such positions (93% of the total). Of these, only four stop codon positions exhibited significantly elevated levels of mismatch (*Fig. 4G*, bottom). All of these were identified in the mRNA and total footprint datasets above as having genomic polymorphisms (*Figs. 4E–F*). Thus, our most stringent dataset contains no positive evidence of RNA editing.

Further, this dataset contains positive evidence against RNA editing. Under the null hypothesis that A-to-I editing drives readthrough, one would expect nearly all footprints (for our purposes, conservatively assuming 90%) in the A-site footprint dataset to contain an edited base. Under this assumption, we used a binomial test to estimate the probability of observing the proportion of A-to-G mismatches in the A-site footprint dataset at each adenine residue sufficiently covered by reads (217 positions, representing roughly 50% of A positions in all readthrough events reported). In this analysis, all positions contained significantly fewer A-to-G mismatches than expected under the hypothesis of A-to-I editing, (Bonferonni-corrected P-value << 0.05 for all transcripts), indicating that A-to-I editing plays no part in any of the readthrough events we could test.

Readthrough occurs in Saccharomyces cerevisiae and Human foreskin fibroblasts

Because we detected far more readthrough events in *Drosophila* than were predicted from phylogenetic data, we collected yeast datasets and examined them for empirical evidence of readthrough. Importantly, because the *[PSI$^+$]* form of the yeast eRF3 homologue is known to promote readthrough, we limited our analysis to data collected from *[psi$^-$]* strains.

In contrast to MNase (which exhibits a 3′ A/T cutting bias, yielding positional uncertainty of the ribosomal P-site, see methods), RNase I show little cutting bias. Therefore, libraries prepared with RNase I (e.g. yeast and mammalian libraries) offer superior spatial resolution along mRNAs. In such libraries, the locations of ribosome-protected footprint fragments in coding regions exhibit a characteristic three-nucleotide periodicity or phasing from which reading frames can be deduced (Ingolia et al., 2009; Ingolia et al., 2011). We therefore tabulated the phasing of ribosome-protected footprint fragments in all annotated coding regions, putative C-terminal extensions, and the 40 codon windows downstream of the putative extensions as an approximation of the portion of the 3′ UTR distal to the putative extension (hereafter called *distal 3′ UTRs*). To control for cloning biases caused by skewed nucleotide frequencies at each phase, we tabulated the phasing of randomly-fragmented mRNA fragments that were cloned using the same protocol and aligned to the same regions. Non-random phasing consistent with translation is apparent in both the coding regions and the putative extensions, but not the distal 3′ UTR ($p = 3.98 \times 10^{-26}$, X^2 test, footprints versus mRNA fragments in extension, dof = 2; *Fig. 5A*). Importantly, the major component of phasing in the putative extensions occurs in the same reading frame as that of coding regions, indicating that readthrough (as opposed to, for example, frameshift)


a
Footprint 28-mers
mRNA fragments
Proportion of reads
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
$p = 1.2 \times 10^{-25}$
Nucleotide in codon: 0 1 2 0 1 2 0 1 2
Region: CDS Extension distal 3' UTR
b
UBC4
Normalized footprints
16
12
8
4
0
200 bp
0.35
0.25
0.15
0.05
50 bp
Next in-frame codon
Annotated stop codon
c
RPS28A
Normalized footprints
120
80
40
0
100 bp
0.8
0.6
0.4
0.2
0.0
50 bp
Next in-frame codon
Annotated stop codon
d
CITED2
Normalized footprints
30
20
10
0
500 bp
0.5
0.3
0.1
25 bp
Next in-frame codon
Annotated stop codon
e
TIMP1
Normalized footprints
20
15
10
5
0
1.0 kb
0.20
0.15
0.10
0.05
0.00
50 bp
Next in-frame codon
Annotated stop codon
f
Readthrough rate
10^{0}
10^{-1}
10^{-2}
10^{-3}
Fly
Human
Yeast

is a major contributor to protected fragment density in 3´ UTRs in yeast. Having found global evidence for readthrough, we manually scored a subset of yeast genes to identify individual examples of readthrough, using the same filtering and scoring criteria we used in the *Drosophila* datasets. We found 30 clear examples of readthrough in yeast (examples in *Figs. 5B* & C; full results in supplemental table *3* at Dryad: Dunn et al., 2013), demonstrating that readthrough is not unique to *Drosophila*.

Because readthrough has been observed in two mammalian genes (Geller & Rich, 1980; Yamaguchi et al., 2012), we collected data from primary human foreskin fibroblasts and sought evidence of readthrough in humans. We identified 42 readthrough events in the human data (*Figs. 5D, 5E*; full results in supplemental table *4*). These events are not explained by selenocysteine insertion, and, as in *Drosophila*, read lengths mapping to extensions in the yeast and human datasets are similar to those mapping to coding regions in these organisms (*Fig. 5*, supplement *1*). Thus, readthrough appears to be prevalent in all three organisms.

To estimate how many of the novel extensions we detected might be translated at a biologically significant level, we estimated a threshold for biological significance as the fifth percentile of readthrough rates for the phylogenetically conserved extensions that were translated in the *D. melanogaster* embryo, a rate of 1.2 percent. Out of all the extensions for which we could measure readthrough rates (i.e. those sufficiently long not to be covered by stop codon peaks, see methods), 61.8% of the novel extensions in *Drosophila*, 94.7% of the extensions in human foreskin fibroblasts, and 40.0% of the extensions in yeast exceeded this threshold, arguing that readthrough might be important in all three organisms (*Fig. 5F*).

Figure 5: Readthrough occurs at specific stop codons in *[psi-]* yeast and in human foreskin fibroblasts
a. Triplet periodicity of 28-mers from yeast data in all non-overlapping coding regions (CDS), putative C-terminal extensions, and distal 3´ UTRs indicates that a signature of translation readthrough is visible in extensions on a bulk scale. Distal 3´ UTRs were estimated as 40 codon windows following putative extensions. Putative extensions and distal 3´ UTRs that overlap annotated coding regions, snoRNAs, snRNAs, tRNAs or 5´ UTRs were excluded from the analysis. **b and c**. Examples of yeast transcripts that undergo readthrough, as in (*Fig. 3b*). **d and e**. Examples of transcripts that undergo readthrough in human foreskin fibroblasts, as in (*Fig 3b*). **f**. Distribution of readthrough rates, by organism, for all extensions of sufficient length not to be covered by bleedthrough from termination peaks (see methods). *Dashed line:* 5th percentile of readthrough rate in conserved extensions in *D. melanogaster,* 1.2%. *Source data in supplementary tables 2, 3, and 4*

Unpredicted C-terminal extensions show signs of recent evolutionary origin

Because 307 of the 350 readthrough events we discovered were not predicted phylogenetically, we sought to determine whether any of them showed signs of protein-coding conservation through the *Drosophila* phylogeny. To this end, we used PhyloCSF, which reports a log-likelihood ratio reflecting the relative probabilities of observing a given alignment of orthologous nucleotide sequences under models of protein-coding or non-coding evolution (Lin et al., 2011). By this metric, only 14 of the 307 novel extensions score positively (*Fig. 6A*), and their distribution of PhyloCSF scores was not markedly different from the global distribution (*Fig. 6*, supplement *1A*), indicating a lack of phylogenetic evidence for amino acid conservation.

The lack of detectable phylogenetic evidence of amino acid conservation among the novel extensions suggests two models: either (1) the novel extensions, on average, are selectively neutral, and occur only because they do not incur too great a fitness disadvantage, or (2) the novel extensions are under selection, but originated after the divergence of *D. melanogaster* from its closest sequenced relatives, making conservation in this group undetectable by cross-species tools such as PhyloCSF. To distinguish these possibilities, we used two tests to detect signs of selection for protein coding specifically within *D. melanogaster*.

To determine whether the nucleotide sequences of novel extensions show signs of selection for protein coding potential, we implemented a Z-curve classifier, a machine-learning technique that separates coding regions from non-coding regions based upon phased differences in nucleotide k-mer frequency (Gao & Zhang, 2004). We trained the classifier to distinguish annotated coding regions from distal 3′ UTRs (see methods for details). Consistent with a long history of protein-coding selection, extensions predicted by phylogenetic conservation showed a nucleotide character indistinguishable from annotated coding regions (*Fig. 6B*). By contrast, novel extensions exhibit a nucleotide character intermediate between coding regions and distal 3′ UTRs ($p = 1.02 \times 10^{-23}$, Mann-Whitney U test, distal 3′ UTR vs novel extensions), which is consistent with an evolutionary trajectory towards coding-like character from a 3′ UTR. This effect is not due

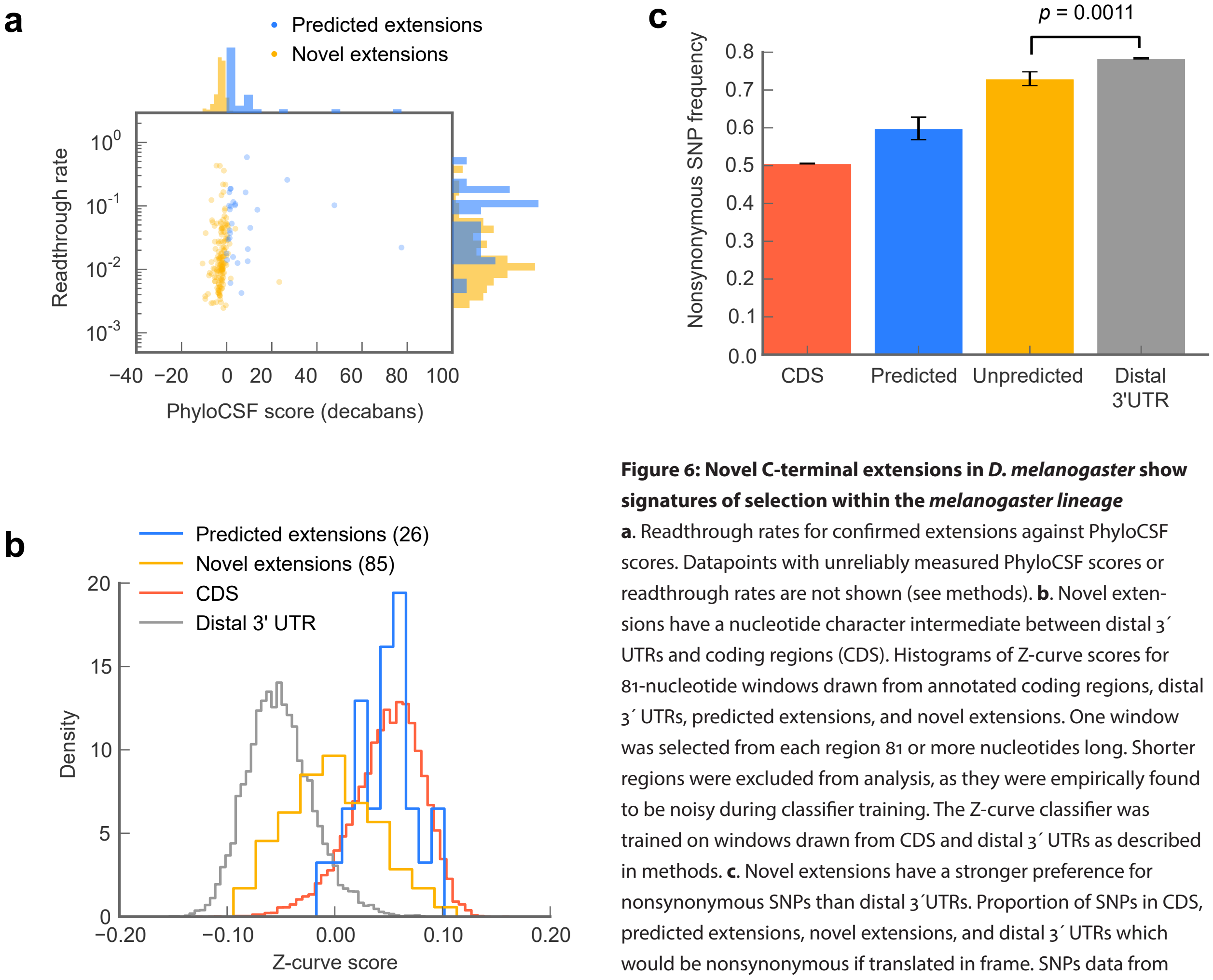


Figure 6: Novel C-terminal extensions in *D. melanogaster* show signatures of selection within the *melanogaster lineage*

a. Readthrough rates for confirmed extensions against PhyloCSF scores. Datapoints with unreliably measured PhyloCSF scores or readthrough rates are not shown (see methods). **b**. Novel extensions have a nucleotide character intermediate between distal 3´ UTRs and coding regions (CDS). Histograms of Z-curve scores for 81-nucleotide windows drawn from annotated coding regions, distal 3´ UTRs, predicted extensions, and novel extensions. One window was selected from each region 81 or more nucleotides long. Shorter regions were excluded from analysis, as they were empirically found to be noisy during classifier training. The Z-curve classifier was trained on windows drawn from CDS and distal 3´ UTRs as described in methods. **c**. Novel extensions have a stronger preference for nonsynonymous SNPs than distal 3´UTRs. Proportion of SNPs in CDS, predicted extensions, novel extensions, and distal 3´ UTRs which would be nonsynonymous if translated in frame. SNPs data from the Drosophila Population Genomics Project, downloaded from Ensembl. *Source data in supplementary table 2*

to specific nucleotide signals found in distal 3´ UTRs ($p = 3.81 \times 10^{-22}$, *Fig. 6*, supplement *1B*), and was robust across Z-curve classifiers trained on different windows drawn from distal 3´ UTRs (see methods).

To obtain more direct evidence for or against protein-coding selection, we analyzed SNP data from 50 individuals of *D. melanogaster* from the *Drosophila* Population Genomics Project (www.dpgp.org). We determined the proportion of SNPs that would be synonymous if translated in-frame in coding regions, predicted extensions, novel extensions, and distal 3´ UTRs. Novel extensions show a modest but significant preference for synonymous SNPs above the background level of distal 3′ UTRs (*Fig. 6C*; $p = 1.42 \times 10^{-5}$, one-sided Fisher's exact test), but below that of the predicted extensions ($p = 8.42 \times 10^{-9}$, one-sided Fisher's exact test). This pattern suggests that a subset of the novel extensions is undergoing selection for protein coding, and that the contribution from this subset to the average SNP preference outweighs the contributions from other subsets of extensions that are selectively neutral or undergoing diversifying selection. Together, these results favor the hypothesis that at least a fraction of the novel extensions are of recent evolutionary origin and have come under selection within the melanogaster lineage.

Readthrough is regulated individually for specific transcripts

In order to determine whether C-terminal extensions might be functional, we sought evidence for biological regulation of readthrough rates. We therefore queried our S2 cell and embryo datasets for evidence of differential regulation of readthrough in all genes that were sufficiently expressed in both datasets and contain only one, unique annotated coding region across all transcripts.

For each gene meeting these criteria, we tabulated the number of ribosome-protected footprints in the corresponding coding region and extension in each tissue type, and calculated a p-value for the observed distribution of counts using Fisher's exact test. Controlling the false discovery rate at 5%, we found 9 of 182 testable transcripts to significantly change between samples, indicating that all 9 should be true positives (table 1; full data in Supplemental table 2 at Dryad: Dunn et al., 2013). Thus, readthrough is differentially regulated between *Drosophila* cell types.

In principle, readthrough could be regulated by: (1) changes in the expression or activities of global factors (e.g. eukaryotic release factors, charged tRNA abundance et c), (2) by gene- or transcript-specific elements, like mRNA structures, or (3) by a combination of both. In the first scenario, readthrough rates for all transcripts should increase or decrease monotonically in one cell or tissue type compared to another. In the latter two scenarios, readthrough rates should increase for some transcripts, but decrease for others. We identified 4 significant increases and 5 significant decreases in readthrough rate in embryos compared to S2 cells, indicating that read-

		Readthrough rate					
Gene ID	Alias	0 –2h Embryo	S2 cell	PhyloCSF	P-value	$\log_{10}$ fold change	Direction
FBgn0036824	CG3902	7.15e-01	2.46e-03	-241.07	6.55e-10	-2.46	↓
FBgn0004362	HmgD	8.82e-03	1.21e-02	-747.85	7.08e-07	0.14	↑
FBgn0035432	ZnT63C	7.17e-03	2.71e-02	181.26	1.14e-06	0.58	↑
FBgn0010409	RpL18A	1.39e-02	2.08e-03	-197.78	5.85e-06	-0.83	↓
FBgn0039218	Rpb10	5.18e-03	2.03e-02	-333.38	8.06e-06	0.59	↑
FBgn0038100	Paip2	2.10e-02	4.60e-03	-497.09	3.71e-05	-0.66	↓
FBgn0261790	SmE	7.55e-03	7.80e-04	-530.28	9.60e-04	-0.99	↓
FBgn0030991	CG7453	2.18e-01	5.28e-02	-164.36	2.00e-03	-0.62	↓
FBgn0043796	CG12219	2.85e-01	1.90	-27.83	2.11e-03	0.82	↑

Table 1. Readthrough is differentially regulated between 0–2 hr embryos and S2 cells
For each transcript, the number of reads aligning to the CDS and corresponding extension were tabulated in both embryo and S2 cell datasets. p-values for significant changes were calculated using Fisher's Exact Test. The false discovery rate was controlled at 5% using the procedures of Benjamini and Hochberg (see methods), yielding 9 transcripts with significant p-values.

through is at least in part regulated on a transcript-by-transcript basis. The distribution of fold-changes in readthrough rate spans several orders of magnitude, indicating that transcripts that are robustly read through in one cell type are not necessarily read through in another (table 1). This result implies that extensions function in specific cellular or developmental contexts, consistent with earlier reports that readthrough of specific genes is regulated in metazoans (Robinson & Cooley, 1997; Yamaguchi et al., 2012).

Because we observe such a large magnitude of regulation, we believe the 350 readthrough events we report here to represent a small subset of a larger group that occur throughout the lifetime of an individual fly. We therefore expect many of the extensions that were phylogenetically predicted but not observed in our samples are in fact translated at other developmental stages in *Drosophila*. Finally, because transcripts with significant p-values are statistically more highly counted in their extensions than those without significant p-values ($p = 2.4 \times 10^{-3}$, Mann-Whitney U test), we surmise that our ability to detect regulation was limited by sequencing depth and that the true number of transcripts whose readthrough rates are regulated in tissue- or condition-specific manners is in fact larger than we report.

Extensions contain functional nuclear localization signals

Many peptide sequences — such as signal sequences, degrons, and phosphorylation sites — affect the localization, stability, or activity of proteins. Because these sequences are frequently short and/or degenerate, a high proportion of even random peptide sequences confer function (Kaiser & Botstein, 1990; Kaiser et al., 1987). Thus, a C-terminal extension produced by termination failure could purely by chance alter the function or behavior of its host protein, and thus come under selection. Indeed, Freitag and colleagues reported two readthrough events in fungi that append peroxisomal localization signals (PTS1) to the C-termini of glyceraldehyde-3-phosphate dehydrogenase and 3-phosphoglycerate kinase, enabling these cytosolic enzymes to function in peroxisomal metabolism (Freitag et al., 2012). We therefore searched our full set of C-terminal extensions for short peptide signals that direct peroxisome localization, nuclear localization, prenylation, or ER retention, or that resemble transmembrane domains (see methods). Notably, ten proteins not annotated as nuclear in FlyBase contain predicted NLSes in their extensions. One extensions contains a predicted transmembrane domains, and another, a C-terminal prenylation signal. No extension contained an ER retention signal (table 2).

To determine whether any of the putative nuclear localization signals (NLSes) function *in vivo*, we constitutively fused C-terminal extensions containing putative NLSes to the C-terminus of a GFP-mCherry-GST reporter, which is excluded from the nucleus (*Fig. 7*, left column; Chan et al., 2007). When expressed in S2 cells, three of four NLSes relocalized the cytosolic reporter to the nucleus at levels above background (*Fig. 7*, columns 3–5), arguing that these extensions can regulate the localization of their endogenous host proteins. Given the large number of short peptide signals (e.g. phosphorylation or degradation motifs, ubiquitination sequences) that have been discovered, and the limited number of reporters we tested here, we likely underestimate the number of extensions that confer function. Nonetheless, our results clearly establish that C-terminal extensions can alter protein function.

Gene ID	Alias	Extension coordinates	PhyloCSF score	Signal predicted
FBgn0000173	ben	X:13892649-13892781(+)	-302.18	NLS
FBgn0005278	Sam-S	2L:113542-113647(+)	-195.30	NLS
FBgn0026144	CBP	X:7235840-7236599(+)	128.52	NLS
FBgn0031897	CG13784	2L:7206347-7208015(-)	4775.49	NLS
FBgn0033712	CG13163	2R:8209607-8209934(+)	-675.02	NLS
FBgn0036272	CG4300	3L:12265284-12265557(-)	-193.87	NLS
FBgn0039213	atl	3R:20459429-20459720(+)	28.43	NLS
FBgn0260934	par-1	2R:15370912-15371608(+)	654.90	NLS
FBgn0261606	RpL27A	2L:4457220-4457289^4457374-4457380(-)	-148.56	NLS
FBgn0262114	RanBPM	2R:6322727-6323228(+)	1045.90	NLS
FBgn0031683	CG4230	2L:5098384-5098573(+)	-5.34	Transmembrane domain
FBgn0033712	CG13163	2R:8209607-8209934(+)	-675.02	Transmembrane domain
FBgn0035498	Fit1	3L:4106386-4106518(+)	-323.36	Transmembrane domain
FBgn0036980	RhoBTB	3L:20374798-20374821^20374891-20374982(+)	154.91	Transmembrane domain
FBgn0037321	CG1172	3R:1221902-1222220(+)	-624.55	Transmembrane domain
FBgn0040813	Nplp2	3L:13350197-13350296(+)	-242.85	Transmembrane domain
FBgn0053523	CG33523	3L:5922386-5922854(+)	383.85	Transmembrane domain
FBgn0263864	Ark	2R:12913933-12914062(+)	-123.89	Transmembrane domain
FBgn0039690	CG1969	3R:25567115-25567154(+)	11.52	PTS1
FBgn0035540	Syx17	3L:4404848-4404983(+)	290.83	Farnesyltransferase signal

Table 2. C-terminal extensions contain predicted localization signals

Peptide sequences of C-terminal extensions were examined using various prediction servers (see methods). Those containing predicted features are shown here. *NLS:* nuclear localization signal. *PTS1:* peroxisome localization signal. Coordinates are zero-indexed and half-open. Splice junctions are denoted with carrots (^). Strands are indicated in parentheses.

Discussion

Here we present the first comprehensive study of stop codon readthrough in a eukaryote. Using empirical data, we identified 350 readthrough events in Drosophila melanogaster, the vast

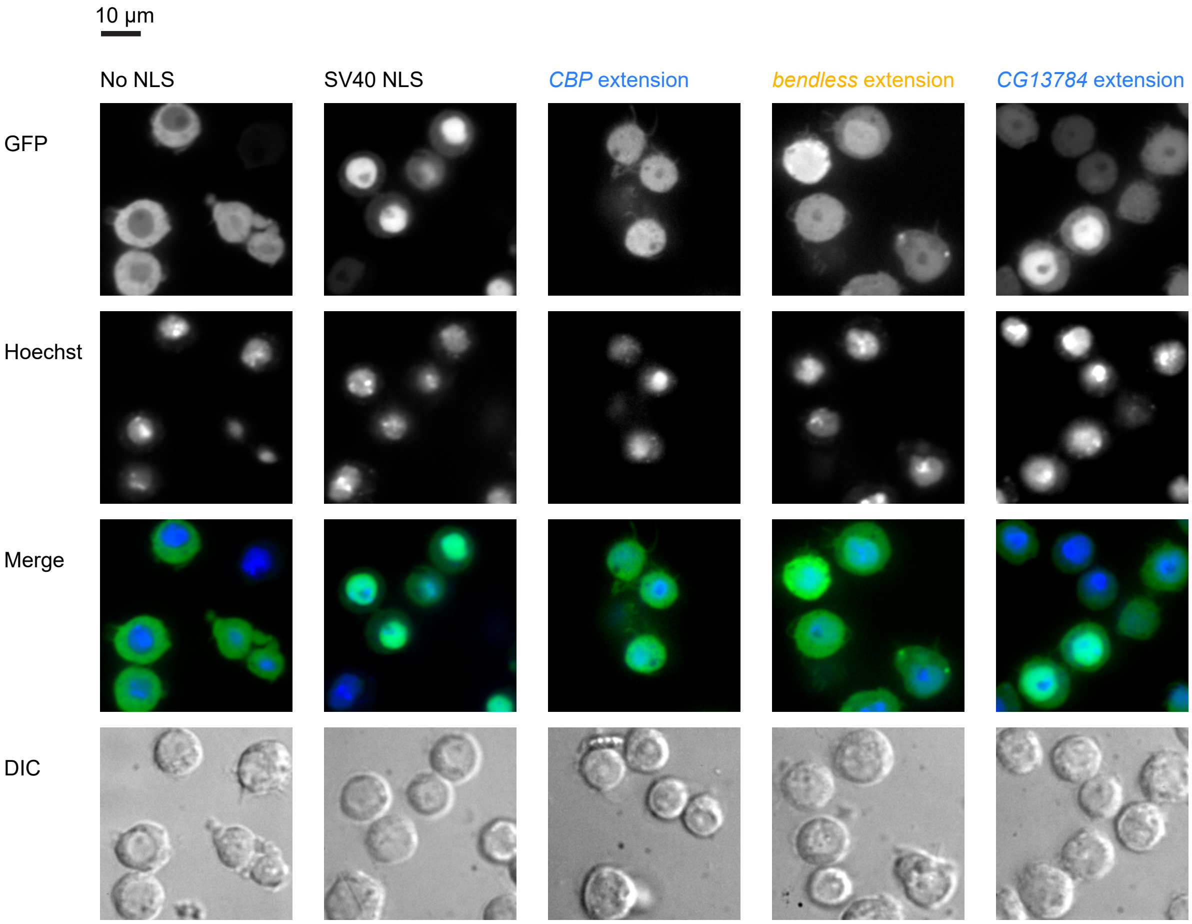


Figure 7: Extensions contain functional localization signals Ordinarily, a GFP-mCherry-GST reporter is excluded from the nucleus (first column). When an SV40 NLS is appended to the reporter, it is predominantly nuclear (second column). Three extensions also contain functional NLSes which at least partially relocalize the reporter to the nucleus when constitutively fused to it (remaining columns). *First row:* GFP reporter. *Second row:* nuclei stained with Hoechst. *Third row:* merged GFP and Hoechst. *Fourth row:* DIC. *Blue labels:* predicted extensions. *Yellow label:* novel extension

majority of which were not predicted from phylogenetic signatures. We further demonstrate that readthrough occurs in yeast and humans. Our studies indicate that readthrough is far more pervasive than appreciated, is biologically regulated, and appends biologically active peptide signals to host proteins. Together, these results argue that readthrough may provide an important mechanism to regulate gene expression and function. Our work further suggests that readthrough provides a means for genes to acquire new functions throughout the course of evolution.

Mechanistic studies of readthrough in various systems have implicated many factors in the modulation of readthrough rates. These include the identity of the stop codon (Chao et al., 2003; Napthine et al., 2012; Robinson & Cooley, 1997), nucleotide context surrounding the stop codon (Bonetti et al., 1995; Cassan & Rousset, 2001; Chao et al., 2003; McCaughan et al., 1995), local or distant RNA structures (Cimino et al., 2011; Feng et al., 1992; Firth et al., 2011; Napthine et al., 2012; Steneberg & Samakovlis, 2001; Wills et al., 1991), specific hexanucleotide sequences (Harrell et al., 2002; Skuzeski et al., 1991), snoRNA-mediated pseudouridylation of stop codons (Karijolich & Yu, 2011), the identity of the tRNA present in the ribosomal P-site (Mottagui-Tabar et al., 1998), the peptide sequence of the nascent chain (Mottagui-Tabar et al., 1998), the concentrations of endogenous suppressor tRNAs (reviewed in Beier & Grimm, 2001), and proteins that bind the ribosome or mRNA (Green et al., 2012; Hatin et al., 2007; Keeling et al., 2004). With the exception of the readthrough signal identified in Tobacco Mosaic Virus (Skuzeski et al., 1991), the majority of readthrough events that have been mechanistically characterized are regulated by two or more such factors, often in complex, context-specific ways. For example, downstream nucleotide contexts which promote readthrough of one stop codon can inhibit readthrough of other stop codons, and these effects can be non-linearly synergistic with upstream nucleotide contexts (Bonetti et al., 1995).

Such complexity is advantageous insofar as it allows readthrough rates to be independently regulated for each transcript, consistent with our own observations. Unsurprisingly, however, this complexity has hindered efforts to identify simple *cis*-acting sequence elements that deterministically predict readthrough, and underscores the importance of having a method to measure readthrough empirically in a physiological setting *in vivo*. By using ribosome profiling to measure readthrough rates over a variety of tissue types and developmental stages, it may be possible to decompose the regulatory complexity into individual components, and then determine the *cis*-acting elements that collaborate to regulate readthrough in tissue-specific manners.

Just as alternative splicing provides a means for proteins to acquire new domains or functional modules, we propose, along with the Lindquist (True & Lindquist, 2000) and Kellis (Jungreis et al., 2011) groups, that stop codon readthrough can provide a mechanism for proteins to evolve at the C-terminus. In this model, transcripts that contain contexts favorable to leaky termination would yield substoichiometric, C-terminally extended populations of cellular proteins. If a particular extension is deleterious, natural selection can favor mutations in the corresponding mRNA that promote efficient termination rather than readthrough. If, instead, the extension provides a fitness advantage, selection can act upon both its amino acid sequence (to tune its function), as well as the nucleotide sequence of its mRNA (to increase or otherwise regulate its readthrough rate). In extreme cases, where an extension is universally advantageous, a mutation that changes a stop codon to a sense codon might become fixed, resulting in a constitutively extended gene. Conceivably, the two C-terminal extensions that we discovered to contain sense codons in place of their annotated stop codons could be the end result of this process.

Several lines of evidence are consistent with this evolutionary model. First, non-zero readthrough rates (0.02–1.4%) have been observed even for control non-readthrough reporter constructs in *[psi$^-$]* yeast (Bonetti et al., 1995; Fearon et al., 1994; Keeling et al., 2004; Namy et al., 2002; Torabi & Kruglyak, 2011) and mammalian cells (Firth et al., 2011; Napthine et al., 2012), arguing that under typical conditions in a variety of eukaryotes, there is a small pool of C-terminally-extended proteins, originating from a wide variety of genes, available for selection to act upon.

Secondly, in specific circumstances, selection appears to favor leaky termination and its extension products. Torabi and colleagues (Torabi & Kruglyak, 2011) reported that in a panel of wild strains of *[psi$^-$]* yeast, allelic combinations of *SUP45* and *TRM10* that promote and inhibit readthrough appear to be in balancing selection, implying that a low baseline level of readthrough is beneficial. Similarly, numerous reports have demonstrated that wild strains of yeast exhibit *[PSI$^+$]*-dependent fitness advantages in a variety of stress conditions, arguing that functions conferred by C-terminal extensions can provide adaptive advantages (Halfmann et al., 2012; True & Lindquist, 2000).

Thirdly, extensions have a high probability of conferring function without prior tuning by natural selection. This point is illustrated by the studies of Kaiser and Botstein, which demonstrated that a large proportion — roughly 30% — of randomly-generated peptide sequences are biologically active, insofar as they can relocalize a cytosolic form of invertase to the nucleus, mitochondrion, or endoplasmic reticulum in yeast (Kaiser & Botstein, 1990; Kaiser et al., 1987). Given the large number of short peptide signals now known (e.g. D-boxes, KEN-boxes, SH3 binding epitopes, phosphorylation sites, et c.), it is likely that a far greater fraction of random peptide sequences contain at least one functional signal. Consistent with this hypothesis, we discovered C-terminal extensions that are not phylogenetically conserved nonetheless contained functional NLSes in *Drosophila*. Furthermore, because these short signals are modular, their addition to the C-terminus of a protein can confer novel function, without requiring modification or coevolution of the host protein. In this way, even novel C-terminal extensions arising purely from termination failure can immediately alter the behavior of their host proteins, in beneficial or deleterious ways, and thus come under selection. Over evolutionary time, this process could yield phylogenetically conserved readthrough events or, in extreme cases, constitutively extended proteins.

Our model predicts that, at any given moment, ribosome profiling should detect a broad spectrum of conservation among readthrough events: at one extreme are ancient, phylogenetically conserved extensions, and, at the other, extensions of recent evolutionary origin. Between these, one would find extensions under varying degrees of age, conservation, and selection. This notion is borne out in our data: in *Drosophila*, a subset of readthrough events are well supported by conservative codon substitutions across the phylogeny (Jungreis et al., 2011; Lin et al., 2007), but a far larger set is not conserved between species. In aggregate, this non-conserved group shows weak but statistically significant signals of selection among fifty wild-type individuals of *Drosophila melanogaster* (*Fig. 6C*), suggesting that a fraction of this group is undergoing protein coding selection. The remainder might include many other different groups of extensions: namely, a group of extensions undergoing diversifying selection, a group of deleterious extensions undergoing counterselection, and a group of selectively-neutral extensions subject to genetic drift.

Finally, our model predicts that conserved extensions should on average exhibit higher readthrough rates than novel extensions, because only a subset of the latter group would have been selected for function and regulation. Our data is also consistent with this prediction: the median readthrough rate for the conserved extensions in *Drosophila* is 5.2%, while the median for the novel extensions is 1.7%. Notably, 62% of the novel extensions we identified in *Drosophila*, 95% of the extensions in human foreskin fibroblasts, and 40% of the extensions in yeast undergo readthrough at rates comparable to those of phylogenetically conserved extensions (*Fig. 5F*), arguing for the importance of these subsets.

Broadly, our work builds upon the growing amount of evidence that eukaryotic genomes and proteomes are far more plastic than previously thought, particularly with regard to translation and coding. In addition to the large number C-terminal extensions we report, various groups have used ribosome profiling to determine that large numbers of genes are regulated by uORFs that initiate at near-cognate start codons (Brar et al., 2012; Ingolia et al., 2009), that many genes can be N-terminally extended in a regulated manner (Fritsch et al., 2012), and even that many parts of mammalian genomes are decoded in multiple frames (Michel et al., 2012). Given the preeminence of *Drosophila* as a developmental model and the abundance of conditional genetic tools available, we anticipate that ribosome profiling in *Drosophila* will be useful in deciphering the biological roles of not only readthrough, but all non-canonical translation events, throughout development.

Materials & methods

Cell culture

Wild-type (*y w*) flies were cultured according to standard procedures. S2 cells were cultured in Schneider's (Gibco by Life Technologies, Carlsbad, California) media supplemented with 10% heat-inactivated FBS (UCSF cell culture facility, San Francisco, California) and antibiotics (UCSF cell culture facility). S2 cells were transfected using Effectene reagent (Qiagen) following the manufacturer's instructions. For stable transfectants, the plasmid of interest was co-transfected at a 10:1 molar excess with pCoPuro. Stable integrants were selected and maintained in Schneider's media supplemented as above, but additionally containing 10 μg/ml puromycin.

Lysate preparation

s2 cells. 16–20 hours before an experiment, cultures were diluted to 1.5–1.8 million cells/ml. To start the experiment, cells were treated for two minutes with 0.01 volumes 2 mg/ml emetine (Sigma-Aldrich, St Louis, Missouri), pelleted for 2 minutes at 1600 rpm in a tabletop centrifuge, resuspended in 4–6 cell volumes cold polysome lysis buffer (50 mM Tris pH 7.5, 150 mM NaCl, 5 mM $MgCl_2$, 0.5% Triton x-100, 1 mM DTT, 20 U/ml SuperaseIn (Ambion by Life Technologies), 20 µg/ml emetine), and homogenized on ice in a pre-chilled dounce homogenizer. The resulting lysate was clarified by spinning 10 minutes at 20000 × *g* at 4°C in a microcentrifuge. Clarified lysate was aliquoted, flash-frozen in liquid nitrogen, and stored at -80°C. Experiments used 12–96 ml s2 cell culture, depending on the application. For ribosome profiling, a single 24 ml culture is sufficient.

Embryos. 0–2 hour old wild-type (*y w*) embryos were collected from egg laying dishes directly into a 50 ml conical tube full of liquid nitrogen using a rubber policeman. Multiple collections were pooled until roughly 200 µl embryos had been collected for each sample. The liquid nitrogen was then decanted, the tube capped, and the pooled embryos stored at -80°C. Frozen pellets of a modified polysome lysis buffer additionally including 50 µM GMP-PNP (Sigma-Aldrich) were prepared by dripping buffer into a conical tube of liquid nitrogen. The nitrogen was decanted, the tube capped, and buffer pellets stored at -80°C. Frozen embryos and 4–6 volumes of frozen buffer pellets were ground together 6 times for 2 minutes each at 15 Hz in a Tissue-Lyser (Qiagen), chilling the canisters in liquid nitrogen before and after each round of grinding. Grindate was either stored at -80°C, or thawed immediately under running tepid water. Thawed grindate was clarified by spinning at 3000 × *g* in a tabletop centrifuge. Avoiding the wax and fat layers at the top, the supernatant was collected into pre-chilled microcentrifuge tubes, and clarified again by spinning 10 minutes at 20000 × *g* at 4°C. Lysates were aliquoted, flash-frozen in liquid nitrogen, and stored at -80°C.

Ribosome footprinting

Concentrations of total RNA in lysates were determined using the RiboGreen kit (Molecular Probes by Life Technologies). For each sample, 35–100 µg total RNA was diluted 2:1 in digestion buffer (50 mM Tris pH 7.5, 5 mM $MgCl_2$, 0.5% Triton X-100, 1 mM DTT, 20 U/ml SuperaseIn, 20 µg/ml emetine, 15 mM $CaCl_2$, and 3 U micrococcal nuclease (Roche Applied Science, Indianapolis, Indiana) per µg of total RNA in the sample), to bring the final concentration of NaCl to 100 mM and $CaCl_2$ to 5 mM. Samples were digested for 40 minutes at 25°C in a Thermomixer (Eppendorf, Hamburg, Germany). Digestions were quenched by adding EGTA to a final concentration of 6.25 mM and placing the reactions on ice. 1 U MNase is defined as previously (Oh et al., 2011) as an increase of 0.005 A260 per minute, measured in a Spectramax M2 plate reader (Molecular Devices, Sunnyvale, California) using 10 µg/ml salmon sperm DNA (Sigma-Aldrich) with 5 mM Ca^{2+} and 20 mM Tris, pH 8.0 in a 0.1ml reaction at 25°C.

Sucrose gradients

10–50% sucrose gradients were prepared in polysome gradient buffer (250 mM NaCl, 15 mM $MgCl_2$, 20 U/ml SuperaseIn, 20 µg/ml emetine) using a GradientMaster (Biocomp Instruments, Fredericton, New Brunswick, Canada) in polyclear centrifuge tubes (Seton Scientific, Petaluma, California). Up to 200 µl of samples was applied to the top of each gradient. Gradients were resolved by spinning for 3 hours at 35 krpm at 4°C in an SW-41 rotor (Beckmann Coulter, Brea, California), and fractionated using the GradientMaster. When appropriate, monosome fractions were collected, flash-frozen in liquid nitrogen, and stored at -80°C.

Sucrose cushions

Up to 0.5 ml of digested sample was layered atop 1.0 ml of a solution of 34% sucrose in polysome gradient buffer. Monosomes were sedimented by spinning for four hours at 70 krpm at 4°C in a TLA-110 rotor (Beckmann Coulter). Pellets were resuspended in 600 µl 10 mM Tris, pH 7.0 and stored at -20°C.

Ribosome profiling of Drosophila melanogaster

Lysates were prepared and footprinted as above. Unless otherwise indicated, monosomes were enriched by sedimentation through 34% sucrose cushions and resuspended in 600 µl 10 mM Tris, pH 7.0. Resuspended monosomes were extracted once with 700 µl 65°C acid phenol and 40 µl 10% SDS, followed by 650 µl acid phenol and a final extraction with chloroform. RNA was precipitated for at least two hours at -30°C, resuspended in 10mM Tris, pH 7.0, and quantitated on a NanoDrop spectrophotometer (Thermo Scientific, Asheville, North Carolina). 5–35 µg RNA was dephosphorylated for one hour at 37°C using T4 polynucleotide kinase (New England Biolabs, Ipswich, Massachusetts) in a 50 µl reaction and resolved on a 15% TBE-urea gel (Invitrogen by Life Technologies). A gel slab spanning 28–34 nt (as measured by oligoribonucleotide size standards in a neighboring lane; see Supplemental file 2) was excised from the gel, eluted, and precipitated. Samples were then carried through all steps of library generation (see below).

Poly(A)+ RNA-seq of Drosophila melanogaster

For each sample, 375 µl of undigested polysome lysate was diluted into 3 volumes Trizol LS (Invitrogen) and total RNA was extracted following the manufacturer's instructions. 20–50 µg Poly(A)+ RNA was selected on oligo-dT25 DynaBeads (Invitrogen) per manufacturer's instructions, and fragmented at 95°C in fragmentation buffer (2 mM EDTA, 100 mM $NaCO_3$ / $NaHCO_3$, pH 9.2) to a mean size of roughly 100 nt. Fragmented RNA was precipitated, dephosphorylated for one hour at 37°C with T4 polynucleotide kinase (New England Biolabs), and resolved on a 15% TBE-urea gel. A gel slab corresponding to 55–65 nt was excised from the gel, eluted, and precipitated. Samples were then carried through all steps of library generation (see below).

Subtractive hybridization to remove rRNA-derived fragments

We performed two sequential rounds of subtractive hybridization on each sample. To 5 µl cDNA the following were added: 1 µl 20× SSC, 3 µl nuclease-free water, and 1 µl of a 60 µM mixture of the biotinylated oligonucleotides OJGD132, OJGD133, OJGD134, OJGD135, OJGD136, OJGD161, OJGD162, OJGD163, and OJGD164 (sequences in Supplemental file 2) mixed in a ratio of 25.5:1:13:17:4:6:2:11:21. Samples were denatured for 90 seconds at 95°C and annealed for 20 minutes at 25°C. MyOne Streptavidin C1 DynaBeads (Invitrogen) were prepared as follows: for each sample, 45 µl of beads were aliquoted into a microcentrifuge tube and washed 3 times in 50 µl 2× binding buffer (10 mM Tris, pH 7.5, 1 mM EDTA, 2 M NaCl), and resuspended in 22.5 µl 2× binding buffer. 10 µl equilibrated beads were added to 10 µl hybridized sample. The mixture was incubated at 20 minutes in a room temperature Thermomixer with shaking at 850 rpm. Beads were then separated on a magnetic manifold (Invitrogen) and the supernatant recovered to a microcentrifuge tube.

For the second round of subtraction, 1 µl 60 µM biotinylated oligo mix and 1 µl 20× SSC were added to the supernatant from the first subtraction, and the denaturation and annealing repeated. 10 µl of equilibrated beads were pelleted on a magnetic manifold. The buffer was removed, and the beads resuspended in the mixture from the second hybridization. Samples were then incubated at 20 minutes in a room temperature Thermomixer with shaking at 850 rpm. The supernatant was recovered on a magnetic manifold, transferred to a microcentrifuge tube, precipitated, and resuspended in 15 µl 10 mM Tris, pH 8.0.

Library generation

RNA concentrations were measured using the Small RNA Series II Bioanalyzer assay (Agilent Technologies, Santa Clara, California). 10–15 picomoles of RNAs were ligated to 1 µg 3´ miRNA cloning linker 1 (Integrated DNA Technologies, Coralvaille, Iowa) for two and a half hours at 25°C in ligase buffer (1× T4 RNA ligase 2 buffer (New England Biolabs), 40% PEG-100 (Sigma-Aldrich), 5% DMSO, truncated T4 RNA ligase 2 K227Q (a kind gift from Calvin Jan)) in a 20 µl reaction. Ligated fragments were precipitated for at least two hours at -30°C, purified on a 10% TBE-urea gel, eluted, and precipitated. Ligation products were then reverse-transcribed using SuperScript

III (Invitrogen) in a 16.7 µl reaction using using the primer o225-link1 (see S8). RNA template was hydrolyzed by addition of 1/10 volume 1 M NaOH and incubation at 95°C for 20 minutes. cDNAs were purified on a 10% TBE-urea gel (Invitrogen), eluted, precipitated, and resuspended in 5 µl 10mM Tris pH 7.0. cDNAs from footprint samples were subjected to two rounds of subtractive hybridization as described above.

Subtracted samples were circularized using CircLigase (Epicentre, Madison, Wisconsin), following manufacturer's instructions in a 20 µl reaction. An additional microliter of CircLigase was then added, and the circularization repeated a second time. Circularized libraries were amplified by 6–12 cycles of PCR using ONTI231 and any of four indexing primers oCJ30–33 (Supplemental file 2) using Phusion polymerase (Finnzymes by ThermoScientific) in a 17 µl reaction. Amplification products were size-selected on 8% TBE gels (Invitrogen), eluted, precipitated, and resuspended in 10 µl 10 mM Tris, pH 8.0. Samples were then quantitated using the Bioanalyzer High Sensitivity DNA assay (Agilent Technologies), diluted to 2 nM, multiplexed as needed, and subjected to 50–57 cycles of single-end sequencing on an Illumina HiSeq sequencer (Illumina, San Diego, CA) using version 3 clustering and sequencing kits with a 6-cycle index read (Illumina).

Sequence processing and alignment

For all *Drosophila* experiments we used revision 5.43 of the FlyBase genome annotation and the corresponding genome assembly (McQuilton et al., 2012). Reads were demultiplexed and cleaned of 3′ cloning adapters using in-house scripts. Reads shorter than 25 nt were discarded. Remaining reads were aligned using Bowtie version 0.12.7 (Langmead et al., 2009) sequentially to Bowtie indices composed of the following sequences: (a) *D. melanogaster* rRNAs (GenBank accession no. M21017 (Tautz et al., 1988) and from FlyBase), (b) *D. melanogaster* tRNAs, snoRNAs, and snRNAs (from FlyBase), (c) cloning oligos, (d) the S288C yeast genome version R64-1-1 (downloaded on June 6, 2011 from http://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases/), (e) Wolbachia (GenBank accession no. AE017196), (f) *D. melanogaster* chromosome arms, and (g) splice junctions (from FlyBase and, in the case of embryos, supplemented with junctions discovered in the pooled embryo mRNA datasets using HMMsplicer 0.95 (Dimon et al., 2010). For all quantitative analyses, we counted only uniquely-mapped reads.

Alignments were assigned to genomic coordinates as follows. Randomly-fragmented poly(A)+ mRNA alignments were counted along the entire length of the alignment. Each genomic position covered by a single RNA fragment was incremented $1/l$, where l corresponds to the length of the alignment. Ribosome-protected footprint alignments were mapped to their estimated P-sites as follows: 12 nt were pruned from each end of the alignment, leaving a fragment n nt long (where $n = l - 2 \times 12$). Each genomic position covered by a nucleotide remaining in the pruned alignment was then incremented by $1/n$. Thus, the P-site of each 25mer was assigned to one unique position, while the P-site of each 26-mer was spread over two positions, each incremented by 0.5 reads, and so on. Alignment statistics are given in supplemental file *1B*.

Attribution of counts to genes and transcripts

Masking of degenerate genomic positions. To determine which positions in the genome give rise to reads that fail to uniquely map, we divided the genome into all possible 29-mers centered on each nucleotide position, and aligned the resulting 29-mer back to the genome allowing zero mismatches. If the 29-mer aligned to multiple sites, the position from which it arose was flagged as degenerate. All such positions were excluded from further analysis.

Attribution of nucleotide positions to loci. Because the genome annotation contains polycistronic transcripts in which each cistron is annotated as belonging to a separate gene — for example tarsal-less / polished rice, which is annotated as four separate genes (FBGN0259730–3) — we collapsed each set of genes whose transcripts share exons (370 genes total) into 179 merged loci. All nucleotide positions in any transcript deriving from a locus were attributed to that locus. Any nucleotide position attributed to multiple loci (e.g. overlapping genes on the same strand), were excluded from further analyses on the gene or transcript levels.

Attribution of nucleotide positions to exons, 5´UTRs, 3´UTRs, and coding regions. For each locus, any position included in any transcript deriving from that locus was included in the list of exonic positions for that gene. Any exonic position which could be labeled as two or more of CDS, 5´UTR, or 3´UTR depending on the transcript isoform was still counted as exonic, but was excluded from analyses that required positions to be uniquely labeled (e.g. comparisons of translation in 5´ or 3´ UTRs to CDS) unless otherwise noted.

Filtering of countable loci. For all analyses, we counted only loci or transcripts deriving from loci that contain at least 95% non-degenerate positions and are at least 60 nucleotides in exonic length, after exclusion of degenerate positions and positions covered by multiple loci. Genes and transcripts that are not translated but which may contaminate the data due to their abundance (e.g. those that encode microRNAs, rRNAs, snRNAs, snoRNAs, and tRNAs) were excluded from analysis. We also excluded the loci *mod(mdg-4)* (which contains transcripts deriving from both strands) and *Yeti*, (for which transcript annotations existed on chromosome arms 3R and 2RHet).

Measurements of gene expression

mRNA abundance and ribosome density for each genomic feature were measured in reads per kilobase of feature length per million reads aligning to chromosomes or splice junctions in the dataset (RPKM), a unit which corrects for both feature length and sequencing depth. Unless otherwise indicated, the RPKM values we report for mRNA abundance reflect the total number of RNA fragments aligning to all countable exonic positions for a given locus. For ribosome density, we report the total number of ribosome-protected footprint fragments aligning to all countable positions of a coding region (CDS) for a given locus. We calculate translation efficiency as the ratio of footprint RPKM in the CDS to the RNA fragment RPKM across the entire locus. When comparing mRNA fragment or footprint density between samples, we restricted our analyses to genes that had at least 128 summed counts between replicates as determined in figure 1 supplement 2. When comparing translation efficiencies between samples, we required at least 128 exonic counts of mRNA for each gene.

Translation efficiency of 5′ UTRs, CDS, and 3′ UTRs

Translation efficiencies for these regions were calculated as the ratio of footprint counts to mRNA counts in each region, for all regions with at least 128 mRNA counts. We excluded all positions that could be labeled as two or more of 5′ UTR, CDS, or 3′ UTR depending upon transcript isoform. To remove variability or bleedthrough introduced by start and stop codon peaks, we additionally excluded the following genomic positions from consideration: 9 nucleotides preceding each start codon, 15 nucleotides following each start codon, the 15 nucleotides preceding each stop codon, and the fifteen nucleotides following each stop codon.

Identification of C-terminal protein extensions

Mapping predicted extensions to transcripts in the modern annotation. C-terminal extensions predicted by Jungreis et al. (Jungreis et al., 2011) were mapped onto the FlyBase annotation 5.43 as follows: first, 26 predicted extensions that overlap regions that are annotated as coding (for reasons other than readthrough) in the present annotation were excluded from further analysis. One additional extension was excluded because it overlapped the 5′ UTR of another gene. Then, the remaining 256 extensions were mapped to transcripts in FlyBase r5.43 that satisfied the following criteria: (a) if the transcript contains an annotated 3′ UTR, it fully covers the extension and (b) the transcript's annotated stop codon must immediately precede the extension in transcript coordinates.

Readthrough rates. Stop codon readthrough rates were evaluated by dividing the ribosome density (in RPKM) for each C-terminal extension by the ribosome density in the corresponding CDS. In cases where multiple transcript isoforms contained the same extension, the transcript that minimized the ratio of ribosome footprint density in the extension to the density in the CDS was reported. To control for variability introduced by start and stop codon peaks (see *Fig. 2*, supplement *1*), we excluded the following genomic positions from our totals: 12 nucleotides following the start codon, the 15 nucleotides preceding the stop codons of the coding region and the extension, and the nine nucleotides following the stop codon of the coding region.

Scoring of predicted extensions. A predicted extension was scored positively if: (a) there existed ribosome density in the extension, (b) ribosome density vanished or unambiguously decreased after the extension's in-frame stop codon, and (c) positions occupied by ribosomes in the readthrough region were evenly-spaced throughout the extension. When ribosome density was sparse in the extension, we relaxed criterion (c) and additionally required a peak of at least two reads at the extension's stop codon. Aside from *Kelch,* which has been demonstrated to undergo readthrough experimentally (Robinson & Cooley, 1997), we did not positively score any extension that contained a methionine in its first three codons, as these could represent downstream cistrons rather than true extensions. Furthermore, we required read density upstream of the first methionine in any extension containing a methionine codon.

Identification of novel extensions. We identified all coding transcripts in FlyBase r5.43 in which: (a) the 3′ UTR was annotated, (b) a C-terminal extension was not predicted by Jungreis et al, (c) there were at least 5 codons between the annotated stop codon of the CDS and the next in-frame stop codon, and (d) the region between these stop codons (the putative extension) did not overlap any annotated CDS, 5′UTR, tRNA, rRNA, snRNA, snoRNA, miRNA, or pre-miRNA. We additionally excluded extensions whose translation could be explained by alternative splice isoforms whose transcripts omitted the stop codon, using splice junctions from FlyBase revision 5.43 and inferred from our on RNA-seq data, as described above.

Following the same scoring criteria we used for the extensions predicted by Jungreis et al., we scored each candidate extension that met the following criteria: (a) a minimum read density of 0.2 RPKM in the extension, (b) a minimum readthrough rate of 0.001, (c) at least 10% of the nucleotide positions in the extension covered by reads, (d) the first read occurring within the first quartile of extension length, (e) the last read occurring within last quartile of the extension length, and (f) a 75% or greater decrease in read density in the first 114 nucleotides of distal 3′ UTR compared to the extension. To calculate this last statistic for transcripts whose distal 3′ UTR was less than 114 nt in length, we extended the distal 3′ UTR in uninterrupted genomic coordinates to 114 nt in length.

Metagene analyses

For each analyses (*Fig. 4B*), we identified regions of interest (ROIs) germane to the analysis. In figure 2 supplement 1, these included roughly 3000 ROIs each for the left and right panels, each of which met the following criteria: (a) all transcripts deriving from that gene had one annotated start codon (left panel) or stop codon (right panel), (b) all transcripts deriving from that locus covered identical genomic positions over the region of interest (ROI) shown, (c) all positions within the ROI were non-degenerate (see methods), and (d) at least 10 reads were present in the coding subregion of the ROI. For coding regions in *Fig. 4C*, we kept the same criteria as above but required only 0.5 reads in the coding subregion of each ROI, yielding roughly 7401 ROI for that set. For C-terminal extensions, we required only that the extension be long enough to cover the interval shown, and have 0.5 reads in the coding subregion, allowing us to include 123 of the 350 extensions.

For each ROI, we then generated a “coverage vector” tallying ribosome density at each nucleotide position. We then normalized each coverage vector to the mean number of footprint reads covering the annotated coding region in the ROI, excluding a 3-codon buffer flanking the start or stop codon to avoid bleedthrough from initiation or termination peaks. We then plotted the median value across all normalized coverage vectors at each position.

Search for genomic polymorphisms and A-to-I editing

To improve our sensitivity in detection, we re-aligned our footprint and mRNA datasets to a

Bowtie database of spliced transcript models, allowing three mismatches (where we previously only allowed two). For the first, second, and third nucleotide position in each unique, annotated stop codon, we counted the number of matching and mismatching nucleotides in each read alignment covering that position. We ignored mismatches that occurred in the first position of the read alignment, because they frequently arise from non-templated nucleotide addition by reverse transcriptase. Considering the first, second, and third positions of each stop codon separately, we calculated a global average mismatch frequency for each. We then searched for individual stop codon positions that far exceeded the corresponding global average using a binomial test, controlling the false discovery rate at 5% following the procedure of Benjamini and Hochberg (Benjamini & Hochberg, 1995). We performed this analysis separately upon each of three datasets: total mRNA, total footprints, and the subset of footprints whose P-sites had passed the nucleotide position in question, following the P-site assignment rules described above.

Immunoprecipitation & western blotting

4–48 ml of transiently or stably transfected cells were harvested 48 or 72 hours post-transfection by centrifuging for two minutes at 1600 rpm in a tabletop centrifuge. All cell pellets were rinsed once in PBS, and flash-frozen in a bath of dry ice in ethanol. Cell pellets were thawed and lysed for at least 15 minutes on ice in 0.5–1.5 ml lysis buffer (150 mM NaCl, 50mM Tris pH 7.5, 1% Triton X-100, 1 mM EDTA and 1× complete protease inhibitor cocktail (Roche Applied Science)), depending upon the pellet size. Lysates were clarified by spinning 10 minutes at 20,000 × *g* in a microcentrifuge and supernatants collected. GFP reporters were immunoprecipitated on 10 μl of anti-GFP beads (Chromotek, Planegg-Martinsried, Germany) equilibrated in IP wash buffer (150 mM NaCl, 50 mM Tris pH 7.5, 1 mM EDTA, 0.05% Triton X-100). The bound fraction was washed three times in IP wash buffer, and finally eluted by boiling for at least five minutes in NuPage sample loading buffer (Invitrogen). Supernatants were collected and transferred to new tubes, and stored at -20°C.

For western blotting, samples were resolved on 4–12% NuPage gels (Invitrogen) in MOPS buffer. Gel lanes were loaded such that the amounts of uncleaved GFP reporter in each lane were loaded as equally as possible. GFP was detected using a mouse anti-GFP antibody (Roche Applied Sciences), and visualized using IR800 anti-mouse antibodies on a LI-COR Odyssey system (LI-COR, Lincoln, Nebraska). FLAG was similarly detected on a separate gel, instead using the M2 Mouse anti-FLAG antibody (Sigma-Aldrich).

PhyloCSF analysis

PhyloCSF analysis was performed on all C-terminal extensions at least 5 codons long, exclusive of the stop codons. Multiple species alignments were obtained from the *Drosophila* 12-way multispecies alignment as downloaded from the UCSC genome browser, and stitched together over regions of interest using the Phast utility *maf_parse* (Hubisz et al., 2011). PhyloCSF was then used to evaluate the extension on the empirical codon model *12flies*. Columns in which the *Drosophila melanogaster* sequence contained gaps were ignored. Alignments that contained no sequence besides that from *D. melanogaster* were not evaluated.

Z-Curve classifier

We calculated the 189-variable Z-curve as previously described (Gao & Zhang, 2004). We empirically determined that the classifier became error-prone if trained on sequences 81 nt or shorter in length. Our training set consisted of 81-nucleotide windows drawn from coding regions (the positive set, 14,507 windows) or from portions of distal 3′ UTRs that did not overlap annotated coding regions or 5′ UTRs (the negative set, 8151 windows). To assay the stability of the classifier's behavior and control for overfitting, we trained the classifier with four-fold cross-validation training on 2200 windows from the CDS set and 2200 windows from the distal 3′ UTR set, yielding an average misclassification error of 6.9–7.3% with each iteration. We repeated this analysis (and cross-validation) several times selecting different 81-nucleotide windows from each CDS and distal 3′UTR, obtaining similar levels of error. The classifier was then trained on the entire training set, and used to evaluate randomly chosen 81 nt windows from observed C-terminal protein extensions that were 81 nt or greater in length. These included 26 extensions predicted by Jungreis et al. and 83 novel extensions.

SNP analysis

We downloaded SNP data from the *Drosophila* Population Genomics Project from ensembl.org and counted the proportion of SNPs that, if translated in frame, would cause synonymous substitutions in coding regions, extensions, and distal 3´ UTRs.

Tests for differential regulation of readthrough rates

To test C-terminal extensions for differential readthrough rates, we examined all extensions which met the following criteria: (1) all annotated isoforms covering the extension contain exactly the same CDS, (2) the CDS had at least 128 total footprint reads in each of the S2 cell and embryo samples, and (3) the C-terminal extension had been scored as positive for readthrough in either the S2 and/or the embryo sample. For those extensions, we tabulated the footprint reads that aligned to the CDS and putative extension, masking out regions normally covered by start and stop codon peaks as described (see section *Readthrough rates,* above). For each extension, this tabulation yielded a 2 × 2 contingency table of reads aligning to the CDS and extension in the S2 cell and 0–2 hour embryo datasets. We evaluated the statistical significance of asymmetry in the contingency tables using Fisher's exact test, and controlled the false discovery rate at 5% using the procedure of Benjamini and Hochberg (Benjamini & Hochberg, 1995).

Human and yeast ribosome profiling data

For human cells, we collected data from uninfected human foreskin fibroblasts and processed it as previously described (Stern-Ginossar et al., 2012). Yeast samples were collected from *[psi$^-$]* W303 and processed as previously described (Ingolia et al., 2009), with the exception that a 3´ linker ligation strategy was used instead of poly(A) tailing for fragment capture. For phasing of yeast footprints, we counted only 28-mers, which have previously been shown to be the best-phased footprint population in that organism (Ingolia et al., 2009).

Motif prediction

C-terminal extensions 20 amino acids or longer were scanned for transmembrane domains using *TmHmm* (Krogh et al., 2001) using default settings. Nuclear localization signals were predicted in extensions 20 amino acids or longer using the *cNLS mapper* (Kosugi et al., 2009), with a score cutoff of 7.0. Peroxisome targeting signals were predicted for all extensions 12 amino acids or longer using *PTS1 Predictor* (Neuberger et al., 2003) with the signal type set to *metazoan.* Prenylation signals were predicted for all extensions 12 amino acids or longer using *PrePS* (Maurer-Stroh & Eisenhaber, 2005). In addition, we searched for ER retention signals using the consensus [KH]DEL*.

We searched 3′ UTRs (including the predicted extension and entire distal 3′ UTR) for selenocysteine insertion elements using *SeciSearch 2.19* (Kryukov et al., 2003) with parameters set as follows: *e1 = 05, e2 = -22, Y_filter = True, O_filter = True, B_filter = True, S_filter = True*. We searched each 3′ UTR using every available SECIS Pattern (*pat_c, pat_Sep20, pat_dm, pat_g, and pat_s*), and considered a 3′ UTR receiving a COVE score above the recommended threshold of 15 in any of the pattern searches to contain a SECIS element. Additionally, we excluded any extensions that were annotated as selenoprotein annotations in SelenoDB (Castellano et al., 2008; for *Drosophila*, yeast, and human data) or FlyBase (McQuilton et al., 2012; for *Drosophila*), For transcripts with no annotated or short 3′ UTRs, we extended the 3′ UTR in uninterrupted genome coordinates until it was 1000 nucleotides in length, an in Jungreis et al. (Jungreis et al., 2011).

Microscopy

S2 cells stably transfected with the reporter of interest were maintained at a density of 1.6–12 million cells/ml. Nuclei were visualized by staining with 1 µg/ml Hoechst 34580 (Invitrogen) for at least 5 minutes. Live cells were imaged in culture media on an inverted spinning disk confocal

Nikon Ti microscope (Nikon Instruments) in glass-bottom culture dishes (MatTek, Ashland, Massachusetts). Images were contrast-adjusted and prepared for presentation in Adobe Photoshop (Adobe Systems, San Jose, CA).

Other software and libraries

We wrote custom scripts in Python 2.7, using the following open-source libraries: Numpy 1.6.0 (http://numpy.scipy.org), Scipy 0.11.0rc2, Biopython 1.59 (Cock et al., 2009), PySam, and HTSeq 0.5.1p2 (http://www-huber.embl.de/users/HTSeq/doc/overview.html). Plots and genome browser snapshots were generated using Matplotlib 1.0.1 (Hunter, 2007).

References

Beier, H. & Grimm, M. (2001). Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. *Nucleic Acids Res.,* 29, 4767-82. doi:10.1093/nar/29.23.4767

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. *Series B (Methodological),* 57, 289–300.

Bonetti, B., Fu, L., Moon, J. & Bedwell, D.M. (1995). The efficiency of translation termination is determined by a synergistic interplay between upstream and downstream sequences in *Saccharomyces cerevisiae. J. Mol. Biol.*, 251, 334-45. doi:10.1006/jmbi.1995.0438

Brar, G. A., Yassour, M., Friedman, N., Regev, A., Ingolia, N. T.et al. (2012). High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science,* 335, 552–7. doi:10.1126/science.1215110

Cassan, M. & Rousset, J.P. (2001). UAG readthrough in mammalian cells: effect of upstream and downstream stop codon contexts reveal different signals. *BMC Mol. Biol.,* 2, 3. doi:10.1186/1471-2199-2-3

Castellano, S., Gladyshev, V. N., Guigó, R. & Berry, M.J. (2008). SelenoDB 1.0 : a database of selenoprotein genes, proteins and SECIS elements. *Nucleic Acids Res.,* 36, D332–8. doi:10.1093/nar/gkm731

Chan, W. M., Shaw, P. C. & Chan, H.Y.E. (2007). A green fluorescent protein-based reporter for protein nuclear import studies in *Drosophila* cells. *Fly (Austin),* 1, 340–2.

Chao, A. T., Dierick, H. A., Addy, T. M. & Bejsovec, A. (2003). Mutations in eukaryotic release factors 1 and 3 act as general nonsense suppressors in Drosophila. *Genetics,* 165, 601–12.

Cimino, P. A., Nicholson, B. L., Wu, B., Xu, W. & White, K.A. (2011). Multifaceted regulation of translational readthrough by RNA replication elements in a tombusvirus. *PLoS Pathog,* 7, e1002423. doi:10.1371/journal.ppat.1002423

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J. et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25, 1422–3. doi:10.1093/bioinformatics/btp163

Dimon, M. T., Sorber, K. & DeRisi, J.L. (2010). HMMsplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data. *PLoS One,* 5, e13875. doi:10.1371/journal.pone.0013875

Dunn J.G., Foo C.K., Belletier N.G., Gavis E.R., Weissman J.S. (2013). Data from: Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster*. *Dryad Digital Repository.* doi:10.5061/dryad.6nr73

Edgar, R., Domrachev, M. & Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.,* 30, 207–10. doi:10.1093/nar/30.1.207

Fearon, K., McClendon, V., Bonetti, B. & Bedwell, D.M. (1994). Premature translation termination mutations are efficiently suppressed in a highly conserved region of yeast Ste6p, a member of the ATP-binding cassette (ABC) transporter family. *J. Biol. Chem.,* 269, 17802–8.

Feng, Y. X., Yuan, H., Rein, A. & Levin, J.G. (1992). Bipartite signal for read-through suppression in murine leukemia virus mRNA: an eight-nucleotide purine-rich sequence immediately downstream of the gag termination codon followed by an RNA pseudoknot. *J. Virol.*, 66, 5127–32.

Firth, A. E. & Brierley, I. (2012). Non-canonical translation in RNA viruses. *J. Gen. Virol.*, 93, 1385–409. doi:10.1099/vir.0.042499-0

Firth, A. E., Wills, N. M., Gesteland, R. F. & Atkins, J.F. (2011). Stimulation of stop codon read-through: frequent presence of an extended 3′ RNA structural element. *Nucleic Acids Res.,* 39, 6679–91. doi:10.1093/nar/gkr224

Freitag, J., Ast, J. & Bölker, M. (2012). Cryptic peroxisomal targeting via alternative splicing and stop codon read-through in fungi. *Nature,* 485, 522–5. doi:10.1038/nature11051

Fritsch, C., Herrmann, A., Nothnagel, M., Szafranski, K., Huse, K. et al. (2012). Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Res.*, 22, 2208–18. doi:10.1101/gr.139568.112

Gao, F. & Zhang, C. (2004). Comparison of various algorithms for recognizing short coding sequences of human genes. *Bioinformatics*, 20, 673–81. doi:10.1093/bioinformatics/btg467

Geller, A. I. & Rich, A. (1980). A UGA termination suppression tRNA^{Trp} active in rabbit reticulocytes. *Nature,* 283, 41–6. doi:10.1038/283041a0

Green, L., Houck-Loomis, B., Yueh, A. & Goff, S.P. (2012). Large ribosomal protein 4 increases efficiency of viral recoding sequences. *J. Virol.*, 86, 8949–58. doi:10.1128/jvi.01053-12

Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. *S.* & Lander, E.*S.* (2013). Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell,* 154, 240-51. doi:10.1016/j.cell.2013.06.009

Halfmann, R., Jarosz, *D.* F., Jones, *S.* K., Chang, A., Lancaster, A. K. et al. (2012). Prions are a common mechanism for phenotypic inheritance in wild yeasts. *Nature,* 482, 363–8. doi:10.1038/nature10875

Hancock, J. M., Tautz, *D.* & Dover, G.A. (1988). Evolution of the secondary structures and compensatory mutations of the ribosomal RNAs of *Drosophila melanogaster*. *Mol. Biol. Evol.,* 5, 393–414.

Harrell, L., Melcher, U. & Atkins, J.F. (2002). Predominance of six different hexanucleotide recoding signals 3′ of read-through stop codons. *Nucleic Acids Res.*, 30, 2011–17. doi:10.1093/nar/30.9.2011

Hatin, I., Fabret, C., Namy, O., Decatur, W. A. & Rousset, J. (2007). Fine-tuning of translation termination efficiency in *Saccharomyces cerevisiae* involves two factors in close proximity to the exit tunnel of the ribosome. *Genetics,* 177, 1527–37. doi:10.1534/genetics.107.070771

Hubisz, M. J., Pollard, K. S. & Siepel, A. (2011). PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinform.,* 12, 41–51. doi:10.1093/bib/bbq072

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9, 90–95. doi:10.1109/MCSE.2007.55

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science,* 324, 218–23. doi:10.1126/science.1168978

Ingolia, N. T., Lareau, L. F. & Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell,* 147, 789–802. doi:10.1016/j.cell.2011.10.002

Jordan, B. R. (1975). Demonstration of intact 26 S ribosomal RNA molecules in *Drosophila* cells. *J. Mol. Biol.*, 98, 277–80. doi:10.1016/s0022-2836(75)80117-3

Jordan, B. R., Jourdan, R. & Jacq, B. (1976). Late steps in the maturation of *Drosophila* 26 S ribosomal RNA: generation of 5-8 S and 2 S RNAs by cleavages occurring in the cytoplasm. *J. Mol. Biol.*, 101, 85–105. doi:10.1016/0022-2836(76)90067-x

Jungreis, I., Lin, M. F., Spokony, R., Chan, C. S., Negre, N. et al. (2011). Evidence of abundant stop codon readthrough in *Drosophila* and other metazoa. *Genome Res,* 21, 2096–113. doi:10.1101/gr.119974.110

Kaiser, C. A. & Botstein, *D.* (1990). Efficiency and diversity of protein localization by random signal sequences. *Mol. Cell Biol.,* 10, 3163–73.

Kaiser, C. A., Preuss, D., Grisafi, P. & Botstein, *D.* (1987). Many random sequences functionally replace the secretion signal sequence of yeast invertase. *Science,* 235, 312–7. doi:10.1126/science.3541205

Karijolich, J. & Yu, Y. (2011). Converting nonsense codons into sense codons by targeted pseudouridylation. *Nature,* 474, 395–8. doi:10.1038/nature10165

Keeling, K. M., Lanier, J., Du, M., Salas-Marco, J., Gao, L. et al. (2004). Leaky termination at premature stop codons antagonizes nonsense-mediated mRNA decay in *S. cerevisiae*. RNA, 10, 691–703. doi:10.1261/rna.5147804

Klagges, B. R., Heimbeck, G., Godenschwege, T. A., Hofbauer, A., Pflugfelder, G. O. et al. (1996). Invertebrate synapsins: a single gene codes for several isoforms in Drosophila. *J. Neurosci.*, 16, 3154–65.

Kopczynski, J. B., Raff, A. C. & Bonner, J.J. (1992). Translational readthrough at nonsense mutations in the HSF1 gene of *Saccharomyces cerevisiae*. *Mol. Gen. Genet.,* 234, 369-78. doi:10.1007/bf00538696

Kosugi, S., Hasebe, M., Tomita, M. & Yanagawa, H. (2009). Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs. *Proc. Natl. Acad. Sci. USA*, 106, 10171–6. doi:10.1073/pnas.0900604106

Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, 305, 567–80. doi:10.1006/jmbi.2000.4315

Kryukov, G. V., Castellano, S., Novoselov, S. V., Lobanov, A. V., Zehtab, O. et al. (2003). Characterization of mammalian selenoproteomes. *Science,* 300, 1439–43. doi:10.1126/science.1083516

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.,* 10, R25. doi:10.1186/gb-2009-10-3-r25

Lasko, P. (2011). Posttranscriptional regulation in *Drosophila* oocytes and early embryos. *Wiley Interdiscip. Rev. RNA,* 2, 408–16. doi:10.1002/wrna.70

Lazarowitz, S. G. & Robertson, H.D. (1977). Initiator regions from the small size class of reovirus messenger RNA protected by rabbit reticulocyte ribosomes. *J. Biol. Chem.,* 252, 7842–9.

Li, G. P. & Rice, C.M. (1989). Mutagenesis of the in-frame opal termination codon preceding nsP4 of Sindbis virus: studies of translational readthrough and its effect on virus replication. *J. Virol.*, 63, 1326–37.

Lin, M. F., Carlson, J. W., Crosby, M. A., Matthews, B. B., Yu, C. et al. (2007). Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. *Genome Res.*, 17, 1823–36. doi:10.1101/gr.6679507

Lin, M. F., Jungreis, I. & Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, 27, i275–82. doi:10.1093/bioinformatics/btr209

Maurer-Stroh, S. & Eisenhaber, F. (2005). Refinement and prediction of protein prenylation motifs. *Genome Biol.,* 6, R55. doi:10.1186/gb-2005-6-6-r55

McCaughan, K. K., Brown, C. M., Dalphin, M. E., Berry, M. J. & Tate, W.P. (1995). Translational termination efficiency in mammals is influenced by the base following the stop codon. *Proc. Natl. Acad. Sci. USA*., 92, 5431-5. doi:10.1073/pnas.92.12.5431

McQuilton, P., St Pierre, S. E., Thurmond, J. (2012). FlyBase 101—the basics of navigating FlyBase. *Nucleic Acids Res.,* 40, D706–14. doi:10.1093/nar/gkr1030

Meijer, H. A. & Thomas, A.A.M. (2002). Control of eukaryotic protein synthesis by upstream open reading frames in the 5′-untranslated region of an mRNA. *Biochem. J.,* 367, 1–11.

Michel, A. M., Choudhury, K. R., Firth, A. E., Ingolia, N. T., Atkins, J. F. et al. (2012). Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res.*, 22, 2219–29. doi:10.1101/gr.133249.111

Mottagui-Tabar, S., Tuite, M. F. & Isaksson, L.A. (1998). The influence of 5′ codon context on translation termination in *Saccharomyces cerevisiae*. *Eur. J. Biochem.,* 257, 249–54. doi:10.1046/j.1432-1327.1998.2570249.x

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

Namy, O., Duchateau-Nguyen, G., Hatin, I., Hermann-Le Denmat, S., Termier, M. et al. (2003). Identification of stop codon readthrough genes in *Saccharomyces cerevisiae*. *Nucleic Acids Res.,* 31, 2289–96. doi:10.1093/nar/gkg330

Napthine, S., Yek, C., Powell, M. L., Brown, T. *D.* K. & Brierley, I. (2012). Characterization of the stop codon readthrough signal of Colorado tick fever virus segment 9 RNA. *RNA,* 18, 241–52. doi:10.1261/rna.030338.111

Neuberger, G., Maurer-Stroh, S., Eisenhaber, B., Hartig, A. & Eisenhaber, F. (2003). Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *J. Mol. Biol.*, 328, 581–92. doi:10.1016/s0022-2836(03)00319-x

Oh, E., Becker, A. H., Sandikci, A., Huber, D., Chaba, R. et al. (2011). Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell,* 147, 1295–308. doi:10.1016/j.cell.2011.10.044

Pavlakis, G. N., Jordan, B. R., Wurst, R. M. & Vournakis, J.N. (1979). Sequence and secondary structure of *Drosophila melanogaster* 5.8 S and 2 S rRNAs and of the processing site between them. *Nucleic Acids Res.,* 7, 2213–38. doi:10.1093/nar/7.8.2213

Pisarev, A. V., Kolupaeva, V. G., Yusupov, M. M., Hellen, C. U. T. & Pestova, T.V. (2008). Ribosomal position and contacts of mRNA in eukaryotic translation initiation complexes. *EMBO J.,* 27, 1609–21. doi:10.1038/emboj.2008.90

Qin, X., Ahn, S., Speed, T. P. & Rubin, G.M. (2007). Global analyses of mRNA translational control during early *Drosophila* embryogenesis. *Genome Biol.,* 8, R63. doi:10.1186/gb-2007-8-4-r63

Ramaswami, G., Zhang, R., Piskol, R., Keegan, L. P., Deng, P. et al. (2013). Identifying RNA editing sites using RNA sequencing data alone. Nat .Methods, 10, 128–32. doi:10.1038/nmeth.2330

Robinson, *D.* N. & Cooley, L. (1997). Examination of the function of two *kelch* proteins generated by stop codon suppression. *Development,* 124, 1405-17.

Skabkin, M. A., Skabkina, O. V., Hellen, C. U. T. & Pestova, T.V. (2013). Reinitiation and other unconventional posttermination events during eukaryotic translation. *Mol. Cell,* 51, 249–64. doi:10.1016/j.molcel.2013.05.026

Skuzeski, J. M., Nichols, L. M., Gesteland, R. F. & Atkins, J.F. (1991). The signal for a leaky UAG stop codon in several plant viruses includes the two downstream codons. *J. Mol. Biol*., 218, 365–73. doi:10.1016/0022-2836(91)90718-l

Steneberg, P. & Samakovlis, C. (2001). A novel stop codon readthrough mechanism produces functional Headcase protein in *Drosophila* trachea. *EMBO Rep.,* 2, 593–7. doi:10.1093/embo-reports/kve128

Steneberg, P., Englund, C., Kronhamn, J., Weaver, T. A. & Samakovlis, C. (1998). Translational readthrough in the hdc mRNA generates a novel branching inhibitor in the drosophila trachea. *Genes Dev.,* 12, 956–67. doi:10.1101/gad.12.7.956

Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V. T. K., Hein, M. Y. et al. (2012). Decoding human cytomegalovirus. *Science,* 338, 1088–93. doi:10.1126/science.1227919

Tautz, D., Hancock, J. M., Webb, *D.* A., Tautz, C. & Dover, G.A. (1988). Complete sequences of the rRNA genes of *Drosophila melanogaster*. *Mol. Biol. Evol.,* 5, 366–76.

Torabi, N. & Kruglyak, L. (2011). Variants in *SUP45* and *TRM10* underlie natural variation in translation termination efficiency in *Saccharomyces cerevisiae*. *PLoS Genet.,* 7, e1002211. doi:10.1371/journal.pgen.1002211

Torabi, N. & Kruglyak, L. (2012). Genetic basis of hidden phenotypic variation revealed by increased translational readthrough in yeast. *PLoS Genet.,* 8, e1002546. doi:10.1371/journal.pgen.1002546

True, H. L. & Lindquist, S.L. (2000). A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature,* 407, 477–83. doi:10.1038/35035005

Tuite, M. F. & Cox, B.S. (2007). The genetic control of the formation and propagation of the *[PSI+]* prion of yeast. *Prion,* 1, 101–9. doi:10.4161/pri.1.2.4665

Wills, N. M., Gesteland, R. F. & Atkins, J.F. (1991). Evidence that a downstream pseudoknot is required for translational read-through of the Moloney murine leukemia virus gag stop codon. *Proc. Natl. Acad. Sci. USA*, 88, 6991–5. doi:10.1073/pnas.88.16.6991

Xue, F. & Cooley, L. (1993). *kelch* encodes a component of intercellular bridges in *Drosophila* egg chambers. *Cell,* 72, 681–93. doi:10.1016/0092-8674(93)90397-9

Yamaguchi, Y., Hayashi, A., Campagnoni, C. W., Kimura, A., Inuzuka, T. et al. (2012). L-MPZ, a novel isoform of myelin P0, is produced by stop codon readthrough. *J. Biol. Chem.,* 287, 17765–76. doi:10.1074/jbc.m111.314468

Yoshinaka, Y., Katoh, I., Copeland, T. *D.* & Oroszlan, *S.* (1985). Translational readthrough of an amber termination codon during synthesis of feline leukemia virus protease. *J. Virol.*, 55, 870–3.

Appendix

Supplemental files

The raw (as FastQ files) and processed sequencing data (wiggle files) are available in NCBI's Gene Expression Omnibus (Edgar et al., 2002) under GEO series accession number GSE49197 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49197). Supplemental tables 1–4 are available at Dryad (doi:10.5061/dryad.6nr73; Dunn et al., 2013). Additional files are described below:

Supplemental table 1. Gene expression measurements in 0–2 hour embryos and S2 cells.

Source data for Figs. 1 and 2, as well as their supplements

Supplemental table 2. Readthrough statistics for Drosophila melanogaster.

Source data for Figs. 3, 4, and 6, as well as their supplements, and annotations of readthrough events in *Drosophila melanogaster*

Supplemental table 3. Readthrough statistics for Saccharomyces cerevisiae.

Source data for Fig. 5 and annotations of readthrough events in *[psi-]* W303 yeast

Supplemental table 4. Readthrough statistics for human foreskin fibroblasts.

Source data for Fig. 5 and annotations of readthrough events in human foreskin fibroblasts

Supplemental file 1. Alignment statistics.

Provides statistics on read alignments by sample and genomic region (e.g. CDS, 5′ UTR, 3′ UTR, intergenic, et c), as well as by sample and alignment type (e.g. chromosomal, spliced, unaligned)

Supplemental file 2. Oligonucleotides used in this study.

For readers who wish to implement the *Drosophila* ribosome profiling protocol

Supplemental figures

Figure 1, supplement 1. Digestion with micrococcal nuclease yields a robust ribosome profiling assay

a. Digestion of polysomes with RNase I degrades ribosomes. A lysate was made from S2 cells using a previous version of our protocol. Aliquots of this lysate were digested with increasing amounts of RNase I, and resolved on 10–50% sucrose gradients. As amounts of RNase I increase, the heights of all peaks – including the monosomal (80S) peak – decrease before polysomes are fully resolved to monosomes.

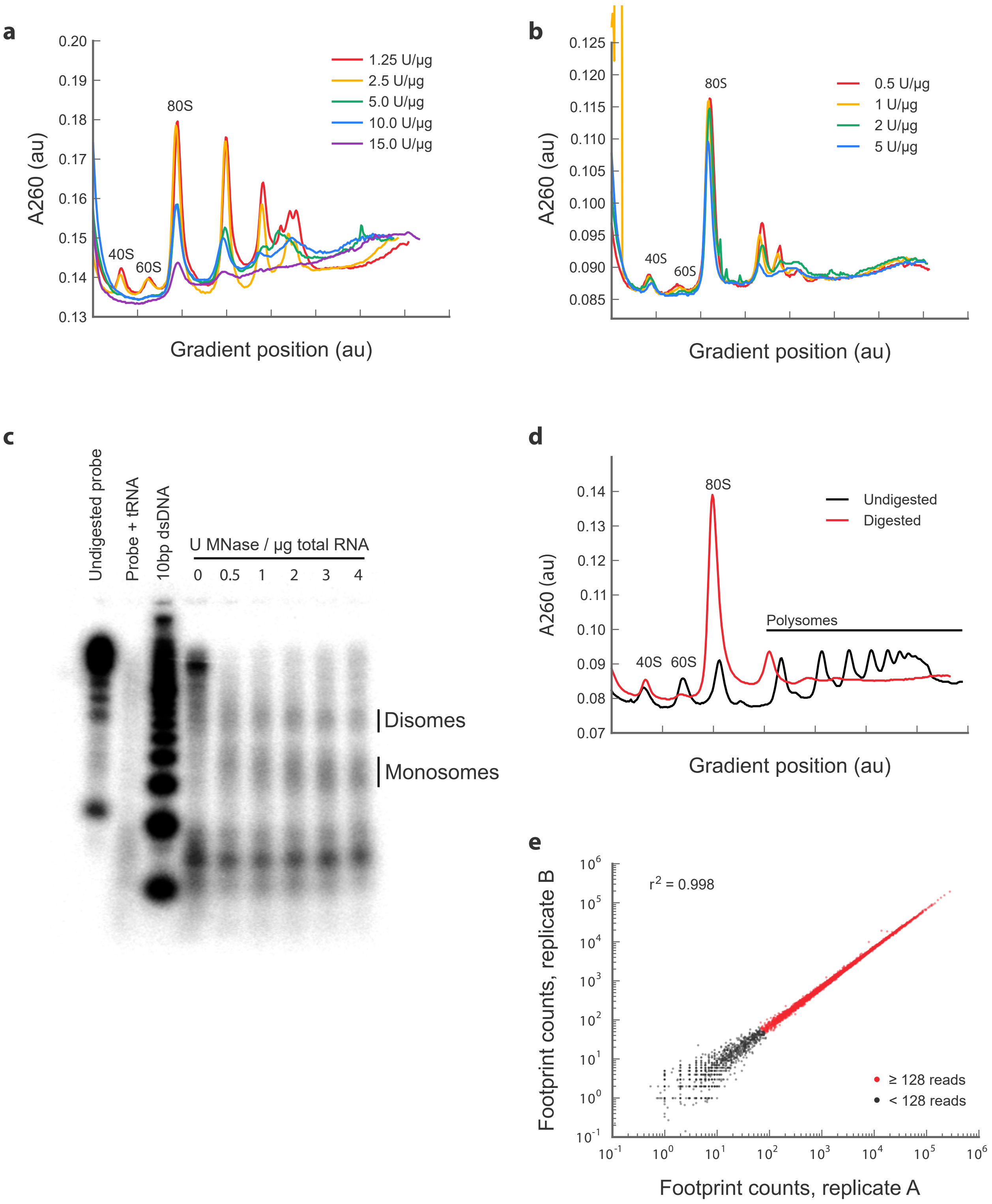
b. As in (a), but using micrococcal nuclease (MNase) and our current protocol. From 0.5 to 2 U MNase / µg total RNA, monosomes are resolved with no reduction in the size of the monosome peak. This result indicates that Drosophila ribosomes are stable to MNase over a broad range of concentrations, whereas the mRNA between ribosomes is digested.

c. Ribosome protection assay. A 320 nucleotide fragment of enolase (FlyBase accession: FBgn0000579) was amplified using oligos oJGD123 & oJGD124 (see Supplemental file 2). A body-labeled probe against this sequence was transcribed from this template using ${}^{\alpha 32}$P-UTP and the T7 MaxiScript kit (Ambion). S2 cell lysates were prepared as in methods and aliquoted. Aliquots were digested as in methods, except with 0, 0.5, 1, 2, 3 or 4 U MNase / µg total RNA. Monosomes were sedimented through a sucrose cushion, resuspended in 600 µl 10 mM Tris pH 7.0, and their RNAs extracted as in methods. Concentrations were determined using a NanoDrop spectrophotometer. 5 µg of each sample was hybridized to 50,000 CPM of probe overnight at 42°C. Single-stranded regions were digested with RNase A/T1 and the remaining footprint:probe duplexes detected using the mirVana micro-RNA detection kit (Ambion), resolved on a 15% TBE-urea gel (Invitrogen), and visualized on a storm phosphorimager (Molecular Dynamics). For size markers, we end-labeled the Novex 10 bp dsDNA ladder (Invitrogen) with ^{32}P. Over two-fold range of nuclease concentrations, the ~30 nt peak corresponding to ribosome-protected footprints remains constant in size and intensity, indicating a lack of degradation consistent with the unchanged monosome peak height across this range of digestion conditions in (b). Also visible is a roughly 60 nt band which we infer to be protected by adjacent ribosomes (disomes) that sterically exclude the nuclease. This interpretation is consistent with the presence of a small disome peak in digested samples (c.f. panels b & d, and Figure 1a).

d. A polysome lysate was prepared from S2 cells and resolved in 10–50% sucrose gradients, with or without prior digestion with 3 U MNase / µg total RNA.

e. A culture of S2 cells was split into aliquots and processed using our current protocol as if they were independent samples. Total counts aligning to the coding region of each gene were tabulated in each replicate. Genes sharing at least 128 footprint counts between replicates (red) are well-correlated, demonstrating the assay is robust (see full discussion in Figure 1, supplement 2).

Source data in supplemental table 1


a
0.20
0.19
0.18
0.17
0.16
0.15
0.14
0.13
A260 (au)
40S
60S
80S
1.25 U/µg
2.5 U/µg
5.0 U/µg
10.0 U/µg
15.0 U/µg
Gradient position (au)
b
0.125
0.120
0.115
0.110
0.105
0.100
0.095
0.090
0.085
A260 (au)
40S
60S
80S
0.5 U/µg
1 U/µg
2 U/µg
5 U/µg
Gradient position (au)
c
Undigested probe
Probe + tRNA
10bp dsDNA
U MNase / µg total RNA
0
0.5
1
2
3
4
Disomes
Monosomes
d
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.07
A260 (au)
40S
60S
80S
Undigested
Digested
Polysomes
Gradient position (au)
e
10^6
10^5
10^4
10^3
10^2
10^1
10^0
10^{-1}
Footprint counts, replicate B
$r^2 = 0.998$
● ≥ 128 reads
● < 128 reads
10^{-1}
10^0
10^1
10^2
10^3
10^4
10^5
10^6
Footprint counts, replicate A

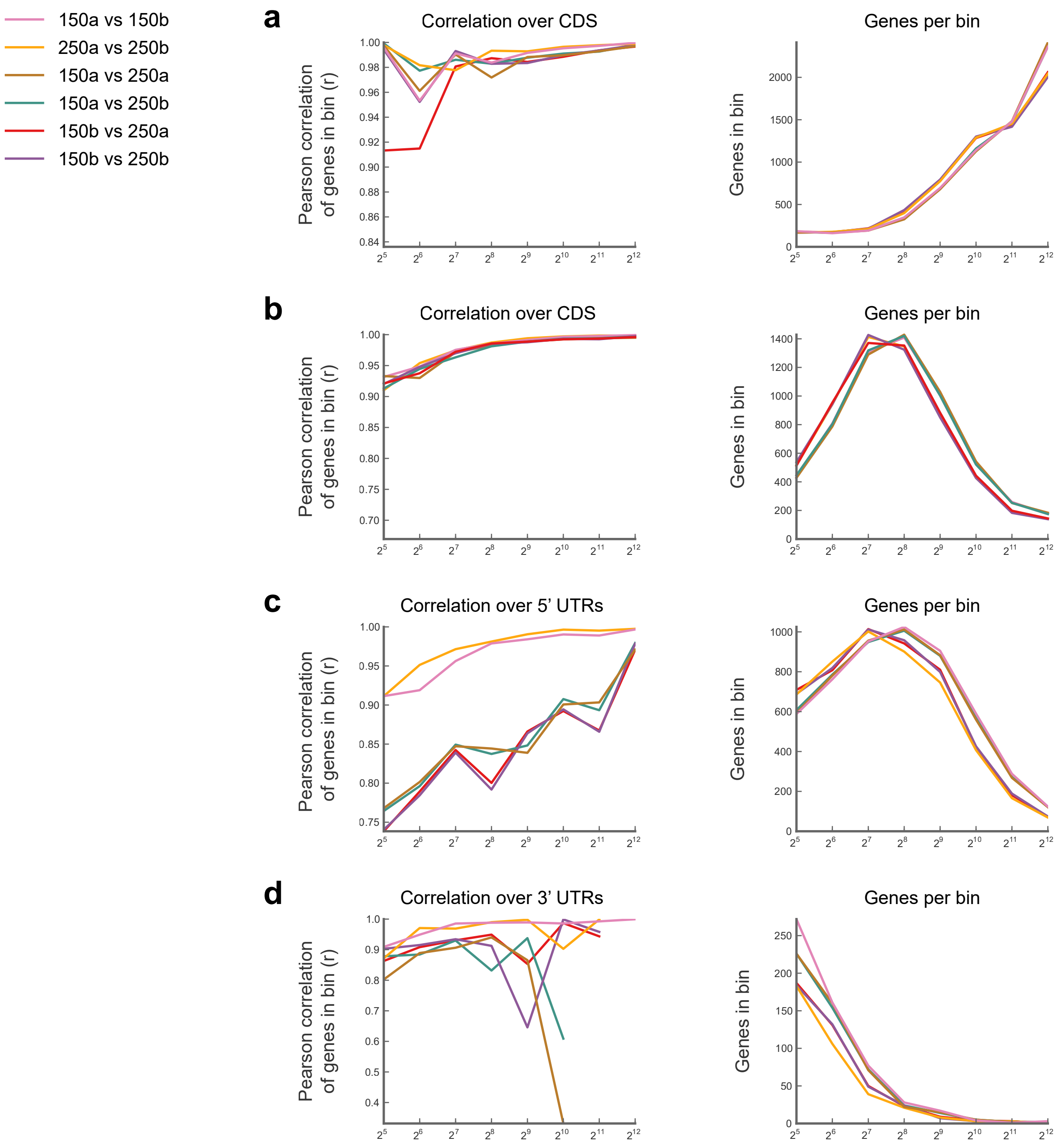

150a vs 150b
250a vs 250b
150a vs 250a
150a vs 250b
150b vs 250a
150b vs 250b
a
Correlation over CDS
Pearson correlation of genes in bin (r)
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
2^{5} 2^{6} 2^{7} 2^{8} 2^{9} 2^{10} 2^{11} 2^{12}
Genes per bin
Genes in bin
2000
1500
1000
500
0
2^{5} 2^{6} 2^{7} 2^{8} 2^{9} 2^{10} 2^{11} 2^{12}
b
Correlation over CDS
Pearson correlation of genes in bin (r)
1.00
0.95
0.90
0.85
0.80
0.75
0.70
2^{5} 2^{6} 2^{7} 2^{8} 2^{9} 2^{10} 2^{11} 2^{12}
Genes per bin
Genes in bin
1400
1200
1000
800
600
400
200
0
2^{5} 2^{6} 2^{7} 2^{8} 2^{9} 2^{10} 2^{11} 2^{12}
c
Correlation over 5' UTRs
Pearson correlation of genes in bin (r)
1.00
0.95
0.90
0.85
0.80
0.75
2^{5} 2^{6} 2^{7} 2^{8} 2^{9} 2^{10} 2^{11} 2^{12}
Genes per bin
Genes in bin
1000
800
600
400
200
0
2^{5} 2^{6} 2^{7} 2^{8} 2^{9} 2^{10} 2^{11} 2^{12}
d
Correlation over 3' UTRs
Pearson correlation of genes in bin (r)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
2^{5} 2^{6} 2^{7} 2^{8} 2^{9} 2^{10} 2^{11} 2^{12}
Genes per bin
Genes in bin
250
200
150
100
50
0
2^{5} 2^{6} 2^{7} 2^{8} 2^{9} 2^{10} 2^{11} 2^{12}

Figure 1, supplement 2. Effects of buffer conditions upon reproducibility.

A culture of S2 cells was divided into four aliquots, and each aliquot carried through the entire ribosome profiling procedure as an independent sample. Two aliquots ("150a" and "150b") were processed using our standard lysis buffer with 150 mM Na^+ and 5 mM Mg^{++} and digested with 3 U MNase / µg total RNA as described in methods. The other two ("250a" and "250b") were processed using an earlier version of our protocol, in which our lysis buffer contained 250 mM Na^+ and 15 mM Mg^{++}, and in which we digested lysates with 30 U MNase / µg total RNA.

We then calculated ribosome density for each gene over coding regions (a), 5' UTRs (c) and 3'UTRs (d), performed pairwise comparisons between samples. For each comparison, we binned genes based upon the summed number of reads in samples A and B, and calculated the correlation coefficients (Pearson's *r*) for the RPKM values for each gene in each bin (left column). The number of genes in each bin are also shown (right column). Correlations between samples for coding regions are robust across buffer regions (a), though some salt-dependence is visible in 5' and 3' UTRs (c, d). (b) is as in (a), but using only 10% of the reads. The high correlation observed at our 128-minimum-count threshold is therefore not a function of the number of genes in each bin

Source data in supplemental table 1

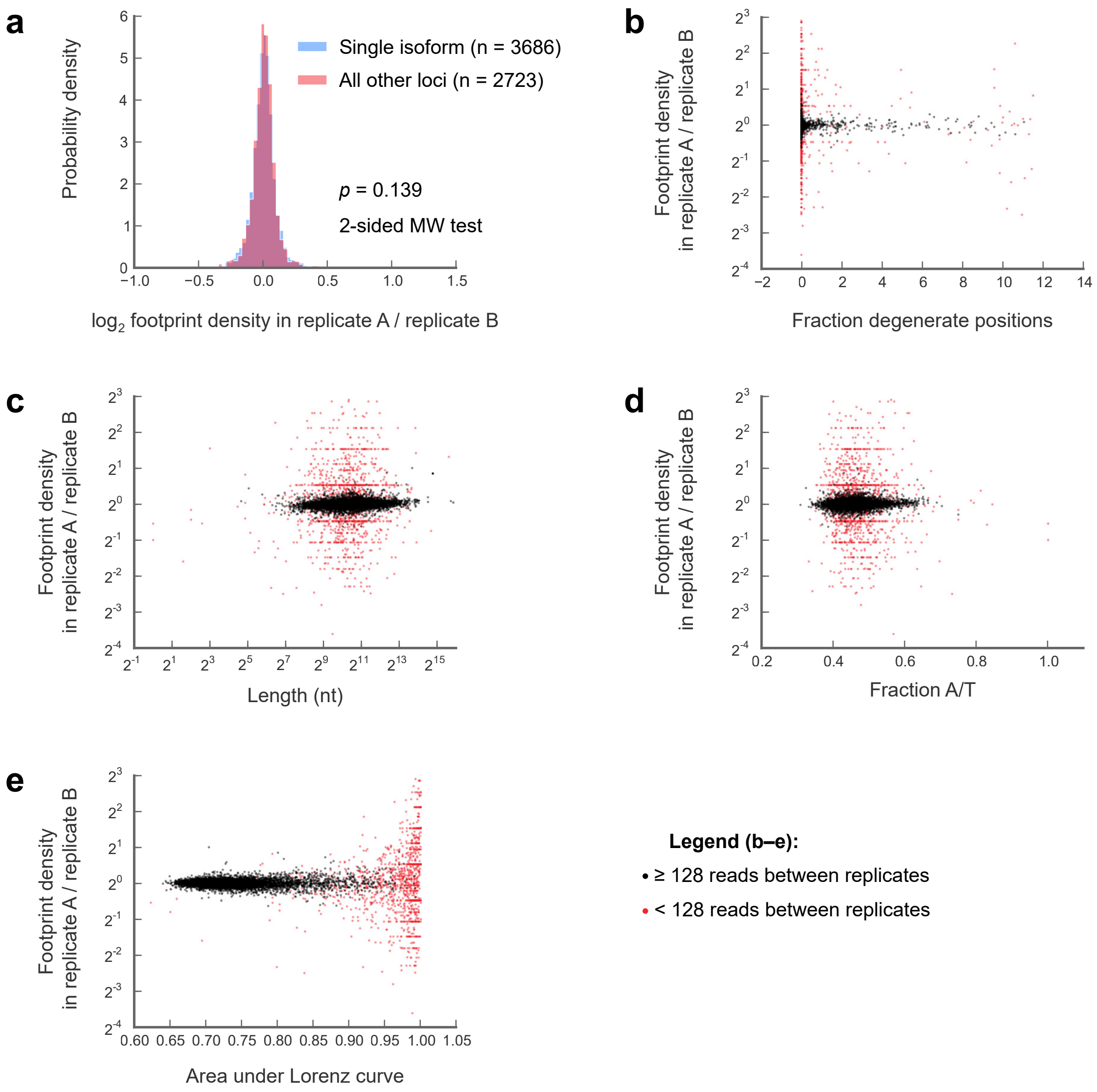


Figure 1, supplement 3. Variability in ribosome footprint density measurements are not correlated with isoform number, sequence degeneracy in the locus of interest, locus length, A/T content, or evenness of coverage

Comparisons are made between S2 cell technical replicates 150a and 150b (see Figure 1, supplement 2)

a. Variability of log2 fold-changes in ribosome footprint densities are no greater for multi-isoform loci (pink) than they are for single-isoform loci (blue)

b. Correlation of the fraction degenerate positions in each locus (see methods) with fold-changes in ribosome density between replicates at that locus. Loci with at least 128 counts between replicates are shown in black, those with less in red

c. As in (b), but correlation of length with inter-replicate fold-changes

d. As in (b), but correlation of A/T content with inter-replicate fold-changes

e. As in (b), but correlation of area under Lorenz curve with inter-replicate fold-changes

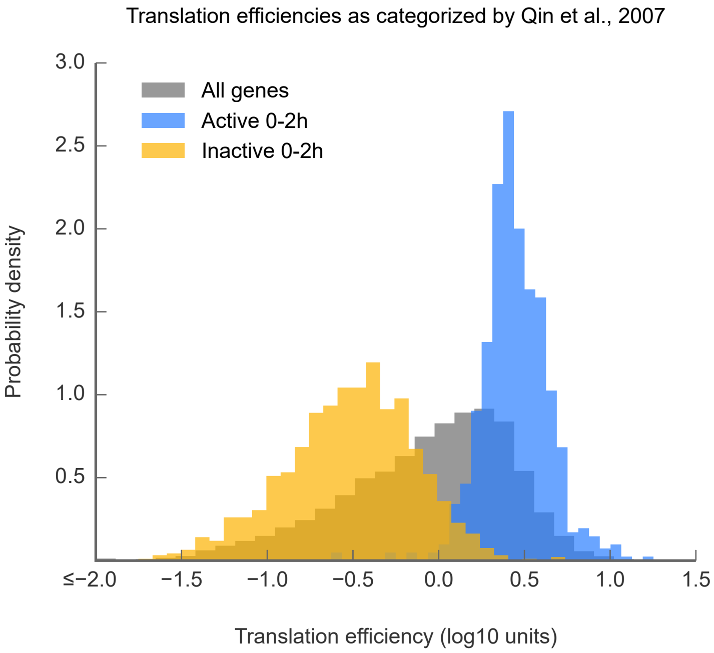


Figure 1, supplement 4. Ribosome profiling is consistent with earlier methods

Measurements of translation efficiency obtained via ribosome profiling are consistent with those made using semiquantitative polysome gradients. Histograms of translation efficiency for genes labeled by Qin et al., 2007 as active (blue) or inactive (yellow) in 0–2h embryos. All genes are shown in gray.

Source data in supplemental table 1

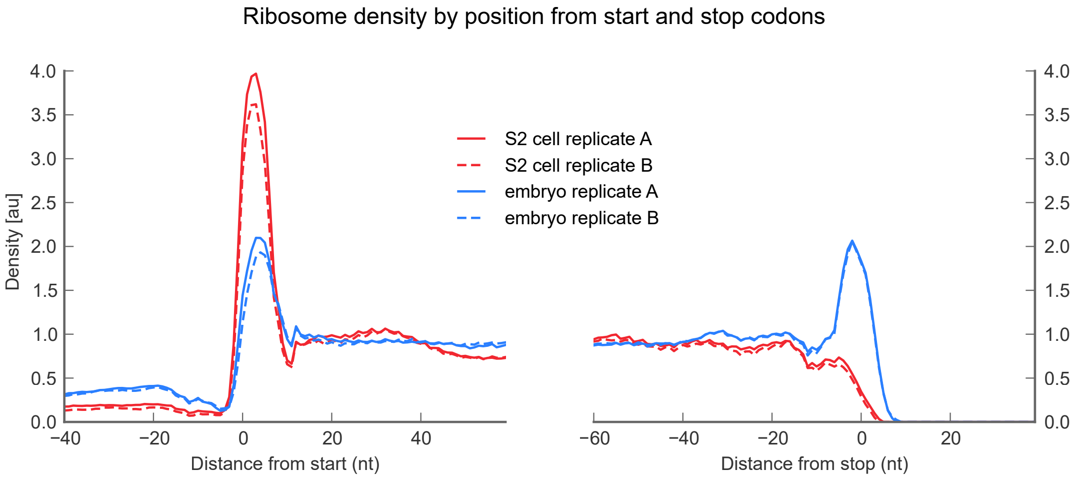


Figure 2, supplement 1
Ribosome density over start & stop codons

Ribosome density across the average gene or "metagene" reveals peaks of ribosome density at start and stop codons. For this analysis we included all genes that met the following criteria: a) all transcripts deriving from that gene had one annotated start codon (left panel) or stop codon (right panel), b) all transcripts deriving from that locus covered identical genomic positions over the region of interest (ROI) shown, c) all positions within the ROI were non-degenerate (see methods), and d) at least 10 reads were present in the coding subregion of the ROI. For each ROI meeting these criteria (2800–3200 ROI per sample), we generated a "coverage vector" tallying ribosome density at each nucleotide position. We then normalized each coverage vector to the mean number of footprint reads covering the annotated coding region in the ROI, excluding a 3-codon buffer flanking the start or stop codon to avoid bleedthrough from initiation or termination peaks. We then plotted the median value across all normalized coverage vectors at each position.

Peaks are visible in the start and stop codons of embryo samples. Consistent with our previous work, stop codon peaks are missing from S2 cell samples because terminating ribosomes release during our 2-minute treatment with translation inhibitors. They are present in our embryo samples, because these are flash-frozen and lysed in the presence of translation inhibitors which block termination as well as initiation and elongation.

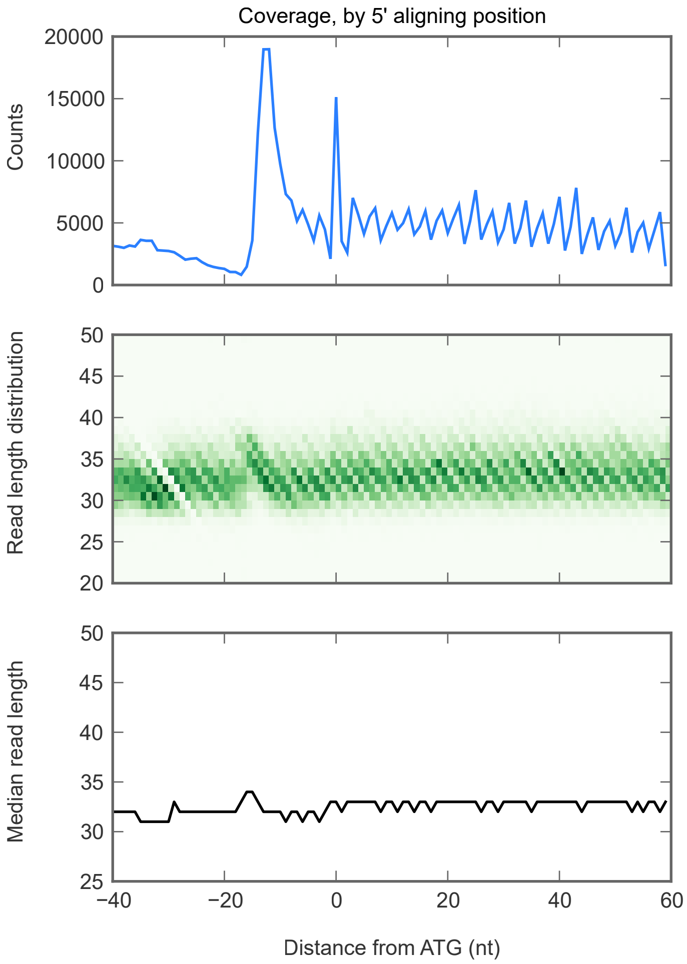


Figure 2, supplement 2
Read lengths are similar in 5′ UTRs and coding regions

We aggregated all ribosome-protected reads aligning to all genes with a single initiation codon, and in which all annotated isoforms cover the same genomic positions in the ROI shown. We plotted the following statistics as a function of the reads whose 5′ end mapped to each position on the x-axis. *Top:* number of reads (y-axis) aligning at each position. Because the 5′ end, rather than the P-site, is plotted, the peak of ribosome density is approximately 13 nucleotides 5′ of the start codon (position 0, x-axis). *Middle:* heatmap of read lengths (y-axis) as a function of position. *Bottom:* median read length (y-axis) at each position.

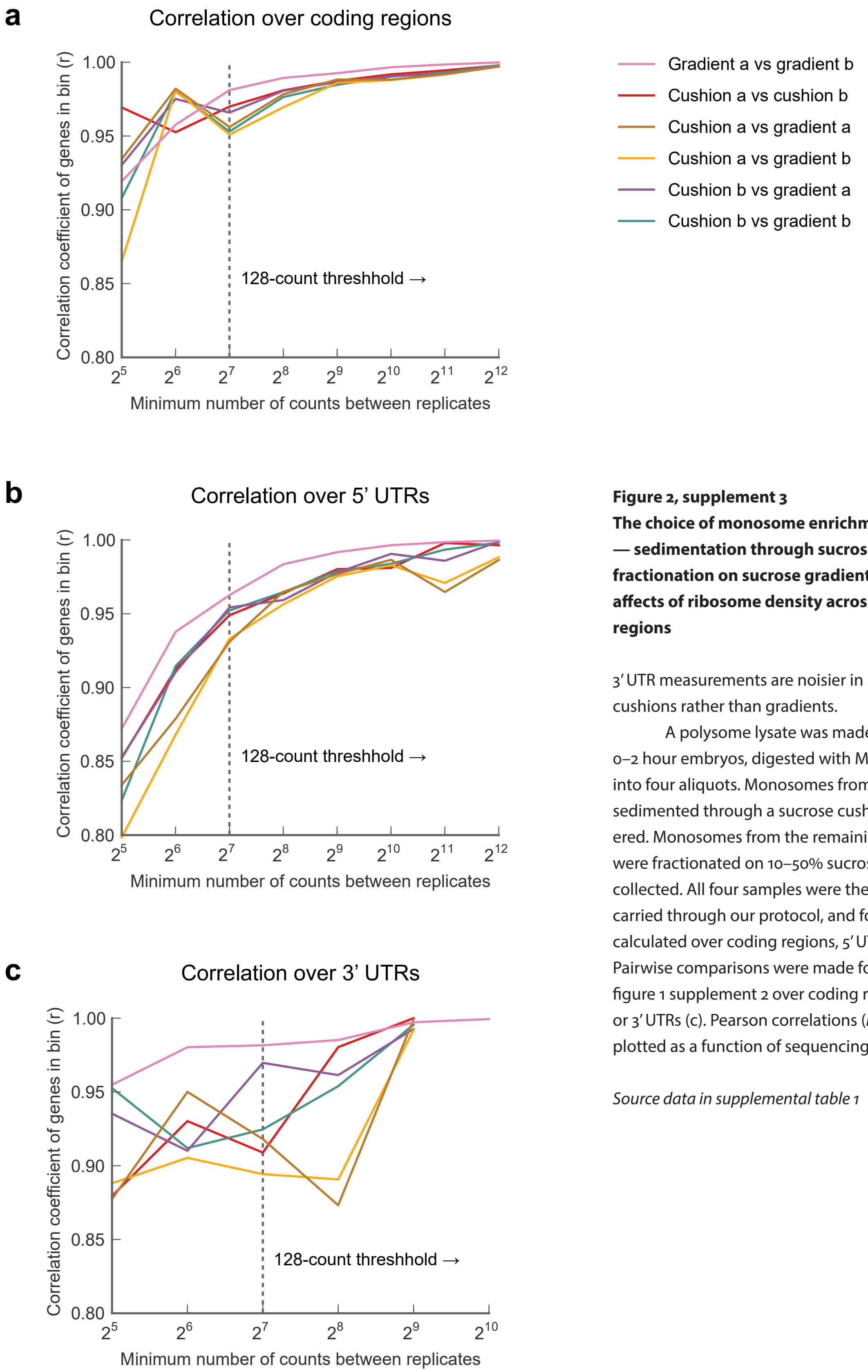


Figure 2, supplement 3
The choice of monosome enrichment technique — sedimentation through sucrose cushions or by fractionation on sucrose gradients — minimally affects of ribosome density across 5' UTRs and coding regions

3' UTR measurements are noisier in samples prepared on cushions rather than gradients.

A polysome lysate was made from collected 0–2 hour embryos, digested with MNase, and split into four aliquots. Monosomes from two aliquots were sedimented through a sucrose cushion and recovered. Monosomes from the remaining two aliquots were fractionated on 10–50% sucrose gradients and collected. All four samples were then independently carried through our protocol, and footprint density was calculated over coding regions, 5' UTRs, and 3' UTRs. Pairwise comparisons were made for each sample as in figure 1 supplement 2 over coding regions (a), 5' UTRs (b), or 3' UTRs (c). Pearson correlations (*r*) for the regions are plotted as a function of sequencing depth.

Source data in supplemental table 1

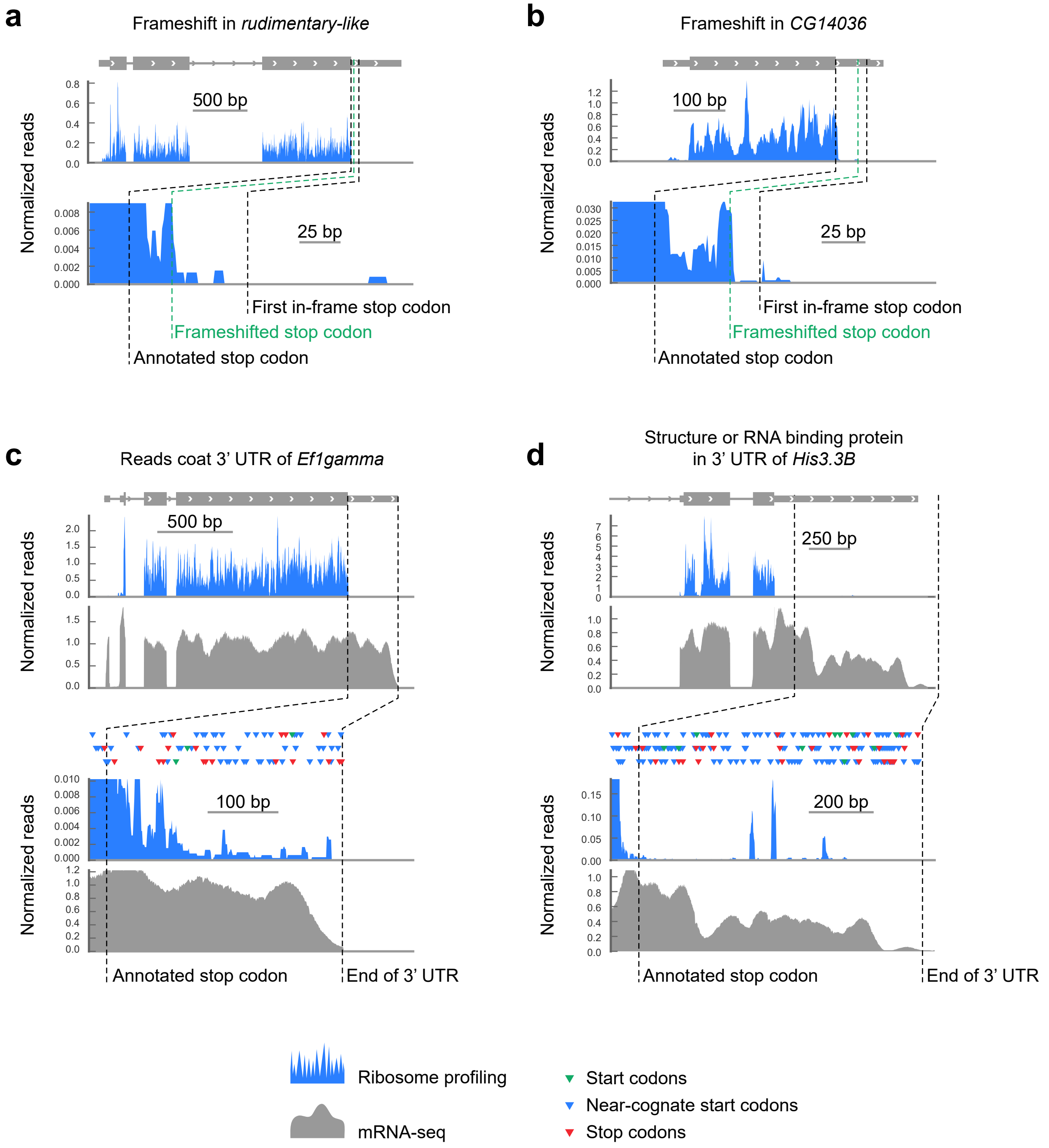


Figure 3, supplement 1. Examples of footprint density in 3´ UTRs attributed to sources other than readthrough

a and b. Transcripts exhibiting translation consistent with translation in alternate frames.

c. Footprint density, potentially caused by RNA binding proteins or structures, coats the 3´ UTR of EF1γ, passing through stop codons (red triangles) in all three frames.

d. The 3´ UTR of HIS3.3B contains highly localized read density consistent with the presence of an RNA binding protein or mRNA structure, but not with translation of an open reading frame. Colors as in (c).

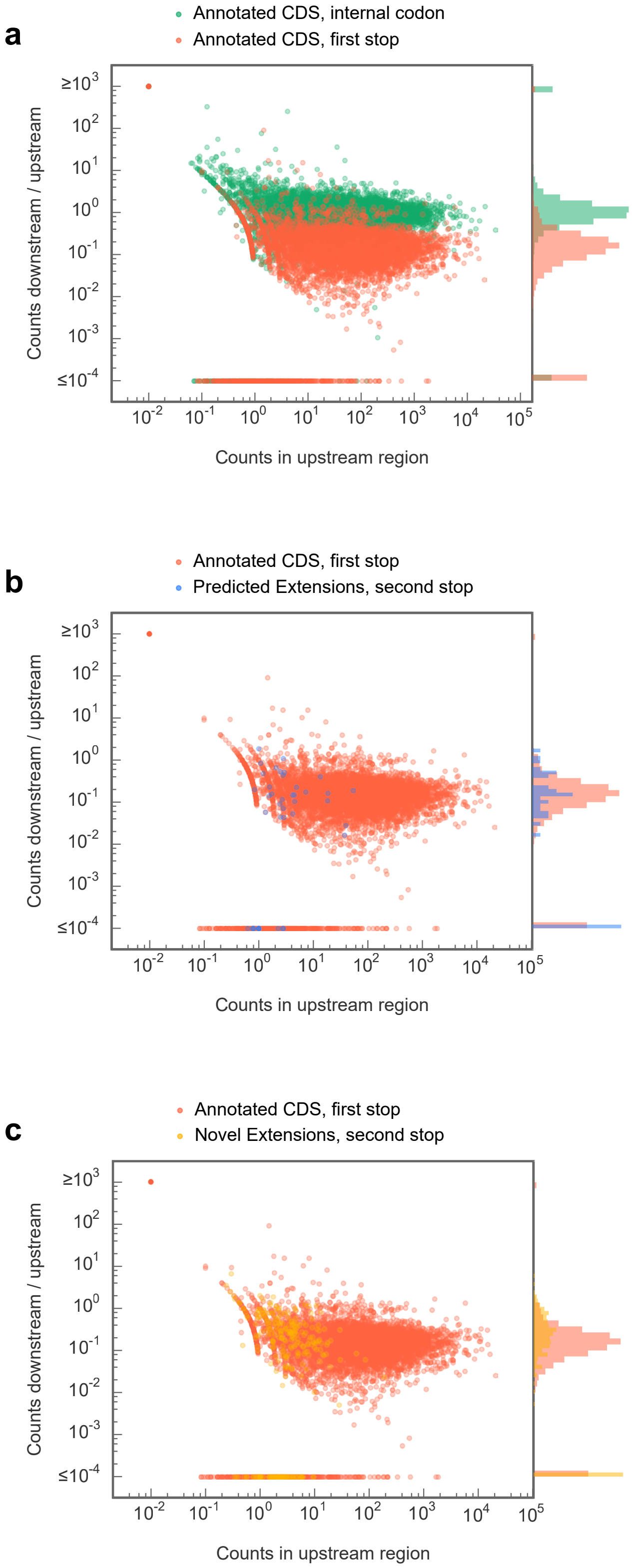


Figure 4, supplement 1. C-terminal extensions in Drosophila melanogaster show ribosome release typical of coding regions, but not of internal codons

For each region of interest, the total number of reads aligning to 5 codon windows immediately upstream of that codon were tabulated, and the ratio (downstream counts / upstream counts) plotted against the total number of counts in the upstream window.

a. Comparison of release scores for termination codons of annotated coding regions and form randomly-selected codons internal to (i.e. at least 10 codons from the annotated start or end) annotated coding regions.

b. As in (a), but stop codons that terminate predicted extensions are compared against those that terminate annotated coding regions.

c. As in (a) but stop codons that terminate novel extensions are compared against those that terminate annotated coding regions.

Source data in supplemental table 2

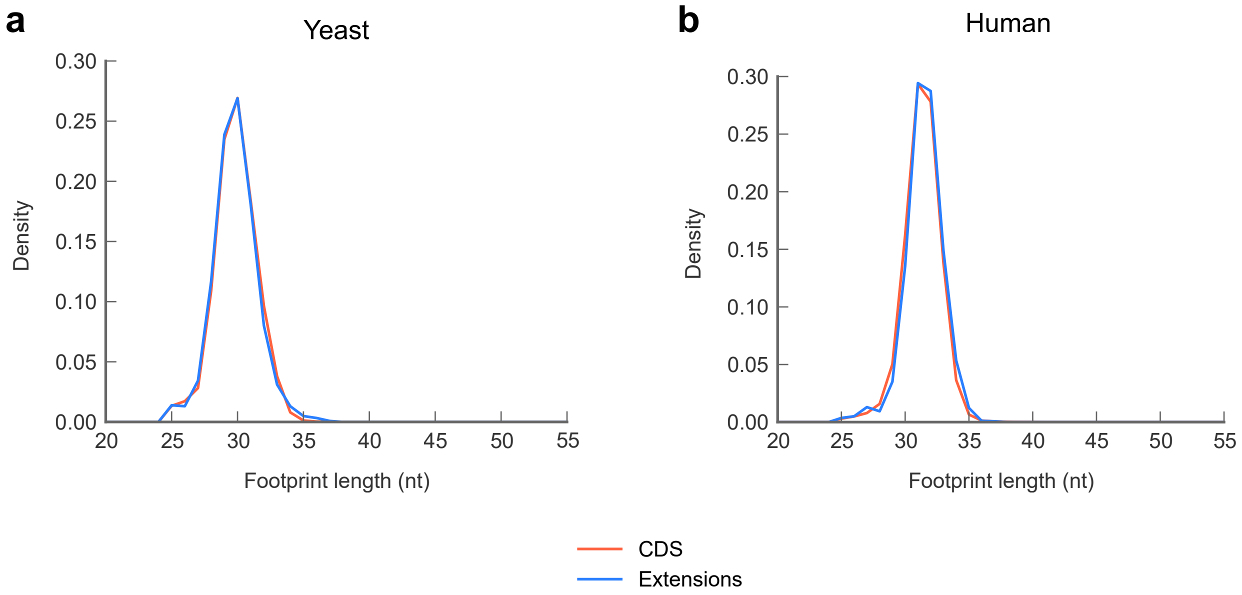


Figure 5, supplement 1
In yeast and human cell lines, reads mapping to C-terminal extensions are drawn from the same length distribution as reads mapping to coding regions

a. Length distributions of reads mapping to coding regions and extensions in yeast.

b. Length distributions of reads mapping to coding regions and extensions in human foreskin fibroblasts.

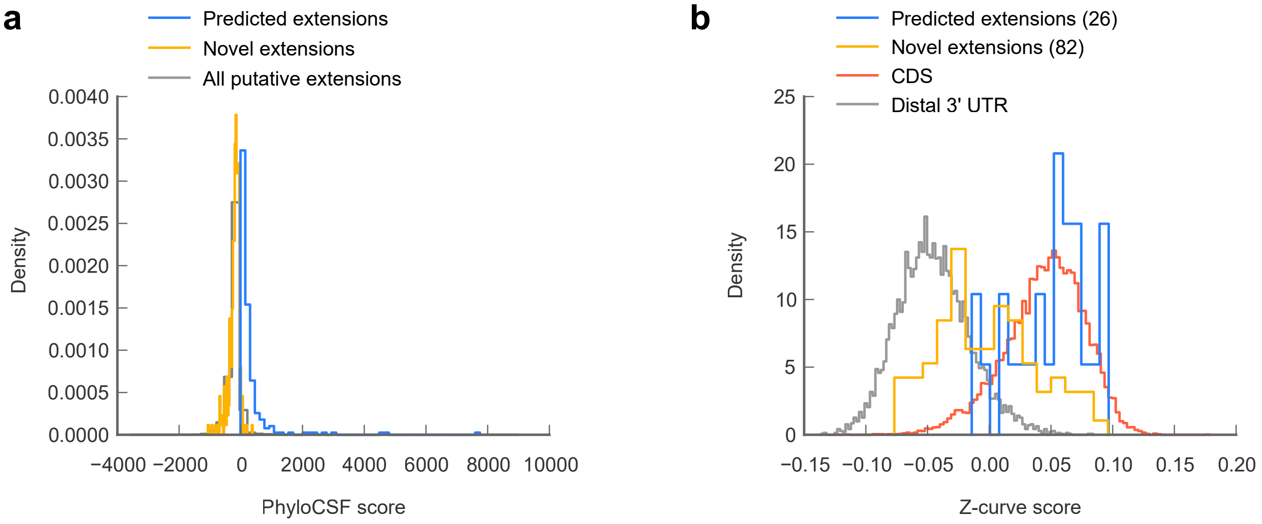


Figure 6, supplement 1. Novel C-terminal extensions in Drosophila melanogaster show signatures of selection within the melanogaster lineage

a. Histogram of PhyloCSF scores for C-terminal extensions. *Blue:* phylogenetically predicted extensions that were confirmed in our datasets. *Yellow:* unpredicted extensions discovered in our datasets. *Gray:* global distribution of all potential extensions. The distribution of novel extensions is not substantially different from the global distribution, suggesting that many of these extensions are not phylogenetically conserved beyond melanogaster.

b. A second Z-curve classifier was trained on 81-nucleotide windows of coding regions, and 81-nucleotide windows of distal 3´ UTRs, but excluding the last 50 bases of annotated UTR to remove potential effects of polyadenylation signals upon classifier scoring. As in figure 6B, predicted extensions overlay coding regions, and novel extensions display a significant shift in median from distal 3´ UTRs ($p = 3.81 \times 10^{-22}$, Mann-Whitney U test), indicating the shift identified in figure 6B is not due to polyadenylation signals.

Publishing agreement

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author

Date 22 December, 2015