# Evidence of Abundant Stop Codon Readthrough in Drosophila and Other Metazoa

Irwin Jungreis[1,2], Michael F. Lin[1,2], Rebecca Spokony[3], Clara S. Chan[4], Nicolas Negre[3], Alec Victorsen[3], Kevin P. White[3], Manolis Kellis[1,2,5]

*[1]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA; [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02139, USA; [3]Institute for Genomics and Systems Biology, University of Chicago, Chicago, IL 60637, USA. [4]MIT Biology Department, Cambridge, Massachusetts 02139, USA.*

***[5]Corresponding author. E-mail manoli@mit.edu; fax 617-452-5034***

## Supplementary Figures:

**Figure S1. Second ORF lengths**. Fraction of in-frame second ORFs of given length (x axis) that are readthrough candidates (red), that show protein-coding constraint as indicated by positive PhyloCSF scores but with alternative plausible explanations (blue), or that have negative PhyloCSF score (green). Counts for each category indicated for each interval. Even among second ORFs greater than 100 amino-acids, only a quarter are readthrough candidates, highlighting the need for comparative evidence to distinguish them. Overall, 2% of second ORFs tested are readthrough candidates.

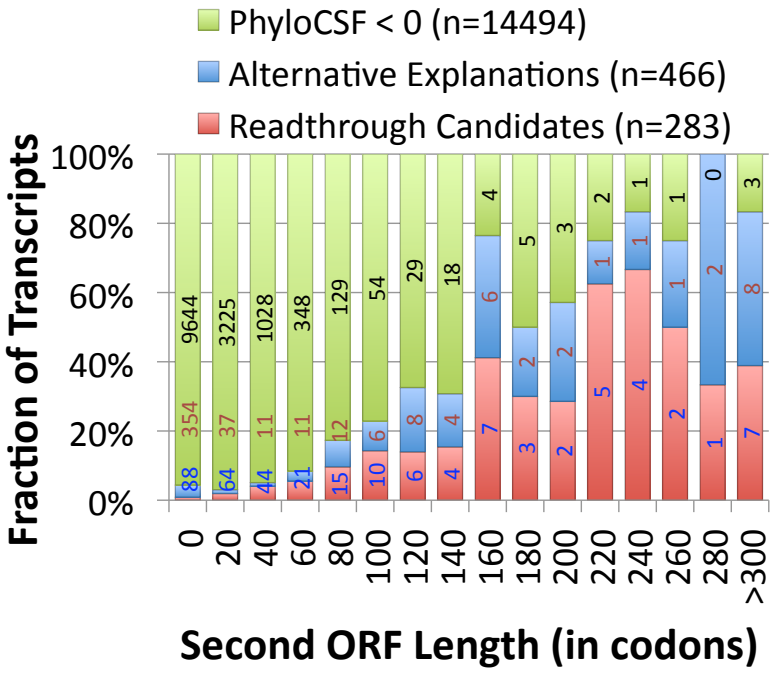


**Figure S2. Periodic base pairing without smoothing.** Identical to Figure 4C except without averaging sets of five consecutive codons. Although periodicity in the readthrough region is more difficult to discern visually without this smoothing, the pairing frequency at the third codon position remains higher than at the second position for almost all codons.

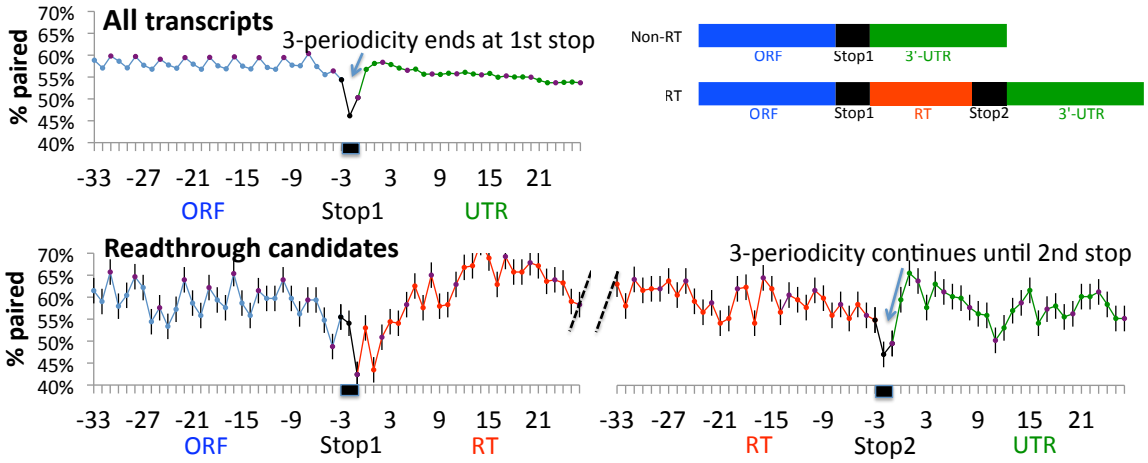


**Figure S3. Close synonymous codons and ribosomal hopping.** Distribution of distances (in bp) between the closest pair of synonymous in-frame codons bracketing the stop codon, for the readthrough candidates and for all transcripts. Error bars show Standard Error of Mean (SEM). The closest synonymous codons bracketing the stop codons of the readthrough candidates are not significantly closer than those of other transcripts, whereas we would expect them to be if ribosomal hopping was programmed in many of these transcripts.

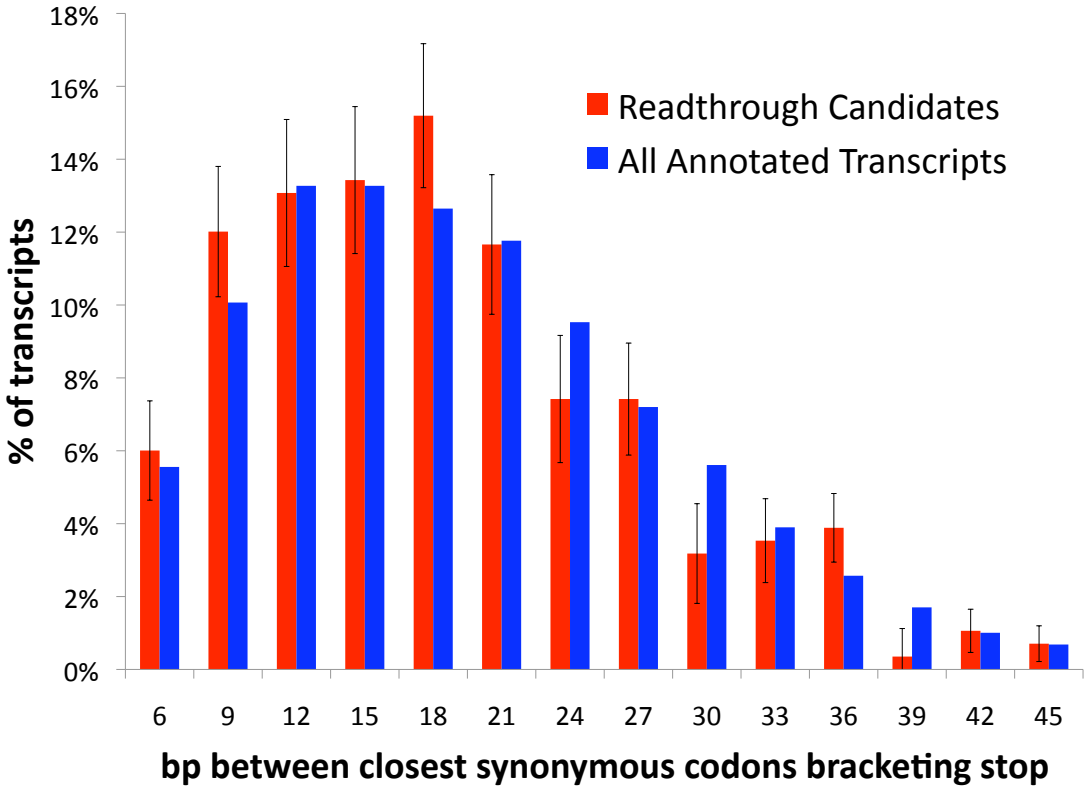


**Figure S4. Stop codon context rules out splicing.** The 4-base contexts are sorted in order of decreasing frequency among the 14,928 non-readthrough stop codons (blue). Context frequencies of the 1315 annotated "spliced-out-stops" (yellow), i.e., annotated stop codons that are within an intron of another annotated transcript of the same gene, are similar to those of the non-readthrough stop codons. Context frequencies for readthrough candidates (red) are opposite of non-readthrough transcripts and of *annotated* spliced-out-stops, whereas if the readthrough candidates were in fact *unannotated* spliced-out-stops we would expect their context frequencies to match those of the annotated ones, providing additional evidence that the evolutionary signatures of our readthrough candidates are not due to splicing.

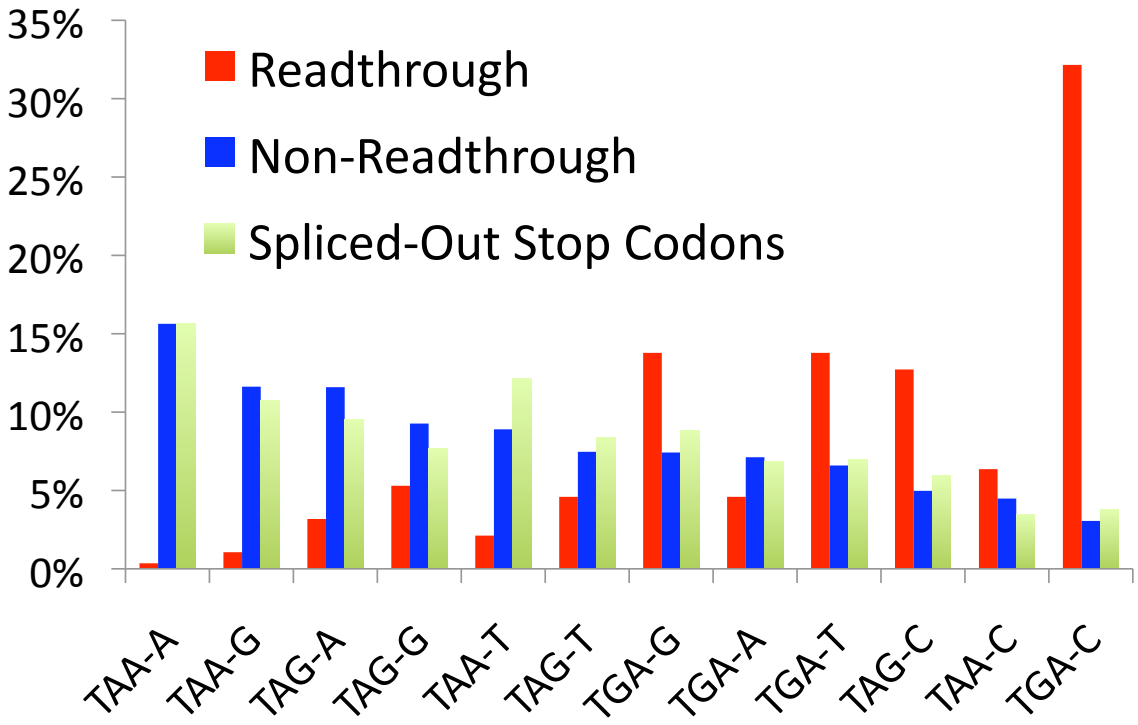
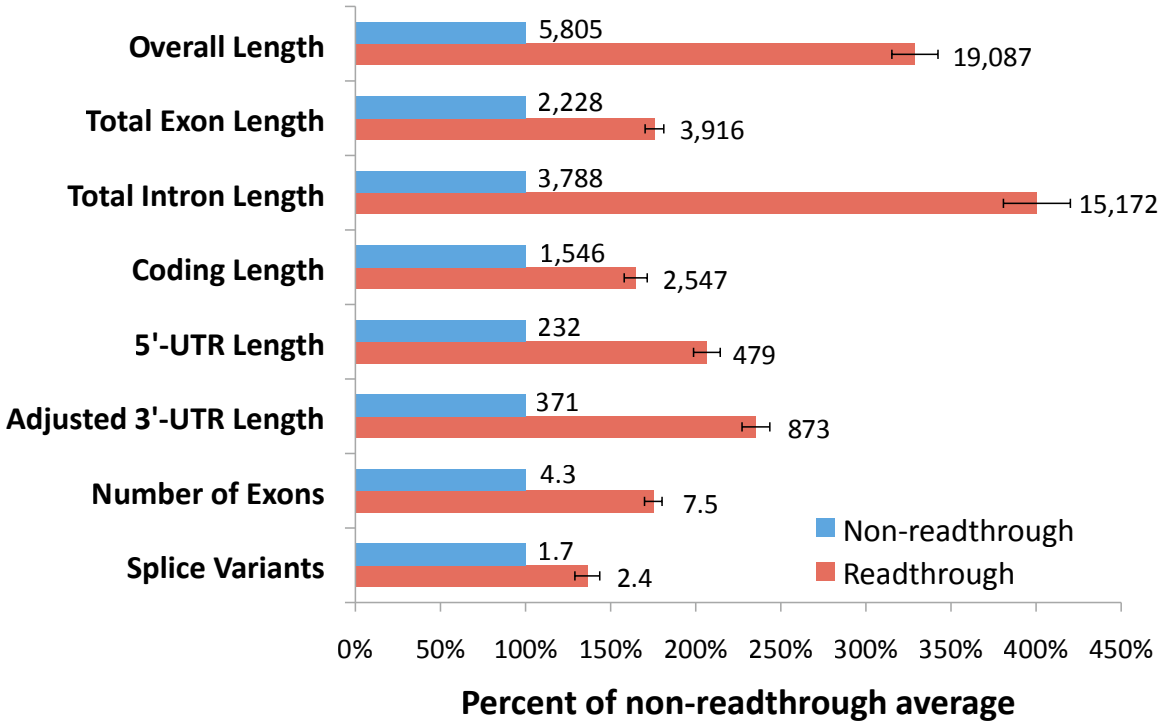


**Figure S5. Length enrichments.** Readthrough candidates as a group (red) show increased length (in nucleotides) and splicing complexity (number of transcript variants) compared to other transcripts (blue). Overall Length: sum of all introns, coding exons, and UTRs. Coding Length: length of first ORF. Adjusted 3'-UTR Length excludes second ORF for readthrough candidates. Splice Variants: number of annotated transcripts of the same gene. Numbers show actual size and bars denote size relative to non-readthrough genes. Error bars show standard error of mean (*SEM*) for readthrough candidates. Enrichments persist even after correcting for total intron length, 5'UTR length, 3'UTR length, or coding exon length (not shown).



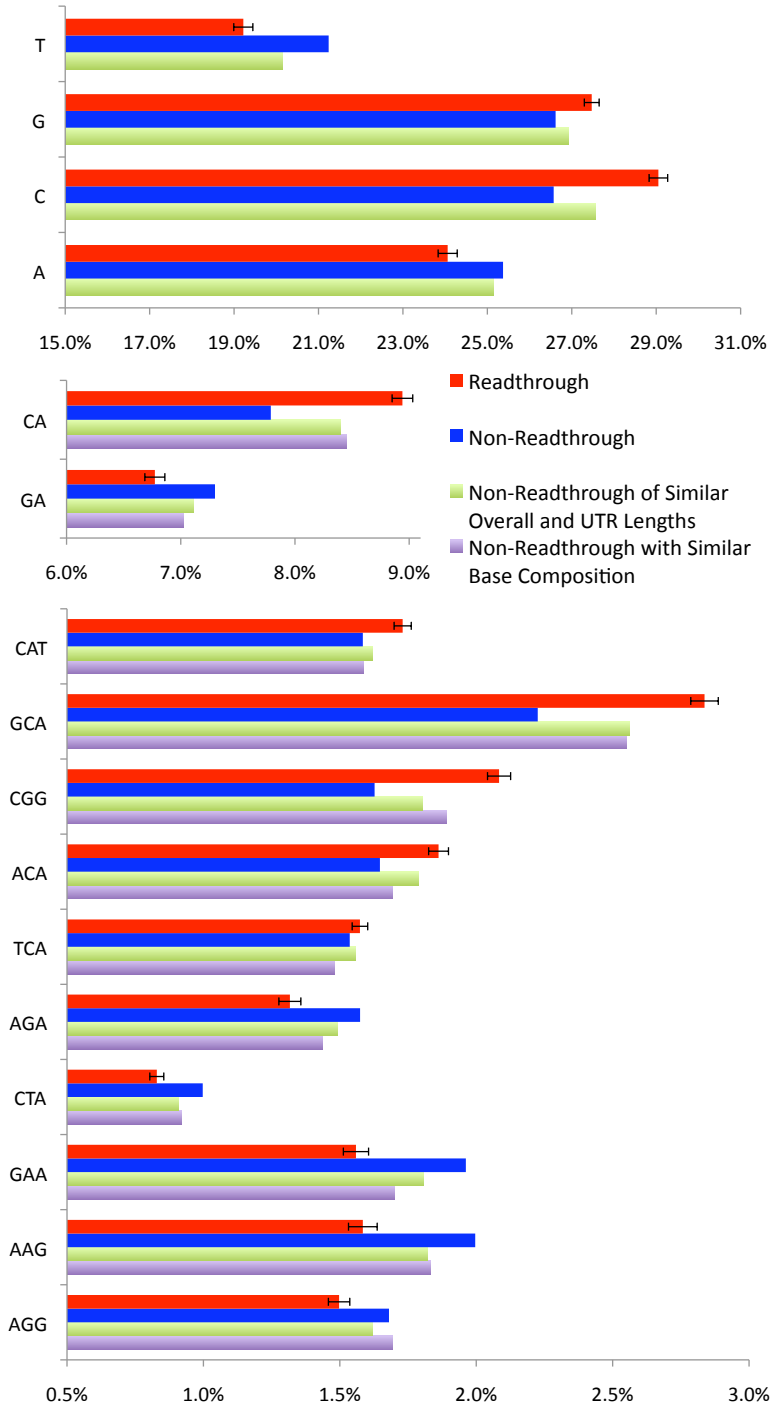



**Figure S6. Unusual 1-, 2-, and 3-mer composition.** Coding regions of readthrough candidates have unusual base composition and are enriched or depleted for some 2- and 3-mers. Average frequency of bases and significant 2- and 3-mers in the coding regions of readthrough candidates (red), non-readthrough transcripts (blue), non-readthrough transcripts with similar primary transcript and UTR lengths to the readthrough candidates (yellow), and non-readthrough transcripts with similar base composition to the readthrough candidates (purple). The 2- and 3-mers shown are those for which the frequencies among readthrough candidates are significantly different from non-readthrough transcripts with similar base composition. Error bars show the Standard Error of Mean (*SEM*) of the readthrough candidates. Correcting for the excess length of readthrough candidates decreases but does not remove the biases, and similarly, correcting for the high GC content of readthrough candidates decreases but does not remove the 2- and 3-mer biases.

**Figure S7. Length distribution biases may confound GO enrichment analysis.** Distribution of total gene length (log scale) for readthrough candidates (red), non-readthrough transcripts (blue), and genes with the GO category "neurogenesis" (green), one of the most enriched among readthrough candidates. Readthrough candidates and neurogenesis genes have very similar length distribution, which would lead to artifactual enrichment if their above-average lengths are for independent reasons. Alternatively, since length could be an indication of more regulatory binding sites, it is possible that neurogenesis genes show increased regulation of many types, only one of which is readthrough.

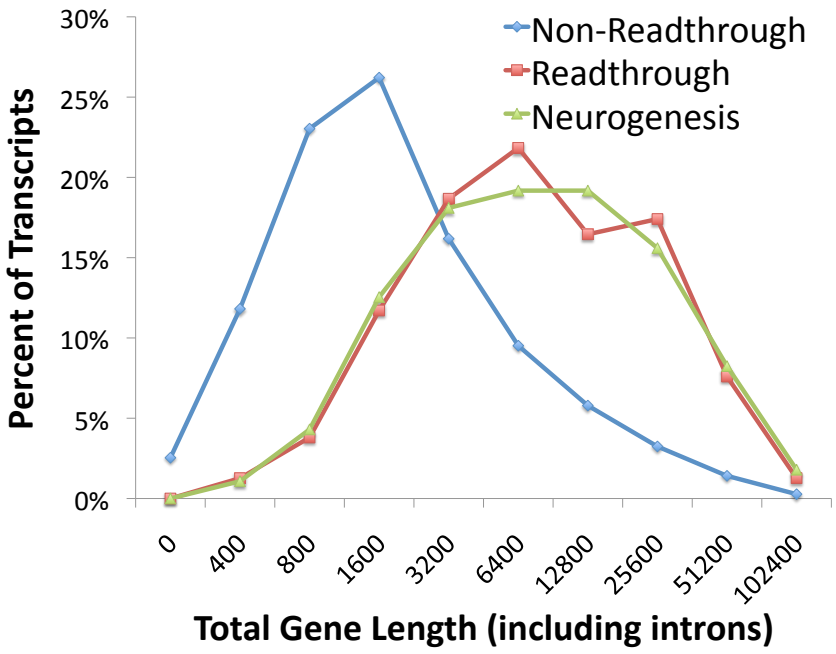
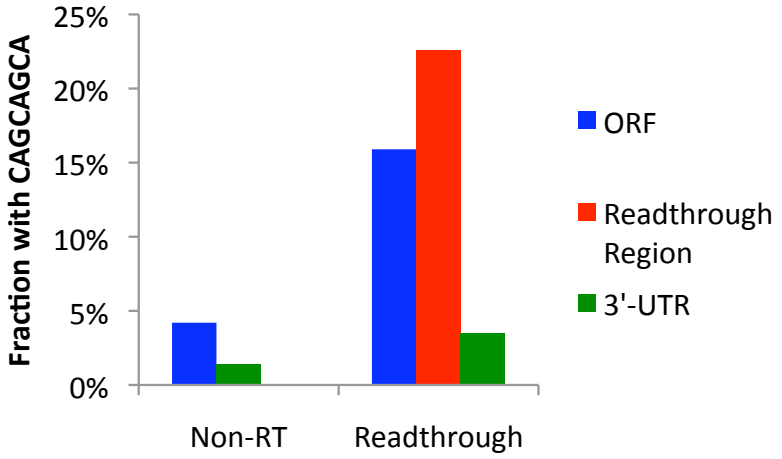


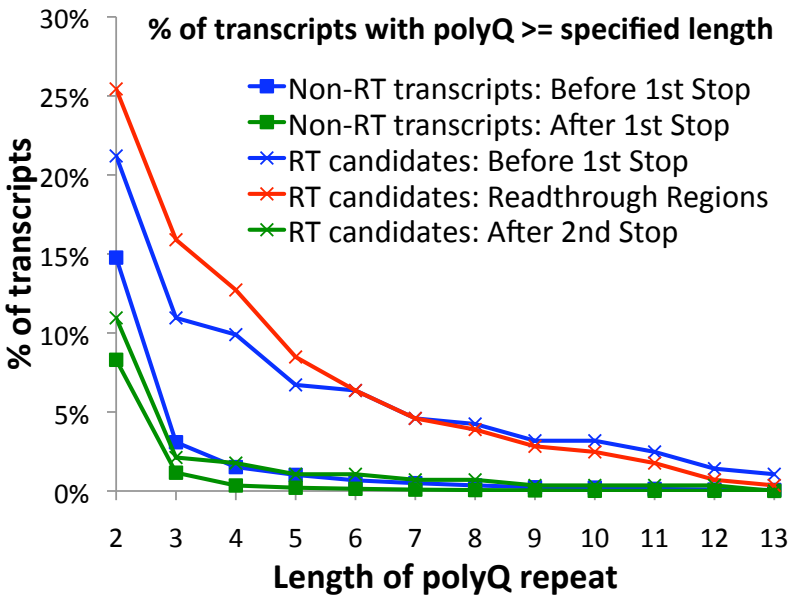
**Figure S8. Enriched 8-mer**. Readthrough regions and corresponding regions before the first stop are highly enriched for the 8-mer CAGCAGCA and long poly-Q repeats. (A) Fractions of readthrough candidates that include the 8-mer CAGCAGCA in readthrough regions, same length regions before the 1st stop (ORF), and same length regions after the second stop (3'-UTR), and the fractions of non-readthrough transcripts with this 8-mer in the regions before and after the 1st stop (region length for non-readthroughs chosen to match the length distribution of the readthrough regions). Readthrough candidates are highly enriched, both before and after the readthrough stop codon, compared to coding or non-coding regions of non-readthroughs. Most of the enrichment persists even among transcripts with similar base composition, or similar frequency of the 3-mer CAG or GCA (not shown). (B) Poly-Q length. Fraction of transcripts for which the specified region contains a sequence of consecutive Q codons of at least the specified length. The regions are the same as for (A). Many of the readthrough candidates have unusually long polyQ repeats either in the readthrough region or region before the 1st stop. The enrichment persists even if they are compared to non-readthroughs with similar frequency of the codon CAG (not shown). (C) Enrichment of 8-mer before the stop. The fraction of transcripts containing CAGCAGCA in 125-base windows before the first stop among readthrough candidates, non-readthrough transcripts, and non-readthroughs with similar frequencies of the 3-mers CAG, AGC, and GCA. The readthrough candidates are significantly enriched for the 8-mer in the last 250 nt before the stop codon, even relative to other transcripts with similar frequencies of these 3-mers. Enrichment of these 3-mers explains the enrichment of the 8-mer in the earlier 90% of the coding region, but not the excess enrichment near the 3' end.

**Figure S8. Enriched 8-mer**.

A




B




C

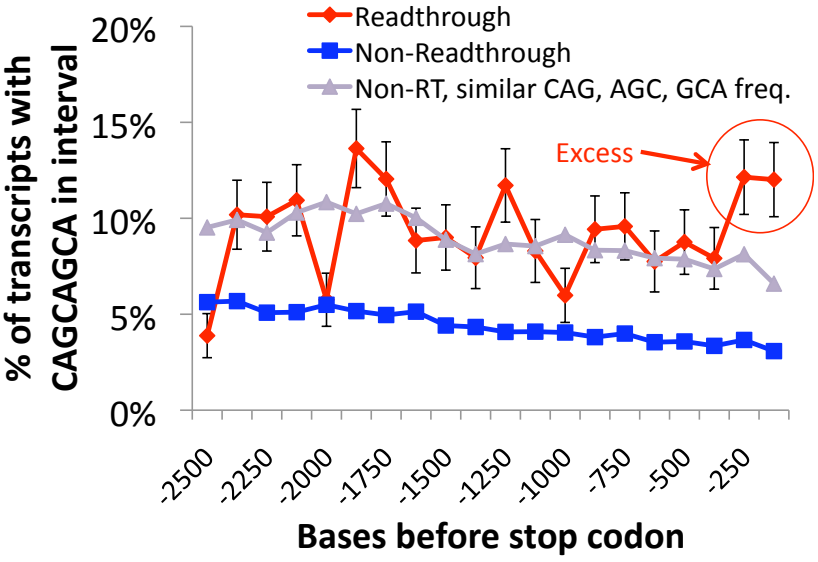


**Figure S9. Evolution by extension.** Second ORF of *HLH4C*, one of our readthrough candidates that might have evolved by the extension scenario since the common ancestor of the 12 flies. Since second ORFs of genes that evolved this way will not be protein coding in some species, we would generally not expect to find them by our method, so they would not be readthrough candidates even if they are read through in *D. melanogaster*.

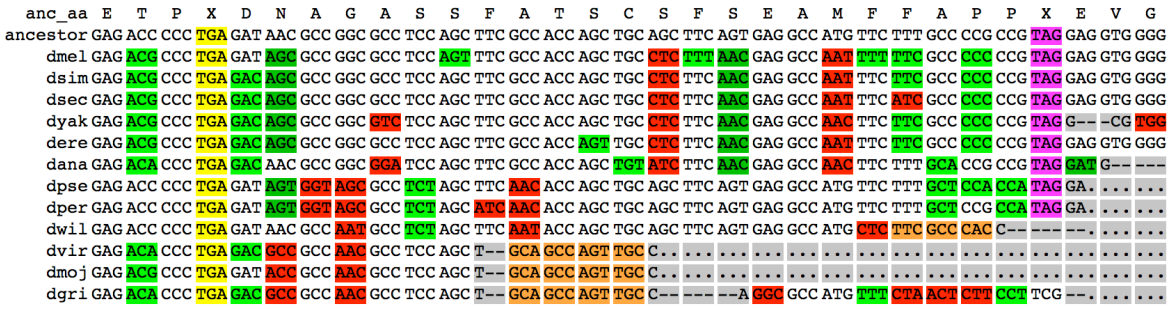


**Figure S10. Rabbit beta-globin.** Mammal alignment after the first stop codon for the known readthrough gene, rabbit beta-globin (some species not shown). The readthrough region in rabbit is not conserved, even compared to its closest relatives among the aligned species, so readthrough could be a species-specific event.

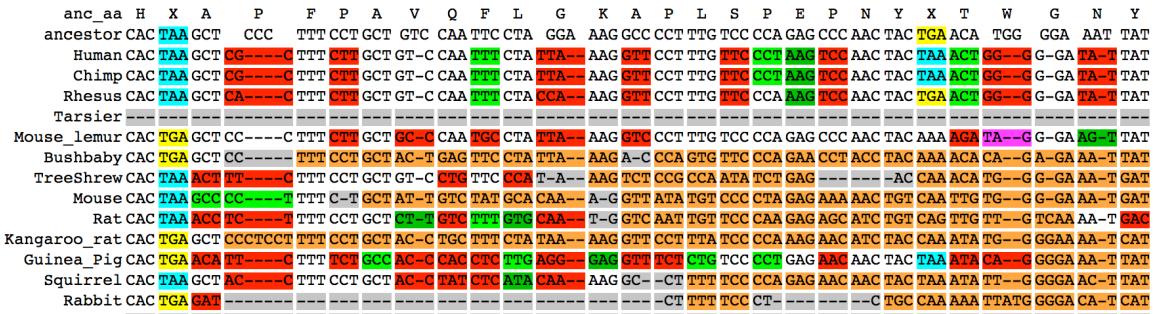


**Figure S11. Estimated extent of readthrough in *C. elegans* using comparative evidence.** Frequency (y-axis) of PhyloCSF scores per codon (x-axis) for *C. elegans* second ORFs in three frames. In contrast to the similar analysis for *D. melanogaster* (Figure 5A) no significant excess is found in frame 0, suggesting lack of abundant readthrough in *C. elegans*. Scores of five readthrough candidates found by manual curation are indicated with stars.

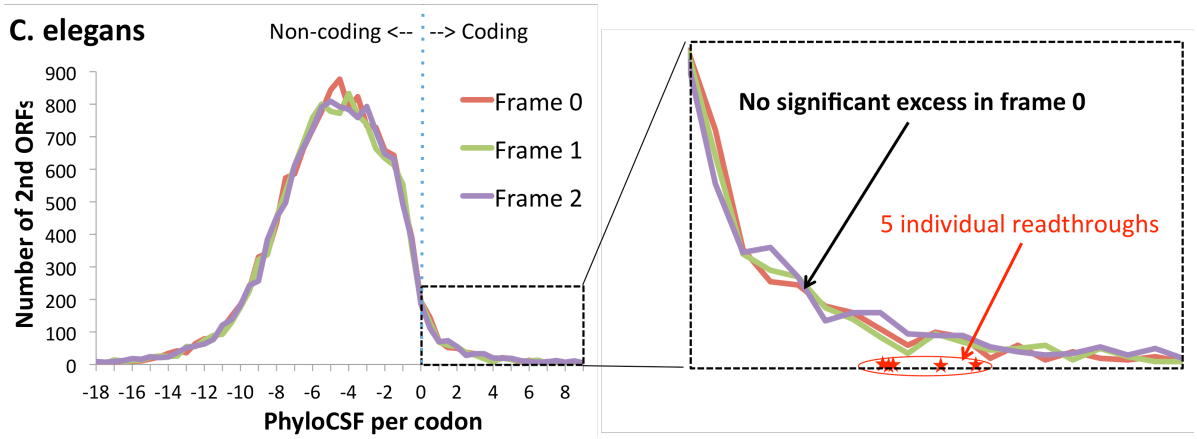
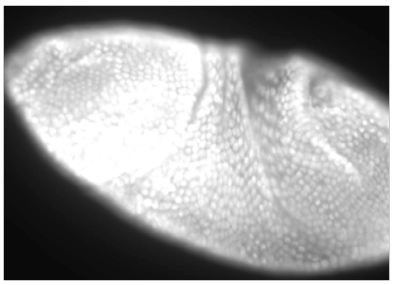


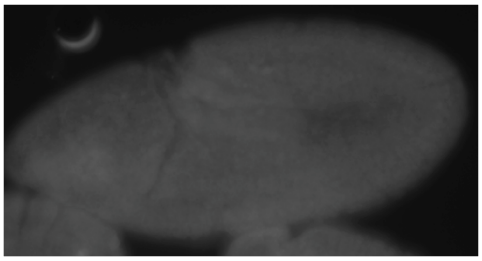
**Figure S12. GFP images with controls.** GFP images of the readthrough strains shown in Figure 3B matched to wildtype embryos or larva with similar development stage and orientation (*y*$^{1}$*; cn*$^{1}$ *bw*$^{1}$ *sp*$^{1}$ strain*)*. Although some staining is seen in the control images, it is neither as bright nor as specific as in images of the readthrough strains.

Readthrough Strains

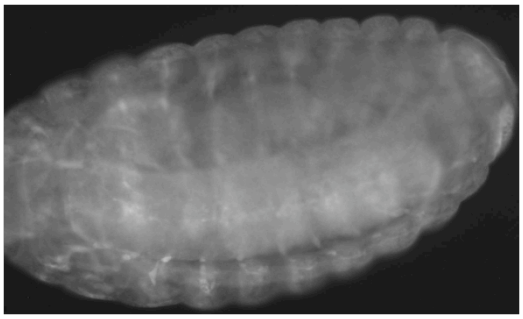


z-RT embryo 3-4 hours

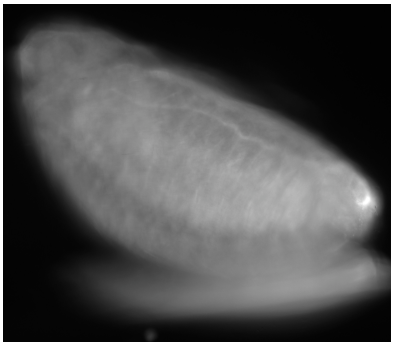
Control Strain



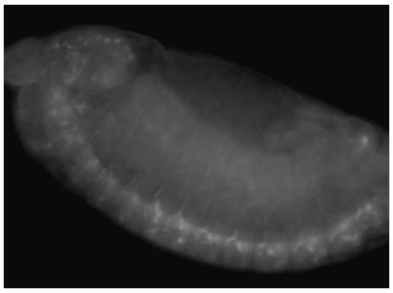
control embryo 3-4 hours



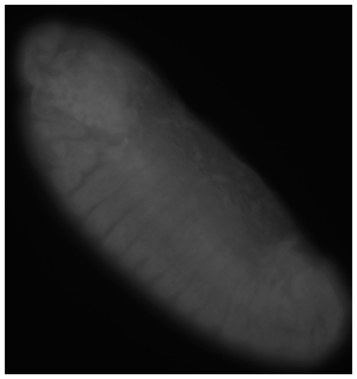
Sp1-RT embryo 13-22 hours



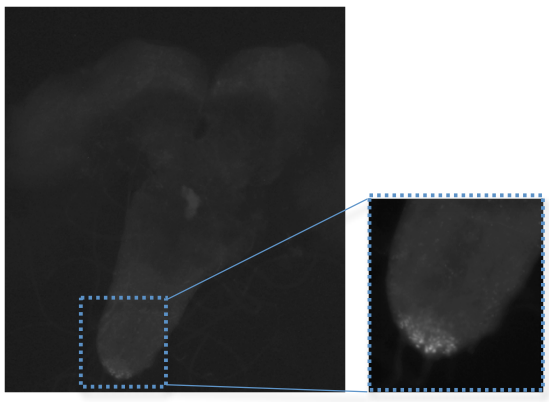
control embryo 13-22 hours



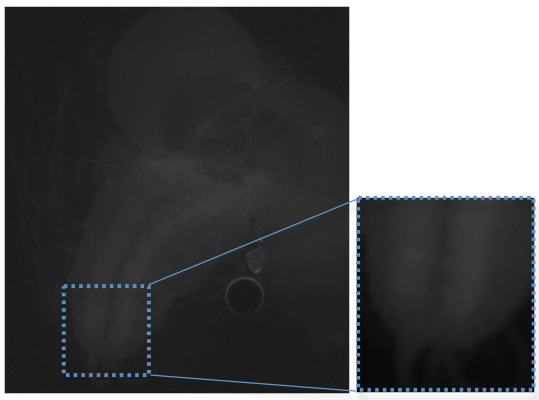
cnc-RT embryo 10-13 hours



control embryo 10-13 hours



Abd-B-RT larva cns



control larva cns