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 Text:

S1 Downstream translation is not a result of RNA editing

We evaluated whether translation beyond the stop codon is a result of A to I editing, which would convert stop codons to UGG tryptophan codons, by analyzing deep RNA-seq data from the modENCODE project (Graveley et al. 2011) for A-to-G single-nucleotide differences between the DNA sequence and the mRNA reads, which provide a median coverage of 1500 reads in the regions evaluated.

Among all annotated stop codons we found only three that are edited to UGG: *DopEcR*, *qvr*, and *CG8481*. The first results in a previously-reported (Stapleton et al. 2002) two-codon extension with positive PhyloCSF score, but below our significance threshold. The other two were also not in our candidate list, as their second ORFs have negative PhyloCSF scores and are unlikely to correspond to functionally-selected translation.

Outside these three stop codons, showing 193, 708 and 601 edited reads respectively, and corresponding to 51%, 41% and 13% of all reads at those loci, we found no evidence of editing. At every other stop codon there were fewer than 4 A-to-G mismatched reads, or such mismatched reads accounted for fewer than 4% of all reads at that locus, with mismatches at candidate readthrough stop codons no higher than at other stop codons, and we believe these are due to sequencing errors.

Other (non-A-to-I) mismatches affecting stop codons also appear to be due to sequencing or mapping errors as well. After excluding the 3 A-to-I editing sites and likely mapping errors, we found 8 stop codon loci at which the number of reads with any mismatch exceeds 100 and the fraction of mismatches exceeds 10%, the highest being 18%. For six of these, the mismatches were A-to-C. In each case, almost all of the mismatched reads were sequenced in the same direction, while sequencing in the other direction produced almost no mismatches, which was *not* true for the 3 A-to-I cases, so we believe the mismatches are due to sequencing errors.

We conclude that downstream translation of our readthrough candidates is not due to RNA editing.

S2 Additional details of manual curation

This section provides further details on how and why we excluded transcripts during manual curation, what evidence we have against alternative explanations for the remaining candidates, and an estimate of PhyloCSF false positives.

We excluded transcripts for which the first stop codon (in *D. melanogaster*) is aligned to a sense codon in most of the other species, because it could have been a sense codon that mutated to a stop codon recently enough that the downstream region continues to display a protein-coding signature even though it is no longer translated.

We excluded transcripts for which the second ORF has a higher score on the opposite strand, when offset so as to align the third codon positions, because in that case it is possible that the opposite strand is in fact protein-coding, and the strand we were considering has a high PhyloCSF score merely because substitutions are concentrated in the third codon position.

We excluded transcripts for which the PhyloCSF score from the stop up to the first ATG codon was low, with a more conservative threshold if the ATG codon was well conserved, in order to eliminate cases where the downstream coding signature is due to translation initiation at an ATG codon within the second ORF, either because of an alternative transcription start site or dicistronic translation at an Internal Ribosome Entry Site (IRES). While this cannot rule out the possibility that translation sometimes initiates at an internal ATG in one of the remaining second ORFs, such initiation would not explain the protein coding signature before the ATG, so some other explanation, such as readthrough, would be required. Because there are *Drosophila* genes known to initiate translation at a GTG start codon (Sugiharas et al. 1990; Takahashia et al. 2005) or a CTG start codon (DeSimone and White 1993), we verified that for all but 44 of the readthrough candidates the PhyloCSF score to the first CTG or GTG is positive, so initiation at these codons could not explain the positive PhyloCSF score. Furthermore, because alternative translation start sites could occur in any of the three frames, the excess of positive scores in frame 0 (Figure 5A) shows that most of the evolutionary signatures are not due to such start sites.

We used several lines of investigation to verify that the readthrough in our candidates is not caused by selenocysteine insertion. First, we found that the first stop codon is TAA or TAG for 101 of our candidates, which cannot trigger selenocysteine insertion. Next, we ran an existing program, SECISearch (Kryukov et al. 2003), to search for SECIS elements in the 3'-UTRs of the remaining 182 (TGA) readthrough candidates, as we had done on our previous set of candidates (Lin et al. 2007), and did not find any with COVE score above the recommended threshold of 15. (Only 3 had COVE score above 0, the largest being 3.73.) Finally, we note that one of the species in our comparisons, *D. willistoni*, is known to lack most of the genes implicated in selenoprotein synthesis, and substitutes a cysteine codon for the TGA stop at all known selenocysteine insertion sites (Chapple and Guigo 2008) and yet it has an aligned TGA stop codon for almost all of our TGA readthrough candidates and retains a protein-coding signature in the second ORF. Moreover, readthrough in *kel* has been experimentally observed in *D. willistoni* (Robinson and Cooley 1997), confirming that the readthrough mechanism is functional. In any case, an exhaustive experimental search for selenocysteine proteins in *D. melanogaster* using radioactive selenium found there to be only 4 such proteins (Martin-Romero et al. 2001), one of them possibly due to bacterial contamination. (The other 3 are annotated as selenocysteine, and are not readthrough candidates.)

To estimate the number of readthrough candidates whose second ORF is a non-coding region that has a high PhyloCSF score simply by chance, we summed the Local False Discovery Rates of the readthrough candidates (Efron et al. 2001). This gave us an estimate, based on p-values calculated from the PhyloCSF score alone, of 34 (12%). However, when compiling the readthrough list we excluded transcripts with borderline p-values unless we had some independent evidence that they are protein-coding that is not captured by PhyloCSF (such as a lack of frame-shifting indels). Thus the transcripts in the readthrough list with largest p-values (hence largest Local FDRs) are more likely to be protein-coding than is indicated by their PhyloCSF score alone. Consequently the actual number of false positives due to chance is probably much less than 34.

We also investigated whether the readthrough candidates could have been detected by the simpler method of looking for long second ORFs. However, we found that while the longest second ORFs are more likely to be readthrough regions, only 51 (25%) of 202 in-frame second ORFs longer than 100 codons are readthrough candidates (Supplementary Figure S1), whereas 232 (1.5%) of 15031 shorter in-frame second ORFs are readthrough candidates, emphasizing the importance of using comparative evidence and our additional metrics.

S3 Evidence that translation of PeptideAtlas matches is not due to alternative splicing or dicistronic translation

Because of the importance of the mass spectrometry-supported readthrough candidates, and because for several of them our evidence suggests changes to previous annotations, in this section we review the specific evidence found during manual curation that second ORF translation in these candidates is more likely due to readthrough than alternative splicing or dicistronic translation.

There were nine readthrough candidates whose readthrough regions had matches to PeptideAtlas peptides, seven of which were not previously annotated as readthrough. Because almost all of the peptides in the database were sequenced by matching to annotated coding regions, six of these seven were found in parts of readthrough regions that are or were previously annotated as protein coding. One of these, CG34318, is annotated as dicistronic, and several of the others are annotated as alternative splice variants.

In each case we verified that between the stop codon and the peptide there are no 3'-splice sites supported by cDNA or EST evidence reported by the UCSC genome browser. Although some of these peptides are within coding exons of currently or previously annotated splice variants, these variants are unsupported by such evidence. Furthermore, none of the modENCODE RNA-seq reads supported the existence of a splice site in any of these regions. There were a minimum of 139 RNA-seq reads covering every locus in each of these readthrough regions, and in most cases several thousand.

For two of the seven, CG42265 and *gish*, there are no ATG codons between the stop codon and the peptide so those peptides cannot have resulted from dicistronic translation (unless it uses a non-standard start codon). While we cannot rule out dicistronic translation in the other five, we observe that in each case the region from the stop codon to the first ATG has a positive PhyloCSF score. This comparative evidence suggests that translation continues from the stop codon, rather than starting at the subsequent ATG.

We conclude that readthrough is the most plausible explanation for translation of the downstream peptides for each of these seven readthrough candidates.

S4 Additional details of frame bias investigation

When using frame bias to estimate the number of readthrough transcripts, we relied on the hypothesis that second ORFs in frame 0 are no more likely to overlap coding exons in the same frame (whether from splice variants, second cistrons, or other genes) than ORFs in the other two frames. We checked this directly for annotated coding exons. The numbers of such overlaps are 92, 133, and 181 for frames 0, 1, and 2, respectively, which supports our hypothesis that there is no preference for frame 0, and in fact suggests an opposite preference. We also found that the reason there are more overlaps in frames 1 and 2 than frame 0 is that a coding exon in frame 1 or 2 can overlap the stop codon itself as well as part of the downstream region (implying that the common genomic region is translated in more than one frame), while a coding exon in frame 0 that overlaps the stop codon would be terminated by the stop codon and thus would not overlap the downstream region.

All regions that overlapped *annotated* coding exons were excluded from the counts we reported in the main text (see Methods), so the hypothesis is only required for the *unannotated* coding exons. While we have no way of directly testing this hypothesis for such as-yet undiscovered coding exons, we see no reason that unannotated coding exons would show a bias towards frame 0 when annotated ones show an opposite bias.

Ribosomal bypassing ("hopping"), in which translation bypasses the region between two synonymous codons, has been offered as an alternative explanation for the downstream protein-coding signals in *Drosophila* (Namy and Rousset 2010). However, the take-off and landing codons need not be in the same frame, so hopping is unlikely to account for the excess in frame 0. Furthermore, hopping efficiency decreases with the length of the bypassed regions (Wills 2010) so if hopping were widespread among our readthrough candidates we would expect close pairs of synonymous codons bracketing the stop codon. However, the closest synonymous codons bracketing the stop codons of the readthrough candidates are not significantly closer than those of other transcripts (Supplementary Figure S3).

S5 Literature on amino acid inserted at readthrough stop codon

Various experiments have determined the amino acids that are inserted when certain stop codons are read through, while others have found a variety of tRNAs can suppress termination of each stop codon in bacteria, yeast, mammals, and plants, though their amino acids are not necessarily the ones that will be inserted when that stop codon is read through (Bertram et al. 2001; Beier and Grimm 2001).

Radioactively labeled amino acids were used to determine that glutamine is inserted when UAA or UAG is read through in vitro in a mammalian virus, while arginine, cysteine, and tryptophan are inserted for UGA (Feng et al. 1990). The arginine and glutamine insertions require G:U base pairing in the first codon position, while others require non-canonical pairing at the third codon position. Glutamine tRNAs have also been found to suppress termination of UAA and UAG in yeast (Lin et al. 1986; Pure et al. 1985; Weiss and Friedberg 1986). On the other hand, protein sequence analysis was used to determine that tyrosine, lysine, and tryptophan are inserted in a ratio of approximately 4 to 2 to 1 at an efficiently suppressed premature yeast UAG codon (Fearon et al. 1994).

Mass spectrometry was used to find that seven different readthrough versions of the rabbit beta-globin protein are created in vivo, corresponding to insertion of serine, tryptophan, cysteine, and arginine, and skipping of the UGA stop codon and 0, 1, or 2 immediately downstream codons (Chittum et al. 1998).

Tyrosine is inserted for a UAG or UAA and tryptophan for UGA in the Tobacco mosaic virus, in vivo (Lao et al. 2009).

In *Drosophila*, readthrough of UGA or UAG at a premature stop codon in two alleles of the *elav* gene produce proteins with different phenotype, suggesting that the inserted amino acid is different for these two stop codons and is functional (Samson et al. 1995).

Termination at a UAG codon can be suppressed by a leucine tRNA in a mammalian cell extract with a viral mRNA, which requires a nonstandard pairing in the second codon position (Valle et al. 1987), however Feng found that leucine was *not* inserted for UAG in their case, which used a different virus, and speculate that the stop codon context could interfere with certain tRNAs as well as the release factor.

Tyrosine tRNA can suppress termination of a UAG stop codon in *Drosophila* but a naturally occurring base modification to this tRNA can prevent this suppression, offering a possible regulatory mechanism (Bienz and Kubli 1981).

S6 Alternative explanations for high conservation of readthrough stop codons

We ruled out two alternative explanations for the high observed conservation of readthrough stop codons. First, we considered whether a particular stop codon might be needed in order for readthrough to occur at a particular stop locus. However, although there is experimental evidence that at certain *Drosophila* loci readthrough is dependent on the choice of stop codon (Robinson and Cooley 1997), we found no reports of cases in which a UAA or UAG can be read through but the weaker UGA can not, so some other explanation is needed for the high conservation of TAA and TAG stop codons. Second, it is unlikely that conservation of readthrough stop codons is due to mRNA secondary structure because the second or third base of the stop codon is predicted to be paired for only about half of the readthrough candidates.

S7 Long readthrough transcripts suggest extensive regulation

As discussed in the main text, readthrough candidates as a group are considerably longer and more complex than non-readthrough genes (Supplementary Figure S5). The length enrichment applies independently to each of the four components that make up the primary transcript -- coding sequence, introns, 5'-UTR, and 3'-UTR (even after subtracting the readthrough region) – in that each of these is larger among the readthrough candidates than non-readthrough transcripts even after controlling simultaneously for all of the other three.

We confirmed that the length enrichments are not due to compositional biases. Indeed, although GC content is high in coding regions of readthrough candidates (Supplementary Figure S6) and other long transcripts, readthrough candidates are still longer than other transcripts with similar composition.

These length enrichments suggest that readthrough genes may be subject to diverse forms of regulation. In the main text we have already discussed how long 3'-UTRs could be involved in the readthrough mechanism. Furthermore, since introns and UTRs often contain transcriptional and splicing regulatory elements, the long introns and UTRs of the readthrough transcripts suggest that readthrough is one more kind of regulation applied to genes that are already highly regulated. The longer 5'- and 3'-UTRs in the readthrough candidates could also contain specific target sites for factors mediating translational regulation, which often bind in UTRs (Wilkie et al. 2003).

Given the strong length biases of readthrough candidates, our previously-reported functional enrichments for neurogenesis proteins (Lin et al. 2007) and several additional functional categories are no longer statistically significant, as neurogenesis proteins are independently associated with longer transcripts (Supplementary Text S8 and Supplementary Figure S7). This suggests neurogenesis proteins could be one of several highly-regulated classes of genes that use readthrough as one of their control mechanisms, but that readthrough is not specific to neurogenesis proteins.

S8 Enrichment for functional roles is confounded by length

To study potential functional roles for readthrough proteins, we studied the enrichment of our candidates for gene ontology (GO) functional terms. We found that readthrough candidates are significantly enriched for 110 GO categories ($p < 0.05$, Bonferroni corrected), including neurogenesis, as we had previously noted (Lin et al. 2007). However, genes in enriched GO categories tend to be longer than average (Supplementary Figure S7), and among genes of similar length, the only enrichments that remain are for transcription factor activity, DNA binding, transcription regulator activity, and regulation of cellular process ($p = 0.0016$, 0.0019, 0.016, and 0.023, respectively, with Bonferonni correction for number of GO categories tested). After also correcting for UTR lengths, no significant GO enrichments remain.

S9 High conservation and unusual base frequencies at positions near the stop codon

In order to obtain clues as to which base positions might play a role in the readthrough mechanism, we analyzed evolutionary conservation across the 12 *Drosophila* species at base positions before the readthrough stop codon. Positions -5, -3, -2, and -1 (where -1 is immediately 5' of the stop codon) are significantly more conserved among readthrough candidates than among non-readthrough transcripts ($p < 10^{-7}$ for positions -3, -2 and -1, $p<0.004$ for position -5). (We restricted the comparison to transcripts with perfectly conserved stop codons to remove biases arising from the high conservation of the stop codon itself.) A similar analysis of the positions after the stop codon would not be meaningful because this region is coding in the readthrough candidates but not in the background transcripts.

Certain bases are known to enhance readthrough in eukaryotes when they are adjacent to the stop codon, namely C in position +1, as already noted, and A in position -1 (Cassan and Rousset 2001). The readthrough candidates are enriched for these bases at these positions, as well as showing other enrichments and depletions. Among positions upstream of the stop codon, position -7 shows high G content, -2 low A, and -1 high A, and among downstream positions, +1 shows high C/low A, +2 high A, and +6 high G/low T ($p < 10^{-5}$ for each position). These differences are significant even when comparing to non-readthrough transcripts with similar overall base composition, or when comparing the positions after the stop to typical coding regions.

S10 Enriched motifs in six downstream positions

In addition to looking at compositional enrichment at individual positions, we also looked for enriched motifs in the six positions immediately downstream of the candidate readthrough stop codons, since particular combinations of bases in these positions have a synergistic effect on termination efficiency in yeast, perhaps by interaction with the 18S rRNA (Namy et al. 2001; Williams et al. 2004).

The most enriched motif, VRD-TYD = (A | G | C) (A | G) (A | G | T) T (C | T) (A | G | T), matches these six positions in 32 of the readthrough candidates whereas only 11 would be expected according to the base frequencies among the readthrough candidates at the individual positions, ($p = 3.2 \times 10^{-8}$, FDR = 0.094)

We verified that this enrichment is not simply a result of a combination of normal codon-pair preferences and the strong preference for C and A at the first two positions. In particular, the fraction of readthrough transcripts containing the motif is significantly higher than the number we would expect given the average frequencies of all consecutive codon pairs in *D. melanogaster* coding regions, adjusted by the frequencies of the first two bases after the readthrough stop codons ($p < 10^{-10}$).

Although one might expect a motif promoting readthrough to be depleted in non-readthrough transcripts, we did not find this motif to be depleted relative to average base frequencies at these positions among those transcripts, even among non-readthrough transcripts with the weak TGA stop codon. Furthermore, although highly expressed non-readthrough transcripts *are* significantly depleted for some signals of weak context, namely TGA stop, and C and A in the first two downstream positions, they are *not* significantly depleted for this motif once those depletions at individual positions are accounted for.

Our motif shows similarities to others that have been found experimentally. Four of the 19 sequences known to cause readthrough of more than 1% in yeast (Namy et al. 2001) match VRD-TYD, including the two that cause the most readthrough. Eight of the readthrough transcripts that match VRD-TYD also match the Skuzeski sequence, CAR-YYA, which has been shown to cause leakage when translating viral mRNA (Skuzeski et al. 1991). Interestingly, all eight of these Skuzeski-sequence matches are perfectly conserved at the base level across the 12 flies whereas only 24% of readthrough candidates show such perfect conservation in these first six downstream positions.

When matches to VRD-TYD are excluded, the most significant match in the remaining readthrough candidates is TYS-WYA = T (C | T) (G | C) (A | T) (C | T) A which matches 8 candidates compared to 0.5 expected, but it is not very significant considering the 1.1 million possible motifs ($p = 4.5 \times 10^{-8}$, FDR = 0.34).

We also looked for enriched motifs in these six base positions allowing only the degeneracies R = (A | G), Y = (C | T) and N = (A | G | C | T) because these are the most appropriate for the RNA-RNA interactions that have been proposed (R to pair with U; Y to pair with G), and because of concern that allowing all possible degeneracies was overfitting the data. The most enriched motifs using only these restricted degeneracies are NAR-TYA, CYA-RNN, TTG-NYR, and TNC-RNN, where in searching for the second and later motifs we excluded transcripts that matched the previous motifs. As with VRD-TYD, these enrichments are not due to codon preferences and are not depleted among non-readthrough transcripts, even highly expressed ones. One of our C. elegans readthrough candidates and two of our human readthrough candidates also match CYA-RNN.

Pattern	Matches	Expected	p-val	FDR
NAR-TYA	12	2	6.9E-7	0.081
CYA-RNN	22	7	1.0E-6	0.12
TTG-NYA	5	<1	1.8E-6	0.13
TYC-AYA	4	<1	6.4E-6	0.38

S11 Highly enriched 8-mer, CAGCAGCA, offers possible cofactor binding site

In order to find potential binding sites of RNA or protein cofactors that might be involved in the readthrough mechanism, we searched within the readthrough regions for all n-mers with $4 \leq n \leq 14$ that were enriched when compared to similar-sized regions *before* the stop codon of non-readthrough transcripts, both with and without adjusting for the base frequency distribution of the readthrough candidates (comparing to regions *after* the stop codon of non-readthrough transcripts would not be meaningful, because they are non-coding).

We found that the most significantly enriched n-mer is the 8-mer CAGCAGCA (Supplementary Figure S8A), which is found in 22.6% of readthrough regions (64) but only 6.8% of similar-length regions before the stop codon of non-readthrough transcripts with similar overall base composition ($p < 4 \times 10^{-17}$, with no multiple hypothesis correction). Without adjusting for base composition the non-readthrough frequency is 4.2%, and in regions after the stop codon it is 1.4%. This enrichment is not merely a consequence of a trinucleotide enrichment. Indeed, among transcripts with similar frequency of the trinucleotides CAG or GCA, the frequency of CAGCAGCA is only 7.3% and 7.8% respectively

We looked for other n-mers that could play a role in the transcripts that lack CAGCAGCA, but although many similar n-mers are also highly enriched, such as the 9-mers GCAGCAGCA and GCAGCAACA they tend to appear in the same 64 transcripts as CAGCAGCA. In fact, if we exclude the 64 readthrough regions containing CAGCAGCA then no 8-mers are significantly enriched among the remaining 219 readthrough regions once multiple hypothesis correction is applied. The most enriched 8-mer in the remainder is AGCGGCAG which is in 11 regions ($p = 2.6 \times 10^{-6}$; Bonferroni adjusted $p = 0.17$).

We investigated the frame bias of this 8-mer to determine if it acts at the amino acid level. The CAGCAGCAs in the readthrough regions are more like to appear in frame 0 than the other two frames (61% versus 12% and 27%, respectively), coding for QQQ or QQH (depending on the subsequent base), suggesting a possible relationship to polyQ repeats. In fact, many of the readthrough regions have unusually long polyQ repeats (Supplementary Figure S8B). However, the preference for frame 0 is not significantly higher in readthrough regions than in other coding regions, and in each of the three frames the number of CAGCAGCAs in the readthrough regions is significantly higher than in coding regions of similar length with similar base composition. This suggests that the relationship of this motif to the readthrough mechanism or function is at the level of nucleic acids, with, perhaps, a small additional effect at the protein level.

We also investigated localization of the 8-mer to see if that could provide clues about its action. The CAGCAGCAs are distributed more or less uniformly through the readthrough regions, rather than being concentrated at some fixed distance after the stop codon. The strong enrichment terminates at the second stop codon – same-sized regions after the *second* stop codon contain this 8-mer for only 10 of the readthrough candidates (3.5%).The readthrough candidates are also enriched for CAGCAGCA in same-sized regions *before* the *first* stop codon. It appears in 43 of them (15.19%), of which 22 include the 8-mer after the stop as well. The enrichment is concentrated in the last 10% of the coding portion of these transcripts, around 250 nt, with no significant enrichment in the first 90% relative to other transcripts with similar frequencies of the 3-mers CAG, AGC, and GCA (Supplementary Figure S8C). The presence of the 8-mer in the coding region before the stop codon suggests that it is not as a binding site for a cofactor that acts at the moment of readthrough, because by that time the ribosome would have already translated the 8-mer and removed any bound cofactor.

In summary, this 8-mer appears within a few hundred nt on one side or the other of the stop codon of about 30% of readthrough candidates and the enrichment is not due simply to chance or to overall prevalence of bases, tri-nucleotides, or codons.

The CAGCAGCA motif has been identified in human cells as the seed for the hsa-miR-424 microRNA (Landthaler et al. 2008), and its initial 6-mer is part of the strong splicing enhancer GACGAC…CAGCAG, which interacts with SRp30 (Tian and Kole 2001), suggesting that the orthologs of these two cofactors might play a role in regulating Drosophila readthrough.

S12 Details of investigation of evolution by truncation or extension

To determine if most *D. melanogaster* readthrough genes evolved primarily by extension rather than truncation, we analyzed alignments at the first and second stop loci. It was not sufficient to simply classify each readthrough candidate individually, due to bias in our method for discovering readthrough genes, so aggregate analysis was required.

We would expect the alignments to show different aggregate behavior depending on which scenario generally applied when readthrough genes evolved in the *D. melanogaster* line since the divergence of the 12 flies. Under the truncation scenario, we would expect any such "new" readthrough transcripts to have sense codons at the first stop locus in the more distantly related species, but stop codons at the second stop locus in all species. Since the second ORF would be protein coding in all 12 species, we would expect such transcripts to be included among our readthrough candidates. On the other hand, under the extension scenario, there would be stop codons in all 12 species at the first stop locus but not at the second stop locus. Although readthrough of these stop codons would occur in *D. melanogaster*, the second ORF would not be protein-coding in some of the species, so they would typically not be found by our method and would not be among our readthrough *candidates*. That is, under the extension scenario, almost all readthrough *candidates* would be "old" readthroughs.

We found that for more than 90% of readthrough candidates, the readthrough state appears to be ancestral to the divergence of the 12 flies, with ancestral stop codons at both stop loci, as expected under the extension scenario and in contrast to what would be expected under the truncation scenario. In several of the remaining candidates there is no stop codon at the second stop locus in the more distantly related species, and these could be examples of evolution by extension even though such transcripts would not usually be found by our method (Supplementary Figure S9).

To look for readthrough genes that have recently evolved by extension, we computed PhyloCSF scores of second ORFs in each frame using only *D. melanogaster* and it 5 closest relatives (*D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, and *D. ananassae*). We found an excess of 188 second ORFs in frame 0 that show both positive 6-species PhyloCSF scores and negative 12-species score, compared to either of the other frames (673 in frame 0, versus 495 and 475 in frames 1 and 2, respectively, after excluding second ORFs that overlap known coding sequences). This suggests that at least ~188 new genes are readthrough in *D. melanogaster* that were not included among our 283 readthrough candidates because the evolutionary signature was not found in all 12 species. (There could be others among the 339 genes that were eliminated during manual curation for having low but positive PhyloCSF score.) For some of these 188 genes, the region might not be protein-coding in all 12 species, while some others might be protein-coding in all 12 species but have negative PhyloCSF score due to chance. To get a lower bound for the former, i.e., the number that have become readthrough since the 12-fly divergence, we used an approximation for the distribution of PhyloCSF scores for protein coding regions of a given length (Lin et al. 2011) to find the subset of these second ORFs whose 12-species PhyloCSF score was not only negative but also lower than the scores of 95% of all regions of that length that are protein-coding in all 12 flies, resulting in 368, 274, and 271 second ORFs in frames 0, 1, and 2, respectively, that are very unlikely to have been protein coding in the 12-fly ancestor. The excess in frame 0 suggests that at least 95 genes have become readthrough by extension in D. melanogaster since the 12-fly ancestor, and probably many more that have 12-fly PhyloCSF score less than 0 but not low enough to meet the 95% cutoff. These 95 genes are not included among our 283 readthrough candidates and the method does not tell us which of the 368 they are.

In contrast, no readthrough candidates were found that matched the truncation scenario. In particular, every readthrough candidate has a stop codon at the first stop locus in at least one of *D. mojavensis*, *D. virilis*, or *D. grimshawi*, the most distant relatives of *D. melanogaster* among the 12 flies, so there was probably a stop codon at that locus in the common ancestor.

A possible concern is that in curating our candidate list we excluded transcripts at which the first stop locus has a stop codon in only one or a few species, since these second ORFs might no longer be protein coding in *D. melanogaster*, and thus might have eliminated legitimate examples of the truncation scenario. To address this, we reviewed the 17 transcripts whose second ORFs have positive PhyloCSF score that were excluded in this way. Of the eight that have stop codons at the first stop locus in more than one species none are convincing examples of readthrough evolved by truncation. Of the remaining nine, for which only *D. melanogaster* has a stop codon at the first stop locus, we cannot determine whether they are actually readthrough in *D. melanogaster* or simply retain the protein coding signature because not enough downstream substitutions have occurred since the nonsense mutation. In any case, these nonsense mutations occurred since the split between *D. melanogaster* and *D. simulans*/*D. sechellia*, which captures less than 10% of the divergence among the 12 flies (Stark et al. 2007), and there are apparently no examples of evolution of readthrough by truncation during the other 90%.

S13 Considerations regarding Z curve test for abundant readthrough

In this section we discuss some considerations related to our Z curve score test for estimating the number of readthrough transcripts in a species.

The estimate our test provides for the number of readthrough transcripts is conservative in four respects. First, the test assumes that all of the readthrough second ORFs will have positive score in frame 0, whereas in most species the Z curve score is positive for only about 60% of short coding regions because of the conservative threshold we used in setting the Z curve origin, so the actual number of readthrough transcripts is probably about 50% more than our estimate. Second, we excluded second ORFs less than 10 codons long, which account for about 7% of the readthrough candidates in *D. melanogaster*. Third, since we have looked only at annotated transcripts, and some of the species tested, such as *P. humanus*, *A. mellifera*, and *A. gambiae* have considerably fewer annotated transcripts than other related species, there might be many unannotated transcripts in these species, some of which could be readthrough. Finally, we did not consider frame-shifting sequencing errors or recent indel mutations which might have inflated the count of positive scores in frames 1 and 2. The magnitude of the underestimate due to all but the last of these considerations would be proportional to the number of readthrough transcripts estimated, so correcting for these three considerations in species in which there was no frame-0 excess would not change the result: there would still not be much readthrough. On the other hand, it is possible that our test would not report abundant readthrough in a species that does have abundant readthrough if the sequence has a large number of frame-shifting sequencing errors. We must keep in mind, though, that our comparative evidence shows that humans and *C. elegans* do not have very many readthrough transcripts, placing limits on the phylogenic extent of the phenomenon.

We see two possible limitations in our approach. First, we have assumed the same frequency of recent nonsense mutations in *D. melanogaster* and other species, whereas in a species with low effective population size the selective force excluding such mutations would be weaker (assuming they are deleterious) while the mutational forces degrading positive Z curve score of the downstream region would remain the same, so our estimate would be too low. Second, we have assumed that the probability that a sequencing mismatch stop codon would be detected by our homology test is the same for simulated sequencing errors as for real errors annotated as stop codons, an assumption that does not account for possible annotation biases. However, the lack of large frame-0 excesses in humans, which have low effective population size, and in tarsiers, which were sequenced with low read coverage (2.1x), gives some reassurance that recent nonsense mutations and sequencing mismatches, respectively, do not account for a large number of positive scores in frame 0 in other species.

S14 Genome Sources

Species	Common Name	Genome Source	Reference
A. aegypti	Yellow Fever Mosquito	vectorbase.org	(Nene et al. 2007)
A. gambiae	Malaria Mosquito	hgdownload.cse.ucsc.edu	(Holt et al. 2002)
A. mellifera	Honey Bee	genome.ucsc.edu	(Solignac et al. 2007)
A. pisum	Pea Aphid	www.aphidbase.com/aphidbase	(International Aphid Genomics Consortium 2010)
B. floridae	Florida Lancelet	ucsc.edu	(Putnam et al. 2008)
B. mori	Silkworm	www.silkdb.org	(Mita et al. 2004)
C. albicans	Thrush Yeast	www.candidagenome.org	(Jones et al. 2004)
C. elegans	Roundworm	www.wormbase.org	(C. elegans Sequencing Consortium 1998)
C. quinquefasciatus	West Nile Mosquito	vectorbase.org	(Arensburger et al. 2010)
D. melanogaster	Fruit Fly	flybase.org	(Adams et al. 2000)
D. mojavensis	Fruit Fly	ucsc.edu	(Drosophila 12 Genomes Consortium et al. 2007)
D. persimilis	Fruit Fly	flybase	(Drosophila 12 Genomes Consortium et al. 2007)
D. pulex	Water Flea	genome.jgi-psf.org	(Colbourne et al. 2011)
D. virilis	Fruit Fly	ucsc.edu	(Drosophila 12 Genomes Consortium et al. 2007)
G. max	Soybean	phytozome.net	(Schmutz et al. 2010)
H. sapiens	Human	ucsc.edu	(Lander et al. 2001)
I. scapularis	Deer Tick	vectorbase.org	(Catherine Hill, personal communication)
M. musculus	Mouse	ensembl.org	(Mouse Genome Sequencing Consortium et al. 2002)
N. vectensis	Sea Anemone	genome.jgi-psf.org/Nemve1	(Putnam et al. 2007)
N. vitripennis	Jewel Wasp	hymenopteragenome.org	(Werren et al. 2010)
P. barbatus	Red Harvester Ant	hymenopteragenome.org	(Smith et al. 2011)
P. humanus	Head Louse	vectorbase.org	(Kirkness et al. 2010)
S. cerevisiae	Baker's Yeast	yeastgenome.org	(Goffeau et al. 1996)
T. castaneum	Red Flour Beetle	bioinformatics.ksu.edu/pub/BeetleBase/3.0	(Tribolium Genome Sequencing Consortium et al. 2008)
T. syrichta	Tarsier	ensembl.org	(Lindblad-Toh et al. 2011)
X. tropicalis	Western Clawed Frog	ucsc.edu	(Hellsten et al. 2010)

Supplementary References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. 2000. The genome sequence of Drosophila melanogaster. *Science* **287**: 2185-95.

Arensburger P, Megy K, Waterhouse RM, Abrudan J, Amedeo P, Antelo B, Bartholomay L, Bidwell S, Caler E, Camara F, et al. 2010. Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics. *Science* **330**: 86-8.

Beier H, Grimm M. 2001. Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. *Nucleic Acids Res* **29**: 4767-82.

Bertram G, Innes S, Minella O, Richardson J, Stansfield I. 2001. Endless possibilities: translation termination and stop codon recognition. *Microbiol* **147**: 255-69.

Bienz M, Kubli E. 1981. Wild-type tRNA-Tyr-G reads the TMV RNA stop codon but Q base-modified tRNA-Tyr-Q does not. *Nature* **294**: 188-190.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**: 2012-8.

Cassan M, Rousset JP. 2001. UAG readthrough in mammalian cells: Effect of upstream and downstream stop codon contexts reveal different signals. *BMC Mol Biol* **2**: 3.

Chapple CE, Guigo R. 2008. Relaxation of Selective Constraints Causes Independent Selenoprotein Extinction in Insect Genomes. *PLoS ONE* **3**: e2968.

Chittum H, Lane W, Carlson B, Roller P, Lung F, Lee B, Hatfield D. 1998. Rabbit Beta-Globin Is Extended Beyond Its UGA Stop Codon by Multiple Suppressions and Translational Reading Gaps. *Biochem* **37**: 10866-10870.

Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, et al. 2011. The ecoresponsive genome of Daphnia pulex. *Science* **331**: 555-61.

DeSimone SM, White K. 1993. The Drosophila erect wing gene, which is important for both neuronal and muscle development, encodes a protein which is similar to the sea urchin P3A2 DNA binding protein. *Mol Cell Biol* **13**: 3641-9.

Drosophila 12 Genomes Consortium, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**: 203-18.

Efron B, Tibshirani R, Storey J, Tusher V. 2001. Empirical Bayes Analysis of a Microarray Experiment. *J Am Stat Assoc* **96**: 1151-1160.

Fearon K, McClendon V, Bonetti B, Bedwell DM. 1994. Premature Translation Termination Mutations Are Efficiently Suppressed in a Highly Conserved Region of Yeast SteGp, a Member of the ATP-binding Cassette (ABC) Transporter Family. *J Biol Chem* **269**: 17802-8.

Feng YX, Copeland TD, Oroszlan S, Rein A, Levin JG. 1990. Identification of amino acids inserted during suppression of UAA and UGA termination codons at the gag-pol junction of Moloney murine leukemia virus. *Proc Natl Acad Sci U S A* **87**: 8860-3.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al. 1996. Life with 6000 genes. *Science* **274**: 546, 563-7.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al. 2011. The developmental transcriptome of Drosophila melanogaster. *Nature* **471**: 473-9.

Hellsten U, Harland RM, Gilchrist MJ, Hendrix D, Jurka J, Kapitonov V, Ovcharenko I, Putnam NH, Shu S, Taher L, et al. 2010. The genome of the Western clawed frog Xenopus tropicalis. *Science* **328**: 633-6.

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JMC, Wides R, et al. 2002. The genome sequence of the malaria mosquito Anopheles gambiae. *Science* **298**: 129-49.

International Aphid Genomics Consortium. 2010. Genome sequence of the pea aphid Acyrthosiphon pisum. *PLoS biology* **8**: e1000313.

Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, Newport G, Thorstenson YR, Agabian N, Magee PT, et al. 2004. The diploid genome sequence of Candida albicans. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 7329-34.

Kirkness EF, Haas BJ, Sun W, Braig HR, Perotti MA, Clark JM, Lee SH, Robertson HM, Kennedy RC, Elhaik E, et al. 2010. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 12168-73.

Kryukov GV, Castellano S, Novoselov SV, Lobanov AV, Zehtab O, Guigó R, Gladyshev VN. 2003. Characterization of mammalian selenoproteomes. *Science* **300**: 1439-43.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Landthaler M, Gaidatzis D, Rothballer A, Chen PY, Soll SJ, Dinic L, Ojo T, Hafner M, Zavolan M, Tuschl T. 2008. Molecular characterization of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. *RNA* **14**: 2580-96.

Lao NT, Maloney AP, Atkins JF, Kavanagh TA. 2009. Versatile dual reporter gene systems for investigating stop codon readthrough in plants. *PLoS ONE* **4**: e7354.

Lin JP, Aker M, Sitney KC, Mortimer RK. 1986. First position wobble in codon-anticodon pairing: amber suppression by a yeast glutamine tRNA. *Gene* **49**: 383-8.

Lin MF, Carlson JW, Crosby MA, Matthews BB, Yu C, Park S, Wan KH, Schroeder AJ, Gramates LS, St Pierre SE, et al. 2007. Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. *Genome Res* **17**: 1823-36.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of evolutionary constraint in the human genome based on 29 eutherian mammals. *Nature* **477**: doi:10.1038/nature10530.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**.

Martin-Romero FJ, Kryukov GV, Lobanov AV, Carlson BA, Lee BJ, Gladyshev VN, Hatfield DL. 2001. Selenium metabolism in Drosophila: selenoproteins, selenoprotein mRNA expression, fertility, and mortality. *J Biol Chem* **276**: 29798-804.

Mita K, Kasahara M, Sasaki S, Nagayasu Y, Yamada T, Kanamori H, Namiki N, Kitagawa M, Yamashita H, Yasukochi Y, et al. 2004. The genome sequence of silkworm, Bombyx mori. *DNA research : an international journal for rapid publication of reports on genes and genomes* **11**: 27-35.

Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520-62.

Namy O, Hatin I, Rousset JP. 2001. Impact of the six nucleotides downstream of the stop codon on translation termination. *EMBO Rep* **2**: 787-93.

Namy O, Rousset JP. 2010. Specification of Standard Amino Acids by Stop Codons. In *Recoding: Expansion of Decoding Rules Enriches Gene Expression* (ed. JF Atkins and RF Gesteland), pp. 79-100. Springer, New York, NY.

Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi Z, Megy K, Grabherr M, et al. 2007. Genome sequence of Aedes aegypti, a major arbovirus vector. *Science* **316**: 1718-23.

Pure GA, Robinson GW, Naumovski L, Friedberg EC. 1985. Partial suppression of an ochre mutation in Saccharomyces cerevisiae by multicopy plasmids containing a normal yeast tRNAGln gene. *J Mol Biol* **183**: 31-42.

Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**: 1064-71.

Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**: 86-94.

Robinson DN, Cooley L. 1997. Examination of the function of two kelch proteins generated by stop codon suppression. *Development* **124**: 1405-17.

Samson ML, Lisbin MJ, White K. 1995. Two distinct temperature-sensitive alleles at the elav locus of Drosophila are suppressed nonsense mutations of the same tryptophan codon. *Genetics* **141**: 1101-11.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* **463**: 178-83.

Skuzeski JM, Nichols LM, Gesteland RF, Atkins JF. 1991. The Signal for a Leaky UAG Stop Codon in Several Plant Viruses Includes the Two Downstream Codons. *J Mol Biol* **218**: 365-73.

Smith CR, Smith CD, Robertson HM, Helmkampf M, Zimin A, Yandell M, Holt C, Hu H, Abouheif E, Benton R, et al. 2011. Draft genome of the red harvester ant Pogonomyrmex barbatus. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 5667-72.

Solignac M, Zhang L, Mougel F, Li B, Vautrin D, Monnerot M, Cornuet JM, Worley KC, Weinstock GM, Gibbs RA. 2007. The genome of Apis mellifera: dialog between linkage mapping and sequence assembly. *Genome biology* **8**: 403.

Stapleton M, Carlson J, Brokstein P, Yu C, Champe M, George R, Guarin H, Kronmiller B, Pacleb J, Park S, et al. 2002. A Drosophila full-length cDNA resource. *Genome biol* **3**: RESEARCH0080.

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, et al. 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* **450**: 219-232.

Sugiharas H, Andrisani V, Salvaterrae PM. 1990. Drosophila choline acetyltransferase uses a non-AUG initiation codon and full length RNA is inefficiently translated. *J Biol Chem* **265**: 21714-9.

Takahashia K, Maruyamaa M, Tokuzawaa Y, Murakamia M, Odaa Y, Yoshikanea N, Makabeb KW, Ichisakaa T, Yamanakaa S. 2005. Evolutionarily conserved non-AUG translation initiation in NAT1/p97/DAP5 (EIF4G2). *Genomics* **85**: 360-71.

Tian H, Kole R. 2001. Strong RNA splicing enhancers identified by a modified method of cycled selection interact with SR protein. *J Biol Chem* **276**: 33833-9.

Tribolium Genome Sequencing Consortium, Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beeman RW, Gibbs R, Beeman RW, Brown SJ, et al. 2008. The genome of the model beetle and pest Tribolium castaneum. *Nature* **452**: 949-55.

Valle RP, Morch MD, Haenni AL. 1987. Novel amber suppressor tRNAs of mammalian origin. *EMBO J* **6**: 3049-55.

Weiss WA, Friedberg EC. 1986. Normal yeast tRNA(CAGGln) can suppress amber codons and is encoded by an essential gene. *J Mol Biol* **192**: 725-35.

Werren JH, Richards S, Desjardins CA, Niehuis O, Gadau J, Colbourne JK, Nasonia Genome Working Group, Werren JH, Richards S, Desjardins CA, et al. 2010. Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. *Science* **327**: 343-8.

Wilkie GS, Dickson KS, Gray NK. 2003. Regulation of mRNA translation by 5'- and 3'-UTR-binding factors. *Trends Biochem Sci* **28**: 182-188.

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

Wills NM. 2010. Translational Bypassing -- Peptidyl-tRNA Re-pairing at Non-overlapping Sites. In *Recoding: Expansion of Decoding Rules Enriches Gene Expression* (ed. JF Atkins and RF Gesteland), pp. 365-381. Springer, New York, NY.