doi:10.1038/nature18308

Translation readthrough mitigation

Joshua A. Arribere[1], Elif S. Cenik[1], Nimit Jain[2], Gaelen T. Hess[3], Cameron H. Lee[3], Michael C. Bassik[3] & Andrew Z. Fire[1,3]

A fraction of ribosomes engaged in translation will fail to terminate when reaching a stop codon, yielding nascent proteins inappropriately extended on their C termini. Although such extended proteins can interfere with normal cellular processes, known mechanisms of translational surveillance[1] are insufficient to protect cells from potential dominant consequences. Here, through a combination of transgenics and CRISPR–Cas9 gene editing in *Caenorhabditis elegans*, we demonstrate a consistent ability of cells to block accumulation of C-terminal-extended proteins that result from failure to terminate at stop codons. Sequences encoded by the 3′ untranslated region (UTR) were sufficient to lower protein levels. Measurements of mRNA levels and translation suggested a co- or post-translational mechanism of action for these sequences in *C. elegans*. Similar mechanisms evidently operate in human cells, in which we observed a comparable tendency for translated human 3′ UTR sequences to reduce mature protein expression in tissue culture assays, including 3′ UTR sequences from the hypomorphic 'Constant Spring' haemoglobin stop codon variant. We suggest that 3′ UTRs may encode peptide sequences that destabilize the attached protein, providing mitigation of unwelcome and varied translation errors.

Failure to terminate translation at a stop codon can lead to ribosomes translating into a 3′ UTR. In some cases, translation may proceed through the 3′ UTR and into the poly(A) tail, triggering a process termed 'nonstop decay' and destabilizing both the mRNA and nascent protein (reviewed in ref. 1). However, for the majority of 3′ UTRs, a stop codon is encountered before the poly(A) tail[2,3]. Readthrough events that encounter a subsequent termination codon are outside the scope of known translational surveillance pathways including nonstop[1]. Depending on the 3′ UTR and the frame in which the ribosome enters, the late stop codon can be several, tens, or even hundreds of codons into a 3′ UTR, producing variant proteins with potentially problematic C-terminal appendages. This issue is highlighted by several pathologies caused by late frameshifts or stop-codon mutations in which 3′-UTR-encoded C-terminal extensions effect protein mislocalization[4,5], aggregation[6,7], and instability[8–12], with severe consequences for organisms. Depending on sequence, genetic background, conditions, and organism, estimates of readthrough efficiency vary from <1% to 10% or more, posing a potential problem of nontrivial magnitude[10,13].

We investigated whether, and to what extent, 3′ UTR translation has an effect on gene expression using a fluorescent reporter system in *C. elegans*. Initially, we selected 3′ UTRs from three genes: *unc-54* (encoding a muscle myosin), *tbb-2* (a beta tubulin), and *rpl-14* (a ribosomal protein). For each gene, fusion of the 3′ UTR to a green fluorescent protein (GFP)-driven by the *myo-3* promoter resulted in robust fluorescence in body-wall muscle (Fig. 1a). Next, by mutating stop codons, we created GFP reporters for each gene, which caused translation to read past the normal termination point, terminating instead at a stop codon part-way through the 3′ UTR (Fig. 1b). In each case, the 'late stop' reporter accumulated substantially less GFP, with differences in signal of at least tenfold. As a control, a co-injected mCherry marker was robustly expressed in the same cells. We conclude that translation into the 3′ UTR can confer substantial loss of protein expression for at least these three 3′ UTRs in *C. elegans*.

To test whether translation into 3′ UTRs could confer a loss of protein expression more generally, a two-fluorescent-reporter system with each fluorophore transgene containing an identical 3′ UTR was used. Nine genes were chosen to reflect a variety of functions and expression levels: *rps-17* (small ribosomal subunit component), *r74.6* (*dom34/pelota* release factor homologue), *hlh-1* (muscle transcription factor), *eef-1A.1* (also known as *eft-3*, translation elongation factor), *myo-2* (a pharyngeal myosin), *mut-16* (involved in gene/transposon silencing), *bar-1* (a beta catenin), *daf-6* (involved in amphid morphogenesis), and *alr-1* (neuronal transcription factor). A criterion in choosing these genes was the presence (common for *C. elegans* genes, Extended Data Fig. 1) of an in-frame stop codon in the 3′ UTR at least 30 bases beyond the normal stop but upstream of known poly(A) sites. We fused the 3′ UTRs of each gene separately to GFP and mCherry, removing the canonical termination codon in the GFP construct. For each of the nine genes tested, the observed GFP signals were extremely faint, with raw GFP:mCherry fluorescence ratios of less than 0.1 (Fig. 1c, Extended Data Fig. 2). As a control, versions of the GFP reporter with the normal termination codon intact provided robust GFP expression, at least tenfold higher than the corresponding readthrough constructs (GFP:mCherry fluorescence ratios in the range of 0.3 to 0.9).

Several observations suggest how translation into 3′ UTRs might reduce protein levels. (1) Experiments with specific mutagenesis support a role for the eventual protein sequence. Shortening readthrough peptides (tested for *unc-54* and *tbb-2*) increased GFP expression (Fig. 2a). Extending this analysis, an equal-length non-synonymous substitution in the *unc-54* 3′ UTR restored GFP expression, whereas synonymous substitution with multiple base differences did not. (2) Mutagenesis analysis of constructs using a constant 3′ UTR reinforced the inference of peptide sequence as the primary determinant of GFP loss. We found that the nucleotide sequence between the normal termination codon and the first in-frame termination codon was sufficient to confer GFP loss if inserted at the end of the GFP coding region for *unc-54*, *tbb-2*, *hlh-1*, *daf-6*, *rps-20*, or *rps-30* (Fig. 2b). The *rps-30* readthrough region had the weakest effect on GFP, and was the shortest (nine amino acids). We performed further mechanistic dissection by synonymous variation of readthrough regions from *unc-54*, *tbb-2*, and *rps-20*, with GC contents from 35–60%, in some cases mutating >50% of bases. Each synonymously substituted variant conferred robust loss of GFP. (3) Decreased expression following translation into the 3′ UTR required peptide linkage between the upstream protein and the 3′-UTR-encoded segment. To assess the relationship between covalent linkage with the translated C-terminal peptide and the outcome for the larger protein, we took advantage of a picornavirus-derived oligopeptide sequence that causes cleavage and release of the nascent chain, after which ribosomes continue translation of the downstream sequence[14,15]. Insertion of the T2A peptide (EGRGSLLTCGDVEENPGP) between GFP and the *unc-54* 3′-UTR-encoded sequence rescued GFP expression, whereas an uncleavable T2A* point mutant did not (Fig. 2b). Restoration of GFP levels by T2A to the level of no-insert controls

[1]Department of Pathology, Stanford University School of Medicine, Stanford, California 94305, USA. [2]Department of Bioengineering, Stanford University, Stanford, California 94305, USA. [3]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA.

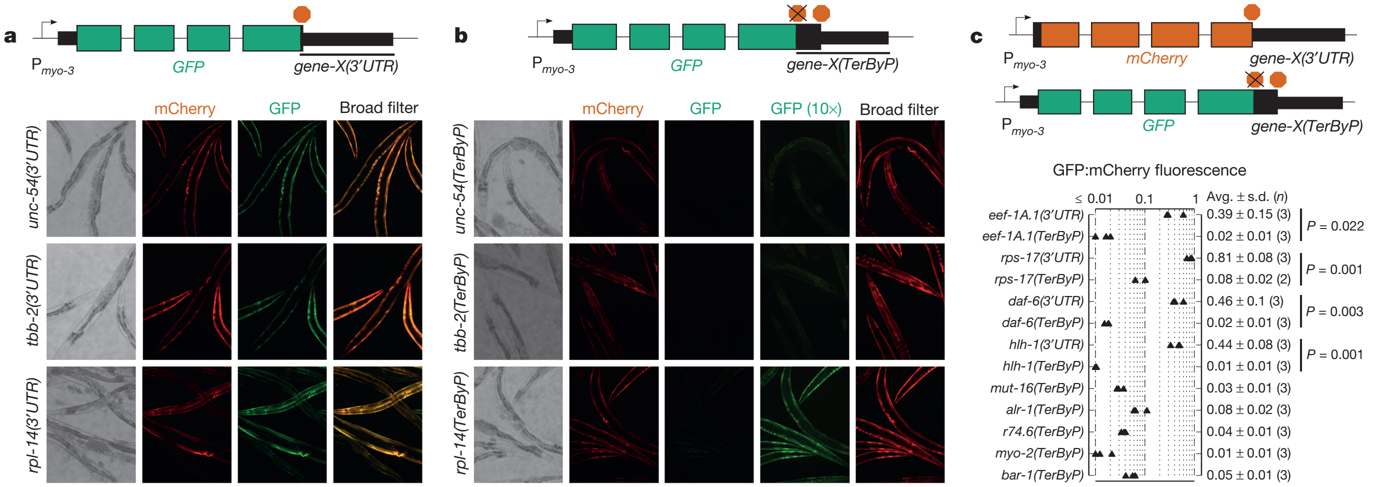


Figure 1 | Translation into 3′ UTRs results in substantial loss of protein expression. a, Dual fluorescence reporter assay to test expression with different 3′ UTRs. Transgenic arrays of each GFP construct were created using *pha-1* selection and mCherry (pCFJ104) as a coinjection marker. Broad filter detects GFP and mCherry signals simultaneously; deviation from yellow towards red or green shows more mCherry or GFP fluorescence, respectively. Three independent transgenic lines were made for each (two for *tbb-2(TerByP)*); transgenic lines with similar mCherry expression are shown. 200 ms exposure, 10 × objective. **b**, Dual fluorescence reporter assay to test expression of readthrough for different 3′ UTRs. The stop codon of each 3′ UTR was mutated, allowing translation to proceed into the 3′ UTR. TerByP refers to 'Termination ByPass', the region between the canonical termination codon and first in-frame termination codon in the 3′ UTR. Images were collected as in **a**. 'GFP (10×)' is a 2 s exposure. The dim yellowish fluorescence in 'GFP (10×)' for *unc-54(TerByP)* and *tbb-2(TerByP)* is autofluorescence. **c**, For each gene, the 3′ UTR was fused to mCherry and GFP. GFP expression was tested with the stop codon mutated to a sense codon (*TerByP*). For each of *eef-1A.1*, *rps-17*, *daf-6*, and *hlh-1*, GFP expression was also tested with the normal stop codon in place (3′ UTR). The ratio of GFP to mCherry fluorescence under a broad fluorescence filter was used as a metric (Extended Data Fig. 2, Methods). Each triangle represents an independently generated transgenic line; mean and s.d. of *n* (shown in parentheses) lines shown. Student's *t*-test two-tailed *P* value.

also argues against mRNA destabilization as a substantial factor in the protein loss observed upon readthrough.

The above results could be explained if GFP was generally incompatible with C-terminal fusions in our system. To address this, we inserted a variety of sequences downstream of GFP: 3 × Flag, 3 × haemagglutinin (HA), three random sequences created *in silico*, and six arbitrary fragments of in-frame coding sequence from *C. elegans* genes, approximately length-matched to 3′-UTR-encoded sequences (Fig. 2b). GFP expression varied between constructs but was generally higher than 3′-UTR-encoded sequences: 3 × HA, 3 × Flag, 2 out of 3 random sequences, and 4 out of 6 coding-derived fragments exhibited GFP:mCherry fluorescence ratios of >0.13, higher than all nine tested 3′-UTR-derived C-terminal extensions and significant statistically ($P = 0.004$, Kolmogorov–Smirnov test). Thus the effects of

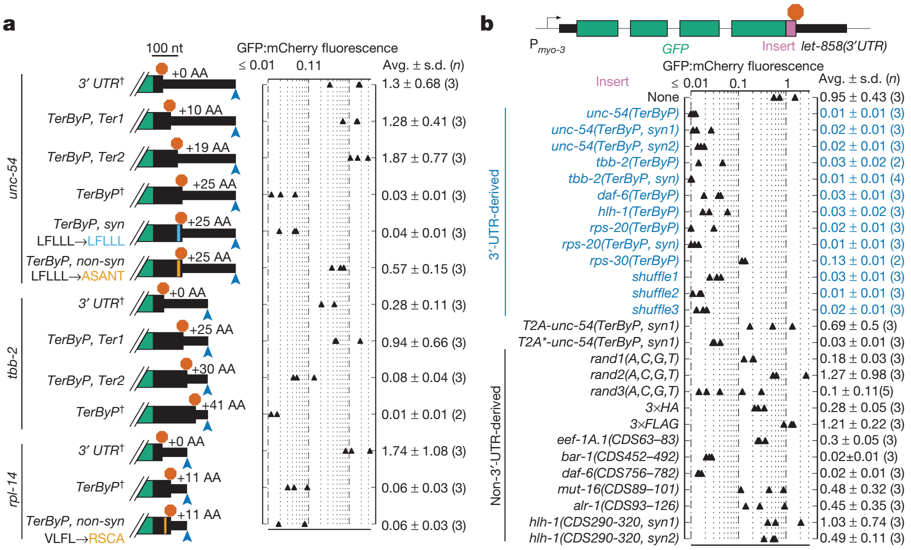


Figure 2 | Identification of determinants for product loss upon translation into the 3′ UTR. a, Shortening or non-synonymous mutations of the readthrough region can restore GFP expression. Stop codons and/or mutations were inserted into each GFP::3′-UTR fusion as shown, with stop codons (red stop sign) and poly(A) site (blue arrowhead). †indicates the constructs shown in Fig. 1. mCherry (pCFJ104) was used as a coinjection marker. '+*X* AA' indicates amino acids added relative to cognate control ('+0 AA') construct. Constructs and mutated regions drawn to scale, scale bar at top. Mean and s.d. of *n* (shown in parentheses) lines shown. **b**, 3′-UTR-encoded peptides are sufficient to confer GFP loss. Sequences were inserted upstream of the *let-858* 3′ UTR. 'syn', synonymously substituted variants. *Shuffle1–3* contain shuffled codons of *unc-54*, *tbb-2*, and *rpl-14*(VLFL to RSCA) TerByP regions (Extended Data Fig. 4). T2A, 'self-cleaving' peptide which releases the upstream nascent chain; T2A*, a non-cleaving variant[14,15]. *Rand1–3*(A, C, G, T) are random combinations of A, C, G, and T created *in silico*. CDS N–M is an arbitrary fragment of the respective coding DNA sequence of the gene (from amino acid N to M).

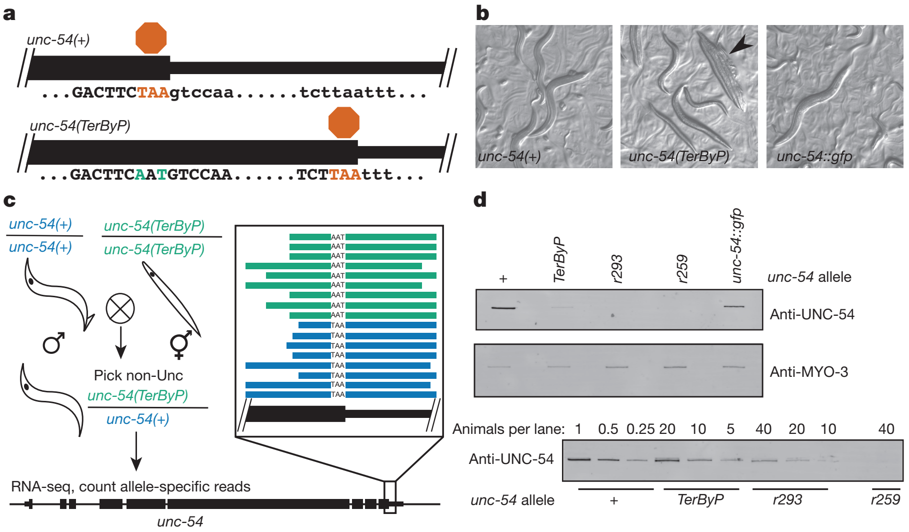


Figure 3 | Translation into the 3′ UTR at an endogenous locus decreases protein levels. a, Schematic of wild-type and readthrough alleles of *unc-54*, the latter made using CRISPR–Cas9 genome editing[17]. See Extended Data Table 1 for additional loci and edits. **b**, Brightfield images of *unc-54* alleles. Arrowhead indicates a 'bag of worms', the shell of an egg-laying-defective mother consumed by its retained progeny. **c**, RNA-seq from *unc-54(TerByP/+)* heterozygotes showed no differential effect on RNA levels. *unc-54(TerByP/+)* heterozygotes were chosen among progeny of *unc-54(+)* males crossed with *unc-54(TerByP)* homozygotes and allele-specific reads identified. The framed inset shows individual allele-specific RNA-seq reads (bars) from *unc-54(TerByP)* (AAT, green) and *unc-54(+)* (TAA, blue). See also Extended Data Figs 5, 6. **d**, Quantification of UNC-54 protein levels. Immunoblotting was performed on homozygous populations of the indicated animals. *unc-54(r293)* encodes a nonsense-mediated decay allele of *unc-54*, producing <5% of normal UNC-54 protein. *unc-54(r259)* contains a >17 kb deletion spanning most of the *unc-54* locus. For the lower blot, the number of animals loaded per lane is indicated. For gel source data, see Supplementary Fig. 1.

3′-UTR-encoded sequences are not explained by a general intolerance of GFP to C-terminal extensions (see also Methods, Extended Data Figs 3, 4).

It was conceivable that peculiarities of GFP and/or transgene expression systems might underlie the above observations. To establish effects of 3′ UTR translation at endogenous genes, we sought loci where (1) a loss of protein would be detectable phenotypically, (2) C-terminal fusions are known to be functional, (3) the next in-frame stop codon of the endogenous locus is ≥10 amino acids past the annotated stop codon, yet upstream of annotated poly(A) sites[16], and (4) there is little or no autoregulation/feedback. *unc-54* and *unc-22* meet all of the criteria, and *pha-4*, *unc-45*, and *tra-2* at least the first three points (Methods). For each locus, we mutated the stop codon to allow translation into the 3′ UTR[17] (Fig. 3a). In parallel, we analysed small insertions/deletions generating late frameshifts for *unc-22* and *unc-54*. Additional controls had length-matched sequences and/or GFP tags at the C terminus (Extended Data Table 1). For each of *unc-22*, *unc-45*, *unc-54*, and *tra-2*, translation into the 3′ UTR in at least one frame generated a strong hypomorphic (near null) phenotype specific to each locus. Other C-terminal tags for each gene did not cause loss of expression, although one *tra-2* C-terminal tag did produce a Tra phenotype. The ability to place alternative tags on the C terminus without obvious phenotypic consequences argues against a general sensitivity of the C terminus to tagging. For *unc-22* and *unc-54*, that elongation into the 3′ UTR in only some frames elicited a hypomorphic phenotype suggests that ribosome elongation into the 3′ UTR is not detrimental *per se*.

To determine the consequences on gene expression upon translation into 3′ UTRs, we analysed the effects of the *unc-54(cc3389)* TAA(stop) to AAT(Asn) mutation on RNA, translation, and protein output. We analysed mRNA expression in *unc-54(cc3389/+)* heterozygotes (phenotypically wild type to avoid complications from an Unc phenotype). RNA-seq revealed that the *unc-54(cc3389)* and wild-type alleles were at approximately equal amounts in the mRNA pool, suggesting that 3′ UTR translation does not appreciably destabilize the *unc-54* mRNA (Fig. 3c). In parallel, we detected an ~20-fold reduction in UNC-54 protein in immunoblots in *unc-54(cc3389)* mutants (Fig. 3d). To look for possible alterations in translation for *unc-54(cc3389)*, we examined the distribution of RNase-protected mRNA fragments with ribosome footprint profiling[18]. We observed no significant difference in the loading of ribosomes on *unc-54* mRNA (Extended Data Fig. 5), nor on the number, distribution, frame, or fragment size of ribosomes in the extended region (Extended Data Fig. 6).

A model that arises from these observations is that 3′-UTR-encoded peptides mark their resulting products for destruction, either co- or post-translationally. Conceivably this process might operate either in a specific cell/tissue type or in a broad spectrum of different contexts. A broadly-expressed reporter bearing a readthrough extension would be expected to highlight any tissue that failed to destabilize the C-terminal peptide. Using a broadly expressed promoter (*unc-37*) driving GFP with and without the *unc-54* 3′-UTR-encoded peptide, we observed no cells where GFP was robustly retained (data not shown).

We likewise considered the possibility that 3′-UTR-encoded peptides might act to limit protein levels in human cells, developing a specific assay using a lentiviral dual fluorescence reporter encoding puromycin *N*-acetyl-transferase tethered to mCherry-T2A, followed by eGFP and a multiple cloning site (Fig. 4a). The resulting reporter expresses both fluorophores from the same mRNA, yet as two disjoint polypeptides, allowing consideration of the effect of a peptide tag on eGFP expression independent of effects on mCherry and mRNA expression. We validated the split dual fluorophore approach in K562 cells using tags known to be destabilizing (d1ODC, d4ODC[19]) or not (3 × Flag, 3 × HA) (Fig. 4b). We selected 13 genes of varying expression and function, and inserted the region between the annotated termination codon and first-in-frame termination codon downstream of eGFP. For 9 of 13 genes, the readthrough region reduced the eGFP:mCherry fluorescence ratio between 3- and 30-fold, a stronger reduction than the degron d4ODC (Fig. 4c). Although not universal, the substantial loss of eGFP fluorescence for a majority of readthrough regions opens up the possibility that translation into 3′ UTRs may be generally inhibitory to expression across systems.

We hypothesize that a function of 3′ UTRs is to minimize the accumulation of extended protein products that could be produced through translational readthrough.

This feature may prove generally important in causes of genetic disease. For example, readthrough alleles (such as stop to Gln) of the *HBA2* locus in humans produce a fraction (~1%[9]) of normal HBA2

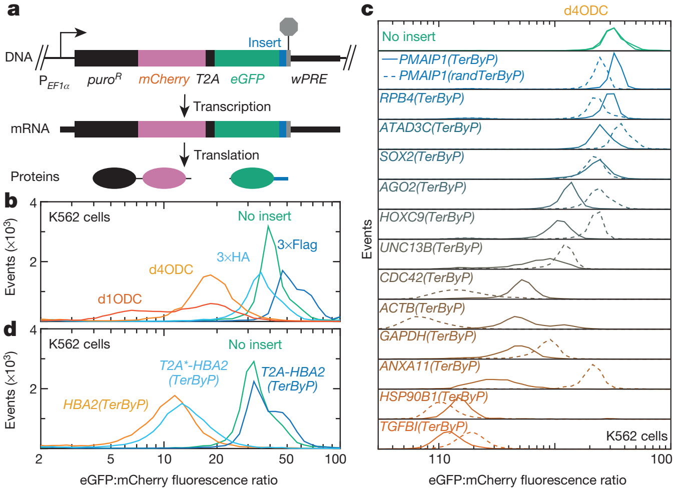


Figure 4 | Translation into 3′ UTRs results in protein loss for several genes in humans. a, Lentiviral reporter schematic. A puroR–mCherry fusion was co-translationally cleaved from eGFP-insert by T2A. Constructs in **b**–**d** were integrated into K562 cells via lentiviral infection and puromycin selection. **b**, Validation of dual fluorescence reporter. Inserts downstream of eGFP were 3× Flag, 3× HA, and degrons d4ODC (half-life, ~4 h[19]), d1ODC (half-life, ~1 h[19]). **c**, The sequence between the annotated and first in-frame termination codon (TerByP) from each gene was inserted downstream of eGFP (solid line). For comparison, nucleotides of each TerByP region were randomized, producing a length- and nucleotide-frequency-matched construct (randTerByP, dashed line). Cells with eGFP lacking an insert and grown a week apart (top, green solid lines) and approximate fluorescence ratio of d4ODC (orange line) are shown. **d**, The first 30 amino acids of the *HBA2* 3′ UTR were inserted downstream of eGFP (orange). Insertion of a self-cleaving T2A peptide restored expression (blue), whereas an uncleavable mutant (T2A*) did not (light blue).

protein (α-globin), causing thalassemia. Translation into the *HBA2* 3′ UTR is known to destabilize the *HBA2* mRNA[20], but it is unclear what effect the appended C-terminal 31 amino acids have on HBA2 protein. We considered the possibility that the *HBA2* 3′-UTR-encoded peptide might prevent protein expression in humans, contributing to the loss of HBA2 protein. When appended to eGFP, the *HBA2* 3′-UTR-encoded peptide decreased the eGFP:mCherry fluorescence ratio in K562 cells (Fig. 4d). Furthermore, eGFP fluorescence was rescued by a self-cleaving (but not an uncleaveable mutant) T2A peptide.

Several observations from the literature support the notion that 3′-UTR-encoded peptides may be detrimental to expression for more genes and organisms than those assessed here. In *Saccharomyces cerevisiae*, translation past a point in the *HIS3* 3′ UTR confers a substantial loss in protein expression, without detectable effects on mRNA levels[11]. Similarly, readthrough of the cyclic AMP phosphodiesterase *PDE2* stop codon produces a destabilized protein variant, and this has been suggested to explain elevated cyclic AMP levels in *PSI*$^+$ yeast[10]. Differential stability by polymorphisms in the readthrough peptide of *SKY1* has been postulated to explain [PSI]-induced strain differences in diamide sensitivity[21]. Particularly intriguing are recent findings that stop codon mutations at the c-FLIP$_L$ locus confer protein instability for this anti-apoptotic factor in mice, leading to embryonic lethality[12]. The same study also noted several hereditary human disease alleles where 3′-UTR-encoded peptides are destabilizing, conferring marked decreases in protein activity and level[8].

Not every case of stop codon readthrough is destabilizing[4–7] (Fig. 4c), and some readthrough events are functional and regulated to defined levels[22–25]. Understanding the mechanisms by which some readthrough events are detected and cleared (whereas others are not) may prove informative in biological contexts and pathological states where inappropriate readthrough occurs. We do not yet know the determinants of a translated 3′ UTR sequence that confer loss of protein, though the ability of numerous sequences (including shuffled and randomized 3′ UTR variants, Fig. 2b, 4c) suggests that a highly degenerate sequence is sufficient. Consistent with the idea that the effects of readthrough peptides may be mediated via their biophysical characteristics, we observed a significant negative relationship between hydrophobicity and expression of GFP (K562 cells, *C. elegans*) and endogenous loci (*unc-22* and *unc-54*, *C. elegans*) (Extended Data Figs 7–9, Supplementary Information).

Destabilization by 3′-UTR-encoded peptides could effectively mitigate at least three types of events in which a stop codon is inappropriately bypassed: (1) stop codon misreading (for example, by suppressor tRNAs). Suppressor tRNAs permit readthrough of up to 30% of ribosomes at a stop codon (UAA, UAG, or UGA)[13]. Whereas some suppressor tRNAs can be toxic, other cells tolerate even high levels of readthrough[13,26–28]. Destabilization of readthrough products by C-terminal appendages may effectively buffer cells from suppressor tRNA-induced proteostatic chaos. (2) A ribosomal frameshift in a coding region which is late enough that no premature termination codon is encountered. In this case, ribosomes would enter the 3′ UTR out-of-frame with the coding region. In our manipulations, translation of 3′ UTRs in multiple frames was detrimental to expression (Extended Data Table 1, data not shown), and similar amino acid and hydropathy biases hold for all three 3′ UTR frames (Extended Data Figs 8, 9). (3) Errors in RNA processing or ribosome dysfunction could produce a variety of other improperly terminated peptides from which translation readthrough mitigation would provide valuable relief.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 28 October 2015; accepted 5 May 2016.
Published online 1 June 2016.

1. Klauer, A. A. & van Hoof, A. Degradation of mRNAs that lack a stop codon: a decade of nonstop progress. *Wiley Interdiscip. Rev. RNA* **3,** 649–660 (2012).
2. Hamby, S. E., Thomas, N. S., Cooper, D. N. & Chuzhanova, N. A meta-analysis of single base-pair substitutions in translational termination codons ('nonstop' mutations) that cause human inherited disease. *Hum. Genomics* **5,** 241–264 (2011).
3. Williams, I., Richardson, J., Starkey, A. & Stansfield, I. Genome-wide prediction of stop codon readthrough during translation in the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **32,** 6605–6616 (2004).
4. Falini, B. *et al.* Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *N. Engl. J. Med.* **352,** 254–266 (2005).
5. Hollingsworth, T. J. & Gross, A. K. The severe autosomal dominant retinitis pigmentosa rhodopsin mutant Ter349Glu mislocalizes and induces rapid rod cell death. *J. Biol. Chem.* **288,** 29047–29055 (2013).
6. Vidal, R. *et al.* A stop-codon mutation in the BRI gene associated with familial British dementia. *Nature* **399,** 776–781 (1999).
7. Vidal, R. *et al.* A decamer duplication in the 3′ region of the BRI gene originates an amyloid peptide that is associated with dementia in a Danish kindred. *Proc. Natl Acad. Sci. USA* **97,** 4920–4925 (2000).
8. Pang, S. *et al.* A novel nonstop mutation in the stop codon and a novel missense mutation in the type II 3beta-hydroxysteroid dehydrogenase (3beta-HSD) gene causing, respectively, nonclassic and classic 3β-HSD deficiency congenital adrenal hyperplasia. *J. Clin. Endocrinol. Metab.* **87,** 2556–2563 (2002).
9. Clegg, J. B., Weatherall, D. J. & Milner, P. F. Haemoglobin Constant Spring—a chain termination mutant? *Nature* **234,** 337–340 (1971).
10. Namy, O., Duchateau-Nguyen, G. & Rousset, J. P. Translational readthrough of the *PDE2* stop codon modulates cAMP levels in *Saccharomyces cerevisiae*. *Mol. Microbiol.* **43,** 641–652 (2002).
11. Inada, T. & Aiba, H. Translation of aberrant mRNAs lacking a termination codon or with a shortened 3′-UTR is repressed after initiation in yeast. *EMBO J.* **24,** 1584–1595 (2005).
12. Shibata, N. *et al.* Degradation of stop codon read-through mutant proteins via the ubiquitin-proteasome system causes hereditary disorders. *J. Biol. Chem.* **290,** 28428–28437 (2015).
13. Capone, J. P., Sharp, P. A. & RajBhandary, U. L. Amber, ochre and opal suppressor tRNA genes derived from a human serine tRNA gene. *EMBO J.* **4,** 213–221 (1985).
14. Ahier, A. & Jarriault, S. Simultaneous expression of multiple proteins under a single promoter in *Caenorhabditis elegans* via a versatile 2A-based toolkit. *Genetics* **196,** 605–613 (2014).
15. Doronina, V. A. *et al.* Site-specific release of nascent chains from ribosomes at a sense codon. *Mol. Cell. Biol.* **28,** 4227–4239 (2008).

16. Jan, C. H., Friedman, R. C., Ruby, J. G. & Bartel, D. P. Formation, regulation and evolution of *Caenorhabditis elegans* 3’UTRs. *Nature* **469,** 97–101 (2011).
17. Arribere, J. A. *et al.* Efficient marker-free recovery of custom genetic modifications with CRISPR/Cas9 in *Caenorhabditis elegans*. *Genetics* **198,** 837–846 (2014).
18. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–223 (2009).
19. Yen, H.-C. S., Xu, Q., Chou, D. M., Zhao, Z. & Elledge, S. J. Global protein stability profiling in mammalian cells. *Science* **322,** 918–923 (2008).
20. Liebhaber, S. A. & Kan, Y. W. Differentiation of the mRNA transcripts originating from the alpha 1- and alpha 2-globin loci in normals and alpha-thalassemics. *J. Clin. Invest.* **68,** 439–446 (1981).
21. Torabi, N. & Kruglyak, L. Genetic basis of hidden phenotypic variation revealed by increased translational readthrough in yeast. *PLoS Genet.* **8,** e1002546 (2012).
22. Steneberg, P. & Samakovlis, C. A novel stop codon readthrough mechanism produces functional Headcase protein in *Drosophila* trachea. *EMBO Rep.* **2,** 593–597 (2001).
23. Freitag, J., Ast, J. & Bölker, M. Cryptic peroxisomal targeting via alternative splicing and stop codon read-through in fungi. *Nature* **485,** 522–525 (2012).
24. Eswarappa, S. M. *et al.* Programmed translational readthrough generates antiangiogenic VEGF-Ax. *Cell* **157,** 1605–1618 (2014).
25. True, H. L. & Lindquist, S. L. A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature* **407,** 477–483 (2000).
26. Waterston, R. H. A second informational suppressor, *sup-7 X*, in *Caenorhabditis elegans. Genetics* **97,** 307–325 (1981).
27. Laski, F. A., Ganguly, S., Sharp, P. A., RajBhandary, U. L. & Rubin, G. M. Construction, stable transformation, and function of an amber suppressor tRNA gene in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **86,** 6696–6698 (1989).
28. Hudziak, R. M., Laski, F. A., RajBhandary, U. L., Sharp, P. A. & Capecchi, M. R. Establishment of mammalian cell lines containing multiple nonsense mutations and functional suppressor tRNA genes. *Cell* **31,** 137–146 (1982).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the Fire Laboratory for critical reading of the manuscript, C. Frøkjær-Jensen and K. Artiles for technical expertise, and T. Schedl, T. Inada, C. Joazeiro, L. Ling, A. Nager, and N. Spies for discussions. A. Sapiro and B. Li were instrumental in developing the RNA-seq2 protocol. This work was supported by grants from NIH R01GM37706, T32HG000044 (G.T.H.), 1DP2HD084069-01 (M.C.B.), 5F32GM112474-02 (J.A.A.), Walter and Idun Berry Foundation (E.S.C.), and NSF DGE-114747 (C.H.L.).

Author Contributions J.A.A., E.S.C., and A.Z.F. designed *C. elegans* experiments. J.A.A. and E.S.C. conducted *C. elegans* experiments. N.J. developed the RNA-seq2 protocol. J.A.A. performed computational analyses. J.A.A. conducted experiments in human cell lines, as designed and aided by J.A.A., G.T.H., C.H.L., M.C.B., and A.Z.F. J.A.A. and A.Z.F. wrote the paper with help from all authors.

Author Information Sequencing data are available at Sequence Read Archive (SRP064516). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.Z.F. (afire@stanford.edu).

Reviewer Information *Nature* thanks J. S. Butler, M. Yarus and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

C. elegans strain construction and husbandry. *C. elegans* were grown at 23 °C on agar plates with nematode growth medium seeded with *Escherichia coli* strain OP50 as described[29]. Some strains were provided by the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). A full list of strains used is available in Supplementary Table 1.

Transgenic array-containing strains were generated as follows: PD5102 (*pha-1(e2123ts)I; rde-1(ne300)V*) young adult hermaphrodites (grown at 16 °C) were injected with a mix of 90 ng ul^{-1} pC1 (containing a rescuing fragment of *pha-1*), 5 ng ul^{-1} of an mCherry-containing vector, and 5 ng ul^{-1} of a GFP-containing vector. Unless otherwise indicated, GFP was driven by the *myo-3* promoter to drive expression in the body-wall muscle[30]. Injectants were shifted to 23 °C to select for F1 progeny animals bearing a transgenic array (selecting for *pha-1*(+) expression[31]). The *rde-1* allele included in this strain avoided a modest degree of secondary siRNA-based silencing observed with many extrachromosomal transgenes[32]. For transgenic lines generating low levels of GFP, we considered the possibility that the GFP protein was toxic and selected against. Under this model, one might expect (1) a subset of sick and/or dead GFP-positive F1 animals, (2) muscle defects due to muscle-specific expression of potentially-toxic GFP derivatives, (3) concomitant low levels of mCherry, and/or (4) a decrease in the efficiency with which transgenic lines were obtained[32]. None of these effects were observed, arguing against any contribution of negative selection to the observed low GFP expression.

For a subset of strains, we deviated from the above protocol to generate *pha-1* arrays as follows: (1) whereas most transgenic lines were generated from independently injected parents, a handful of strains were possibly generated from siblings of an injected parent (PD6480, 6481, 6482, 6483, 6484, 6485, 6486, 6493, 6494, 6495). In these cases, all injectants were pooled together on the same plate, and independent F1 were picked off to generate transgenic lines. Previous work has demonstrated independent F1 from the same injected parent carry distinct transgenic arrays[32,33]. (2) During the course of our analyses, we found some strains with an mCherry-negative subpopulation. The mCherry-positive subpopulation was isolated and propagated to generate the strains PD6401, 6450, 6452, 6456, 6457, and 6464.

CRISPR–Cas9 genome editing was performed in the VC2010 (PD1074) N2 background as described[17]. We selected *pha-4* (refs 34, 35), *unc-45* (refs 36, 37), *tra-2* (refs 38–40), *unc-22* (ref. 29), and *unc-54* (refs 26, 38, 41, 42) based on the criteria in the text (citations indicated). The statement that *unc-22* and *unc-54* exhibit little or no autoregulation/feedback is based on a number of genetic experiments (with heterozygous[29], amber-suppressed[26], and/or *smg*-suppressed[38] alleles) which express either UNC-54 or UNC-22 at stable intermediate levels (between wild type and null). Alleles of *unc-45* were initially generated in the VC2010 background, though the embryonic lethality made *unc-45(TerByP)* difficult to maintain. We subsequently remade all alleles in a balanced heterozygote background (*sC1(s2023)(dpy-1(s2170))* III/+) and considered non-Dpy segregants for phenotypic analyses.

Human cell line construction. K562 cells (obtained from ATCC) were grown at a density of ∼0.5 to 1 × 10^6 cells ml^{-1} in RPMI medium supplemented with penicillin/streptomycin, L-glutamine, and 10% FBS. All cell lines were maintained in a humidified incubator (37 °C, 5% CO$_2$), and checked regularly for mycoplasma contamination. As a means of validating K562 cells, we performed RNA-seq on a subset of lines and observed good correlation with published data sets[43] (data not shown). Viral particles were produced in HEK293T cells in 6-well dishes, and 1 ml of viral supernatant was used to infect ∼100,000 K562 cells by spin infection, 10^3 relative centrifugal force for 2 h. Polybrene was omitted to keep the infection rate low (<10%), ensuring a single incorporation event for most cells. After 3 days of recovery, cells were selected with puromycin at 0.7 μg ml^{-1} for at least 3 days. Fluorescence was examined on a BD Accuri C6 flow cytometer, with appropriate gating for live cell events and investigators blinded to cell line identity. For each construct examined via puromycin selection in K562 cells, similar eGFP and mCherry fluorescence levels were also observed in transient transfection in HEK293T cells in the absence of puromycin, arguing against a puromycin-selected skew in mCherry fluorescence.

Plasmids. Plasmids were constructed by restriction digest or Gibson cloning as detailed in Supplementary Table 2. pJA138/L3785 and pJA137/pCFJ104 were used as the basis of all *C. elegans* GFP- or mCherry-containing vectors, respectively. Portions of pMCB306 and pMCB309 were used to construct pJA291, the parental *puro::mCherry::T2A::eGFP::MCS::wPRE* vector for experiments in human cells. Plasmids were confirmed by both sequencing and restriction digest, and plasmid concentrations determined with the QuBit dsDNA Broad Range kit (Invitrogen). Plasmids that may be useful have been deposited with Addgene: pJA327 (*C. elegans* superfolder GFP in L3785), pJA291, pJA317 (pJA291 with d1ODC insert) and pJA318 (pJA291 with d4ODC insert).

GFP fusions were contructed with a GFP variant that corresponds to wild-type (*Aequora*) GFP with mutations at position 65 (Ser to Thr for human, Ser to Cys for *C. elegans*) known to improve folding and acquisition of fluorescence. Even with these mutations, GFP has a known propensity to misfold under some circumstances, and we therefore examined the effect of a subset of the 3′-UTR-encoded sequences (*hlh-1*, *daf-6*, and *unc-54*) downstream of a faster and more robust-folding GFP variant, superfolder GFP (ref. 44). The observed reduction in superfolder-GFP:mCherry ratios was quantitatively similar to that observed with normal GFP (Extended Data Fig. 3).

Sequences of Flag[45], HA[46], d1ODC[19], and d4ODC[19] were obtained from the indicated publications. For exact sequences, see Supplementary Tables 1 and 2. T2A was previously shown to function in *C. elegans*[14]. Translation elongation through a member of the 2A peptide family (consensus D(V/I)EXNPGP) causes ribosomal pause, then release of the N-terminal peptide up to and including the Gly[15]. Translation elongation resumes, with the C-terminal peptide being produced with an N-terminal Pro.

Microscopy. Animals were immobilized by placing on a slide with a coverslip in 5 mM EDTA, 50 mM NaCl, 1 mM levamisole and imaged on a Nikon Eclipse E6000 microscope using a Nikon super high pressure mercury lamp power supply. Filter cubes for fluorescence images were GFP (96342, Nikon Corp), mCherry (96321, Nikon Corp), and broad (GFP and mCherry, 59022, Chroma Technology Corp). Images were collected with a 3CCD Digital Camera C7780 (Hamamatsu Corp) using HCImage (Version 1.0.2.060107, Hamamatsu Corp). Images of PD4251 and one of PD3363/3364 were taken for each imaging session and compared to ensure consistency between days.

For quantification of GFP to mCherry relative fluorescence, animals were imaged using a 4 × objective with a broad filter and 200 ms exposure. Investigators were blinded during imaging. To avoid image over- or underexposure, a number of exceptionally bright or dim strains were taken with a decreased or increased (respectively) exposure time (PD1798, 40 ms; PD3294, 500 ms; PD3299, 50 ms; PD3395, 500 ms; PD6327, 50 ms; PD6375, 50 ms; PD1786, 50 ms; PD1789, 50 ms; PD1790, 50 ms; PD6460, 500 ms; PD6469, 50 ms; PD6471, 50 ms; PD6472, 50 ms; PD6473, 50 ms; PD6485, 100 ms; PD1787, 40 ms; PD6450, 100 ms; PD6455, 40 ms; PD6477, 40 ms; PD6479, 50 ms; PD6498, 40 ms; PD6499, 40 ms; PD6500, 40 ms; PD6501, 40 ms; PD6502, 40 ms; PD6503, 40 ms; PD6504, 40 ms).

Raw pixel values for the red and green channels were obtained from image files using the tifffile package in python. Pixels below a threshold distance (200) from the median pixel intensity of the entire image were discarded as background. Pixels above a threshold intensity distance (4,000 of a possible 4,095) from the origin were discarded as saturated. The median pixel intensity for the entire image (essentially the black background, given the relatively low density of *C. elegans* tissue) was subtracted from the remaining pixels, and the slope of the linear regression line taken as the GFP:mCherry fluorescence ratio. This metric was robust to different exposure times and neutral density filters.

Statistics. Statistical tests and *P* values are stated throughout the text and figures.

To test statistical significance of C-terminal appendage effects on the GFP:mCherry fluorescence ratio (Fig. 2b), we divided the data into two groups: (1) 3′-UTR-derived (*unc-54(TerByP)*, *tbb-2(TerByP)*, *daf-6(TerByP)*, *hlh-1(TerByP)*, *rps-20(TerByP)*, *rps-30(TerByP)*, *shuffle1-3*); and (2) non-3′-UTR-derived (*rand1-3(A,C,G,T)*, *3 × HA*, *3 × Flag*, *eef-1A.1(CDS63-83)*, *bar-1(CDS452-492)*, *daf-6(CDS756-782)*, *mut-16(CDS89-101)*, *alr-1(CDS93-126)*, *hlh-1(CDS290-320,syn1)*). For each construct, we took the average GFP:mCherry fluorescence ratio of all available lines. We compared the distribution of 3′-UTR-derived and non-3′-UTR-derived GFP:mCherry fluorescence ratio values by Kolmogorov–Smirnov test ($P = 0.004$).

No statistical methods were used to predetermine sample size.

Ribosome footprint profiling. Ribo-seq was performed essentially as previously described[18,47], with a few modifications. Briefly, animals were grown to around L4 stage, and collected by centrifugation and flash-freezing in liquid nitrogen. Animals were ground with a mortar and pestle in liquid nitrogen, after which the powder was thawed in excess volume ice-cold polysome lysis buffer (20 mM Tris (pH 8.0), 140 mM KCl, 1.5 mM MgCl$_2$, 1% Triton) with cycloheximide (100 μg ml^{-1}). RNase 1 and sucrose gradient centrifugation was performed as previously described[47]. Around 2 μg of purified, RNase-1-digested monosomal RNA was run on a urea 15% polyacrylamide gel, and the entire region from ∼15–30 nt was excised for library preparation. At this point, the protocol continued with T4 polynucleotide kinase (PNK) (New England Biolabs) treatment as with the RNA-seq1 protocol (next section).

RNA sequencing. Two RNA sequencing (RNA-seq) procotols were used in this study. The first RNA-seq protocol (RNA-seq1) was performed on homozygote populations of animals (Extended Data Fig. 5). 5 μg of total RNA was treated with the RiboZero kit (Illumina). RNA was fragmented at 95 °C for 30 min by addition of an equal volume of 100 mM sodium carbonate, 0.5 mM EDTA (pH 9.3) buffer. RNA fragments were gel-purified, then treated with T4 PNK. 3′-ligation with AF-JA-34.2 adaptor (/5rApp/NNNNNNAGATCGGAAGAGCACACGTCT/3ddC/, Integrated DNA Technologies) and T4 RNA ligase 1 (New England Biolabs) was performed at room temperature for 4 h with 20% PEG8000 in 3.3 mM DTT, 8.3 mM glycerol, 50 mM HEPES KOH (pH 8.3), 10 mM $MgCl_2$, 10 μg ml^{-1} acetylated BSA. Unligated AF-JA-34.2 was removed by sequential treatment with 5′deadenylase (M0331S, New England Biolabs), then RecJ$_f$ (M0264S, NEB). Reverse transcription was carried out with AF-JA-126 (/5Phos/AGATCGGAAGAGCGTCGTGT/iSp18/CACTCA/iSp18/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT, Integrated DNA Technologies) as a primer. Circular ligase treatment and PCR were as previously described[47].

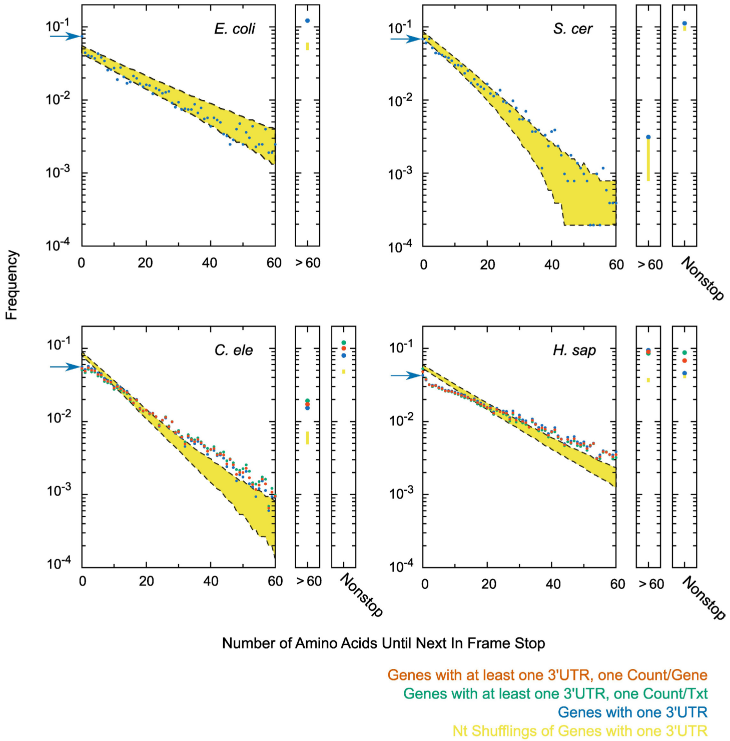
A second RNA-seq protocol (RNA-seq2) was used to examine RNA levels with small numbers of heterozygote animals (Fig. 3C). Around sixty L2–L4 mixed gender animals were picked and flash-frozen in 50 mM NaCl, and RNA extracted with trizol. RNaseH and 94 oligonucleotides complementary to ribosomal RNA were used to deplete rRNA from the sample[48]. Briefly, ~250 ng of a cocktail of DNA oligonucleotides complementary to rRNA (Supplementary Table 3, ordered from Integrated DNA Technologies) was mixed with ~100 ng total RNA in 125 mM Tris (pH 7.4), 250 mM NaCl in 8 μl. The sample was heat-denatured at 95 °C for 2 min, then cooled at −0.1 °C per s to 45 °C. 1 μl of digestion buffer was added (500 mM Tris (pH 7.4), 1 M NaCl, 200 mM $MgCl_2$) with 1 μl (5 units) Thermostable RNase H (Epicentre), and the sample was incubated at 45 °C for one hour. DNA oligonucleotides were removed by treatment with TURBO DNase (ThermoFisher) at 37 °C, and RNA was extracted using an equal volume of phenol/chloroform. An RNA-seq library was prepared using the SMARTer Stranded RNA-Seq kit (Clontech Laboratories, Inc.).

Sequencing. Libraries were sequenced on a MiSeq Genome Analyzer (Illumina, Inc.). Reads were mapped to the *C. elegans* genome (Ensembl70, WBcel215) using STAR (v2.3.1 ref. 49), with the mutated bases of *unc-54(cc3389)* and *unc-54(e1301)* masked. For Ribo-seq and RNA-seq1, reads bearing the same last 6 nucleotides (from NNNNNN, added with AF-JA-34.2) were assumed to be PCR duplicates and collapsed to a single read. For RNA-seq2, multiple reads containing the same start and stop mapping positions were collapsed to a single read count to reduce effects of PCR bias. The removal of PCR duplicates with either protocol only affected ~5–10% of reads and did not adversely impact any of the analyses shown. RNA-seq1 and Ribo-seq were performed once for each strain shown in Extended Data Figs 5, 6.

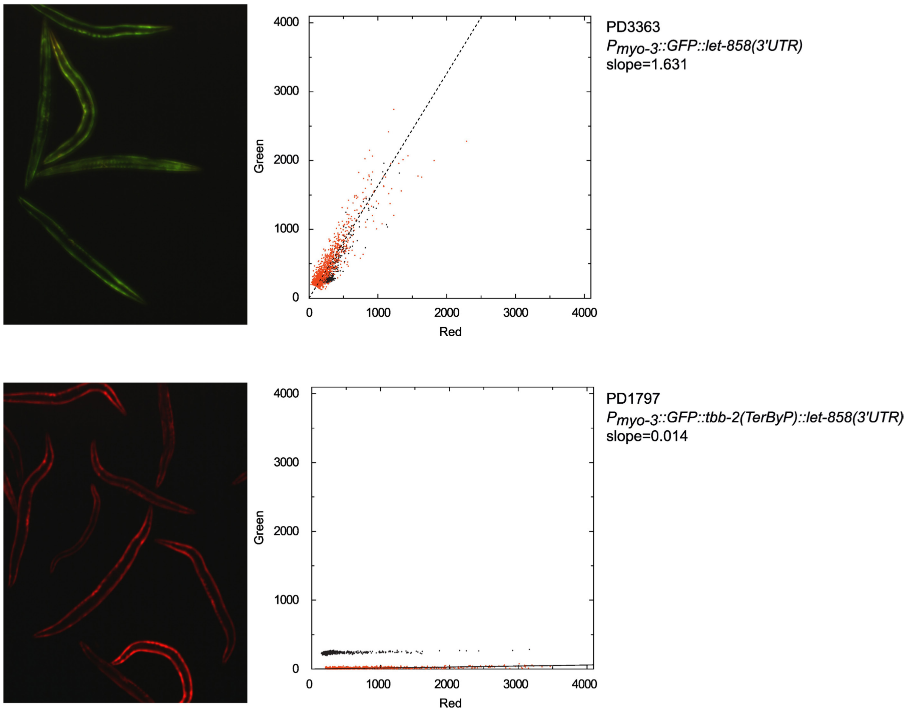
Genomes and annotations. Although we sought to use the latest genome versions and annotations, we found it prudent to take advantage of the care and time with which other researchers annotated and analysed earlier versions of genomes. For whole genome alignment of nematode species, *C. elegans* UCSC genome ce10/WS220 was used. To examine the length of predicted C-terminal extensions upon readthrough (Extended Data Fig. 1), genomes and annotations of each of the indicated species were as follows: *E. coli* Ensembl genome and annotations from assembly GCA_000967155.1.30, *S. cerevisiae* genome S288c (R57-1-1_20071212) and annotations[50], *C. elegans* UCSC genome (WS190/ce6) and annotations[16], *H. sapiens* Ensembl genome release 83 and annotations from TargetScan v7.0[51].

Immunobloting. Animals were boiled in 1 × SDS loading buffer (65 mM Tris (pH 6.8), 10% glycerol, 2% SDS, 2 mM PMSF, 1 × Halt Protease Inhibitor (Thermo), 10% 2-mercaptoethanol) and run on a 7.5% Criterion TGX gel (Bio-Rad Laboratories, Inc.). Protein was transferred to a low background fluorescence PVDF membrane (Millipore). The membrane was blocked in 3% nonfat milk in 1 × PBST with 250 mM NaCl. The 5-6 antibody was used at a 1:5000 dilution to detect *myo-3*, and 5-8 antibody used at a 1:5000 dilution to detect *unc-54* (ref. 52). The 5-6 and 5-8 monoclonal antibodies were produced previously by purification of endogenous myosin proteins. Secondary antibody staining was performed with 1:500 Cy3-conjugated affiniPure goat anti-mouse (Jackson Immunoresearch). Imaging was performed on a Typhoon Trio (Amersham Biosciences), and quantification performed in ImageJ. For the lower blot of Fig. 3d, lysates were made from multiple animals, and serial dilutions performed to titrate the number of animals per lane.

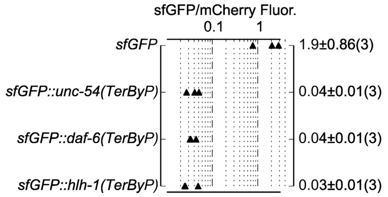
29. Brenner, S. The genetics of *Caenorhabditis elegans*. *Genetics* **77,** 71–94 (1974).
30. Okkema, P. G., Harrison, S. W., Plunger, V., Aryana, A. & Fire, A. Sequence requirements for myosin gene expression and regulation in *Caenorhabditis elegans*. *Genetics* **135,** 385–404 (1993).
31. Granato, M., Schnabel, H. & Schnabel, R. *pha-1*, a selectable marker for gene transfer in *C. elegans*. *Nucleic Acids Res.* **22,** 1762–1763 (1994).
32. Mello, C. C., Kramer, J. M., Stinchcomb, D. & Ambros, V. Efficient gene transfer in *C. elegans*: extrachromosomal maintenance and integration of transforming sequences. *EMBO J.* **10,** 3959–3970 (1991).
33. Stinchcomb, D. T., Shaw, J. E., Carr, S. H. & Hirsh, D. Extrachromosomal DNA transformation of *Caenorhabditis elegans*. *Mol. Cell. Biol.* **5,** 3484–3496 (1985).
34. Mango, S. E., Lambie, E. J. & Kimble, J. The *pha-4* gene is required to generate the pharyngeal primordium of *Caenorhabditis elegans*. *Development* **120,** 3019–3031 (1994).
35. Zhong, M. *et al.* Genome-wide identification of binding sites defines distinct functions for *Caenorhabditis elegans* PHA-4/FOXA in development and environmental response. *PLoS Genet.* **6,** e1000848 (2010).
36. Venolia, L. & Waterston, R. H. The *unc-45* gene of *Caenorhabditis elegans* is an essential muscle-affecting gene with maternal expression. *Genetics* **126,** 345–353 (1990).
37. Ao, W. & Pilgrim, D. *Caenorhabditis elegans* UNC-45 is a component of muscle thick filaments and colocalizes with myosin heavy chain B, but not myosin heavy chain A. *J. Cell Biol.* **148,** 375–384 (2000).
38. Hodgkin, J., Papp, A., Pulak, R., Ambros, V. & Anderson, P. A new kind of informational suppression in the nematode *Caenorhabditis elegans*. *Genetics* **123,** 301–313 (1989).
39. Hodgkin, J. A. & Brenner, S. Mutations causing transformation of sexual phenotype in the nematode *Caenorhabditis elegans*. *Genetics* **86,** 275–287 (1977).
40. Mapes, J., Chen, J.-T., Yu, J.-S. & Xue, D. Somatic sex determination in *Caenorhabditis elegans* is modulated by SUP-26 repression of *tra-2* translation. *Proc. Natl Acad. Sci. USA* **107,** 18022–18027 (2010).
41. Anderson, P. & Brenner, S. A selection for myosin heavy chain mutants in the nematode *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **81,** 4470–4474 (1984).
42. Eide, D. & Anderson, P. The gene structures of spontaneous mutations affecting a *Caenorhabditis elegans* myosin heavy chain gene. *Genetics* **109,** 67–79 (1985).
43. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science* **306,** 636–640 (2004).
44. Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T. C. & Waldo, G. S. Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.* **24,** 79–88 (2006).
45. Hopp, T. P. *et al.* A short polypeptide marker sequence useful for recombinant protein identification and purification. *Nat. Biotechnol.* **6,** 1204–1210 (1988).
46. Field, J. *et al.* Purification of a RAS-responsive adenylyl cyclase complex from *Saccharomyces cerevisiae* by use of an epitope addition method. *Mol. Cell. Biol.* **8,** 2159–2165 (1988).
47. Stadler, M. & Fire, A. Wobble base-pairing slows *in vivo* translation elongation in metazoans. *RNA* **17,** 2063–2073 (2011).
48. Morlan, J. D., Qu, K. & Sinicropi, D. V. Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. *PLoS One* **7,** e42882 (2012).
49. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
50. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320,** 1344–1349 (2008).
51. Agarwal, V., Bell, G. W., Nam, J. W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4,** 1–38 (2015).
52. Miller, D. M. III, Ortiz, I., Berliner, G. C. & Epstein, H. F. Differential localization of two myosins within nematode thick filaments. *Cell* **34,** 477–490 (1983).
53. Thompson, O. *et al.* The million mutation project: a new approach to genetics in *Caenorhabditis elegans*. *Genome Res.* **23,** 1749–1762 (2013).
54. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157,** 105–132 (1982).



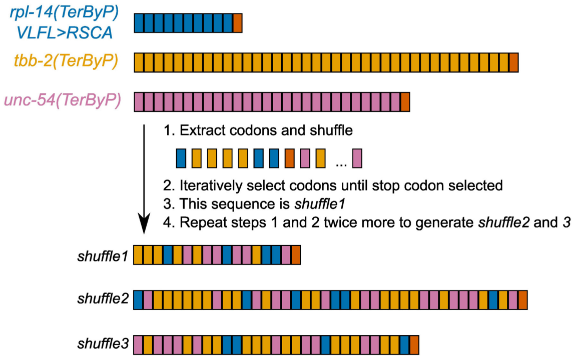
Extended Data Figure 1 | Distribution of C-terminal extensions upon stop codon readthrough. Annotations and genomes were as described in Supplementary Methods. Each 3′ UTR was translated starting one codon after the stop codon until the next in-frame stop codon. For metazoans, counting was performed in three different ways: including only genes for which exactly one 3′ UTR was annotated (blue), counting each annotated 3′ UTR separately (green), or counting each gene once and splitting gene counts with multiple 3′ UTRs equally amongst the 3′ UTR isoforms (red). ‘Nonstop’ denotes 3′ UTRs for which no stop codon was encountered before the poly(A) tail. For each species the distribution of next in-frame stop codons was calculated for 1,000 nucleotide shuffling of 3′ UTR sequences for genes with a single 3′ UTR annotated, and 95% confidence interval shown (yellow). A similar ‘randomized’ distribution was obtained upon shuffling 3′ UTR sequences and preserving dinucleotide frequency. The frequency of stops immediately after the annotated stop codon (amino acid length 0) is highlighted with a blue arrow in each species. The distribution of peptide lengths follows an exponential decay curve, where the slope is related to the probability of encountering a stop codon at each position. In the simplest model, the probability of encountering a stop codon is constant throughout the 3′ UTR, accounting for the roughly linear shape of each plot (previously noted[2,3]). Notable exceptions are a tendency towards second in-frame stops in *E. coli* (blue arrow), and a tendency towards peptides >60 amino acids in length in all species. In *E. coli*, the enrichment towards longer downstream peptides is at least partially explained by the operonic layout of genes.



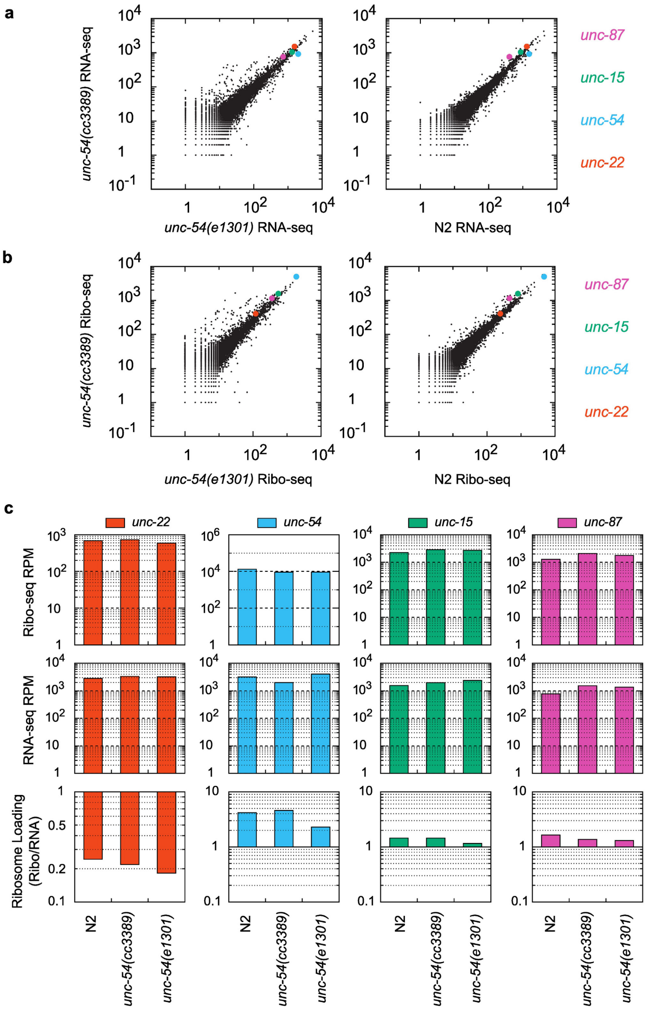
Extended Data Figure 2 | Example quantification of the GFP:mCherry fluorescence ratios of images. Images were taken under a broad excitation and emission filter to allow for simultaneous capture of GFP and mCherry fluorescence. Intensities of each pixel in the red and green channels were extracted in python. Unfiltered pixel intensities are shown as black dots. Pixels were filtered, background subtracted, and linear regression performed (red dots and line, see Methods). For simplicity, the green–red intensities from 1,000 random pixels are shown. The GFP:mCherry fluorescence ratio was taken as the slope of the linear regression line. 10 × objective.



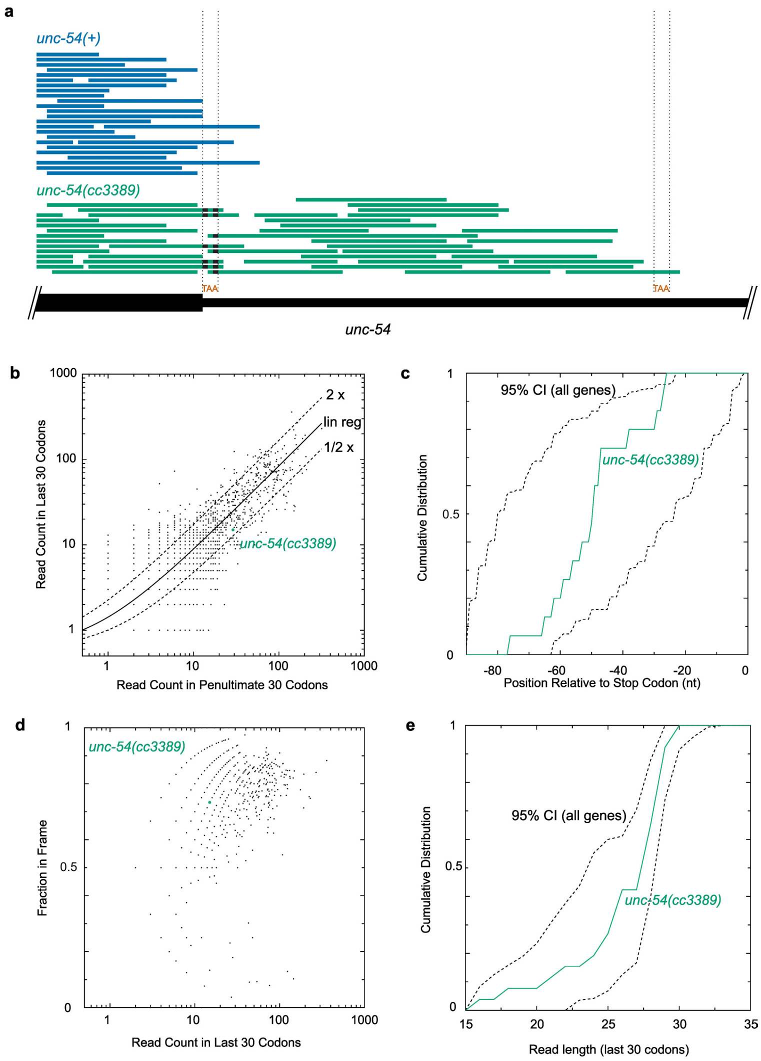
Extended Data Figure 3 | Readthrough regions confer a loss of superfolder GFP fluorescence. Each of the indicated TerByP regions were inserted downstream of superfolder (sf) GFP, upstream of the *let-858* 3′ UTR. TerByP is the region after the annotated stop codon, up to and including the first in-frame stop codon in the 3′ UTR. Quantification was performed as described in Extended Data Fig. 2.


rpl-14(TerByP)
VLFL>RSCA
tbb-2(TerByP)
unc-54(TerByP)
1. Extract codons and shuffle
...
2. Iteratively select codons until stop codon selected
3. This sequence is shuffle1
4. Repeat steps 1 and 2 twice more to generate shuffle2 and 3
shuffle1
shuffle2
shuffle3

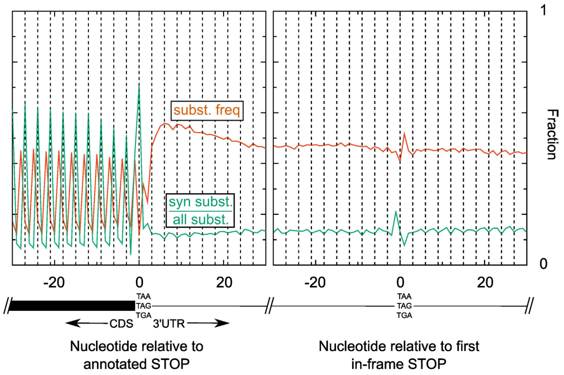
Extended Data Figure 4 | Explanation of 'shuffle' sequences. Trinucleotide codons from each TerByP region are colour-coded by gene (top). Codons were extracted and randomly shuffled in python. A codon was iteratively selected until a stop codon was encountered, defining shuffle1. The process was repeated twice more to define shuffle2 and shuffle3. The resulting shuffle peptides are a combination of all three TerByP regions. Lengths and colour-coding of codons for shuffle1–3 accurately reflect the sequences they are derived from.



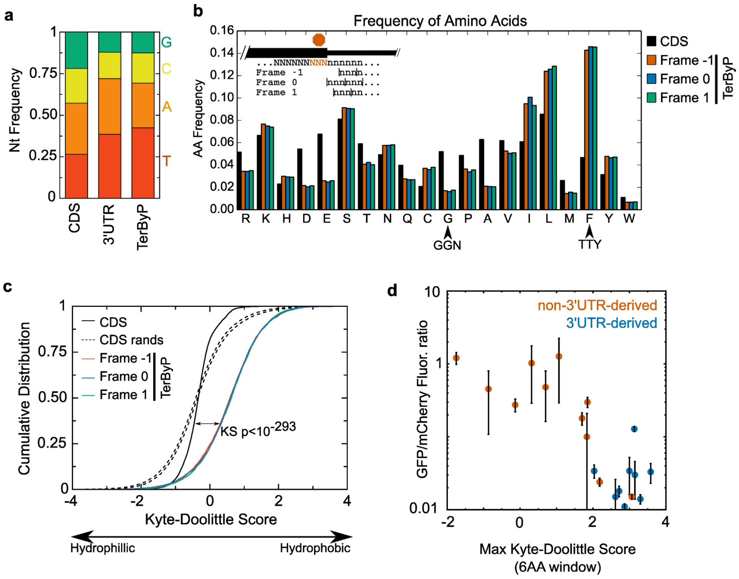
Extended Data Figure 5 | RNA-seq and ribo-seq from *unc-54* mutants. **a**–**c**, RNA-seq (**a**) and ribosome footprint profiling (ribo-seq) (**b**) library mRNA counts, with summary counts (**c**) for the indicated strains and mRNAs. Libraries were prepared from L4 animals, as described Methods. 'N2' is wild type (PD1074, VC2010 (ref. 53)). *unc-54(cc3389)* bears a TAA (stop) to AAT (Asn) mutation, *unc-54(TerByP)*. *unc-54(e1301)* has a GGA (Gly387) to AGA (Arg387) point mutation that confers a temperature-sensitive Unc phenotype with minimal discernible effects on UNC-54 protein levels. *unc-54(e1301)* was included as a control for the Unc phenotype of *unc-54(cc3389)*, though *e1301* confers a less severe Unc phenotype than *cc3389*. Values for *unc-54* mRNA (blue) are highlighted throughout, and for comparison, three additional transcripts known to be at least partly expressed in the body-wall muscles are also highlighted: *unc-87* (pink), *unc-15* (green), and *unc-22* (red).



Extended Data Figure 6 | Ribo-seq of *unc-54(cc3389)* shows an unexceptional progression of ribosomes in the readthrough region. **a**, Raw ribo-seq reads for *unc-54*(+) (blue) and *unc-54(cc3389)* (green) animals, plotted as read pile-ups. Mismatched bases are indicated with black bars. Location of the normal stop codon and the first in-frame stop codon are indicated with 'TAA' and dotted lines. The extension in *unc-54(cc3389)* is 30 amino acids. **b**, The number of ribo-seq reads in the last 30 codons, compared to the previous 30 codons, for all mRNAs. Linear regression was performed on all points (solid line), and twofold difference shown (dashed lines). **c**, The distribution of ribo-seq reads in the last 30 codons (90 nt) of *unc-54(cc3389)* is shown in green, and the 95% confidence interval (CI) for all open reading frames in dashed lines. **d**, The fraction of in-frame ribo-seq reads in the last 30 codons is plotted as a function of read counts in the last 30 codons, and *unc-54(cc3389)* highlighted. **e**, The distribution of read lengths in the last 30 codons of *unc-54(cc3389)*, and all open reading frames (95% confidence interval, dashed lines). For **b**–**d**, reads were restricted to 28, 29, 30 nt lengths. For **b**–**e**, a 12 nt offset was performed for the ribosomal *P*-site, and read counts were derived solely from the *unc-54(cc3389)* ribo-seq library. For **c** and **e**, a minimum 15 read counts was imposed to obtain the 95% confidence interval from 'all genes'.



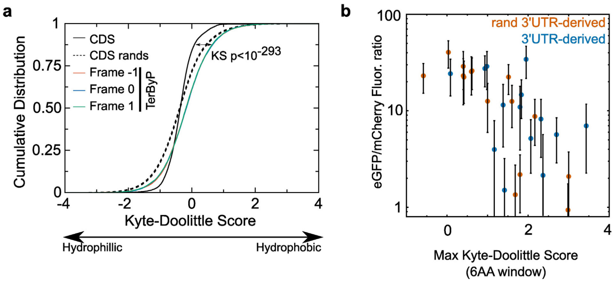
Extended Data Figure 7 | Lack of general conservation of coding potential downstream of stop codons in *Caenorhabditis*. Whole-genome alignment of six nematode species with *C. elegans* genome assembly ce10/WS220 was obtained from the UCSC genome browser. For each annotated transcript, the aligned bases from the multiple species alignment were extracted and compared to the reference (*C. elegans*) genome. The left plot shows summary information of the alignment centred on annotated stop codons; the right plot shows the same centred on the first in-frame stop codon in 3′ UTRs. In red is the substitution frequency, that is, the number of mismatched bases divided by the number of aligned bases at a given position. The enrichment of 'wobble' position mutations is apparent as an increase in substitutions at the third position of each codon in the CDS. In green is the synonymous substitution frequency, that is, for codons beginning at a given position, the fraction of mutations that yield a synonymous substitution divided by all mutations at that position (synonymous and non-synonymous). The tendency to conserve amino acids in the CDS is apparent as a green spike at every in-frame codon. The change in substitution frequency and synonymous substitution frequency about the first in-frame stop codon (right plot) is due to a tendency for NTR codons to be conserved, and for AAN/AGN/GAN codons to not be conserved in 3′ UTRs, regardless of frame.



e

Gene	Genotype	Phenotype	Avg Kyte-Doolitte	Max Kyte-Doolitte (6AA window)
unc-22				
	unc-22(TerByP)	**wild type**	**0.370**	**1.950**
	unc-22(-1,TerByP)	twitcher	1.081	3.283
	unc-22(+1,TerByP)	twitcher	1.685	1.333
unc-54				
	unc-54(TerByP)	paralyzed	0.928	2.866
	unc-54(+1,TerByP)	paralyzed	1.391	3.150
	unc-54(-1,TerByP)	**weak Unc**	**-0.088**	**1.550**

Extended Data Figure 8 | Nucleotide and amino acid composition of readthrough regions (*C. elegans*). CDSs and 3′ UTRs were analysed for various sequence properties. For simplicity, only genes and 3′ UTRs for which a single 3′ UTR was annotated were considered. Similar results were obtained with genes with multiple 3′ UTRs. **a**, Nucleotide frequency of CDS, 3′ UTR, and TerByP (region between annotated stop codon and first in-frame stop codon). **b**, Frequency of amino acids in all three possible frames for the TerByP region. 3′ UTRs were translated one codon past the stop codon of the CDS until the next in-frame stop codon, with nonstop 3′ UTRs ignored. Highlighted are codons with high G content (GGN, Gly) and high T content (TTY, Phe). **c**, TerByP regions tend to be hydrophobic, regardless of frame. Kyte–Doolittle score was used as a measure of hydrophobicity[54]. To reduce noise, only TerByP regions at least 10 amino acids long were considered. *P* value is from Kolmogorov–Smirnov test comparing CDSs and TerByP sequences (each frame has *P* value < 10e-293 for this comparison). As the TerByP sequences are shorter than CDSs on average, the distribution of TerByP hydrophobicity scores will tend to have higher variance than CDSs. Random portions of CDSs were taken, length-matched to TerByP frame zero peptide lengths. This was repeated 100 times, and the 95% confidence interval is shown (dashed lines, 'CDS rands'). **d**, Hydrophobicity of the inserts is correlated with a negative effect on GFP fluorescence. The GFP:mCherry fluorescence ratio (Fig. 2b) was plotted against the maximum Kyte–Doolittle score in a six amino acid window for each insert. (Similar results were obtained using the Kyte–Doolittle score averaged across the entire sequence.) Mean (circle) and s.d. (bars) are shown. 3′-UTR-derived sequences are in blue, and non-3′-UTR-derived sequences are in red. To avoid redundancy or skewing of the data, in cases where multiple constructs were present with the same peptide sequence (for example, *unc-54(TerByP)*, *unc-54(TerByP, syn1)*, and *unc-54(TerByP, syn2)*), only the first of these was used. **e**, Hydrophobicity analysis of the TerByP extensions obtained by CRISPR–Cas9 engineering at the *unc-22* and *unc-54* loci. '+1/-1 TerByP' denotes the gain or loss of a nucleotide, generating a late frameshift and allowing translation to proceed past the annotated stop codon out of frame with the upstream open reading frame. In each case, Kyte–Doolittle hydropathy was used to analyse the C-terminal appendage. The least phenotypically affected strain of the three is shown in bold.



Extended Data Figure 9 | Nucleotide and amino acid composition of readthrough regions (*H. sapiens*). **a**, **b**, Similar analysis of hydrophobicity as in Extended Data Fig. 8c, d, performed in humans.

Extended Data Table 1 | Translation into 3′ UTRs at endogenous loci tends to yield hypomorphs

Gene	Genotype	Phenotype	Isolates
pha-4			
	pha-4(TerByP)	wild type	2
unc-22			
	unc-22(TerByP)	wild type	2
	unc-22(-1,TerByP)	twitcher	1
	unc-22(+1,TerByP)	twitcher	1
	unc-22(unc-22::GFP)	wild type	3
unc-45			
	unc-45(TerByP)	emb lethal	3
	unc-45(3xFLAG::TEV::3xHA)	wild type	2
tra-2			
	tra-2(TerByP)	XX males	2
	tra-2(3xFLAG)	wild type	2
	tra-2(3xFLAG::TEV::3xHA)	XX males	2
unc-54			
	unc-54(TerByP)	paralyzed	2
	unc-54(unc-54::gfp)	wild type	2
	unc-54(+1,TerByP)	paralyzed	1
	unc-54(-1,TerByP)	weak Unc	1

CRISPR–Cas9 editing[17] was used to construct the mutations shown. See Supplementary Table 1 for precise nucleotide sequences of all strains. '−1/+1 TerByP' indicate the loss or gain of one nucleotide relative to the zero frame, generating a frameshift over the stop codon, and translation into the 3′ UTR out of frame with the coding sequence. For *unc-22*, 'wild type' indicates a lack of twitching, even in 1 mM levamisole. For *unc-54(−1,TerByP)*, 'weak Unc' animals were visibly slower than *unc-54*(+), but faster than *unc-54(TerByP)*. All mutant phenotypes were recessive.