How antisense transcripts can evolve to encode novel proteins

Bharat Ravi Iyengar[1,†], Anna Grandchamp[1], Erich Bornberg-Bauer[1,2]

[1]Institute for Evolution and Biodiversity, University of Münster,
Hüfferstrasse 1, 48149 Münster, Germany

[2]Department of Protein Evolution, Max Planck Institute for Biology Tübingen,
Max-Planck-Ring 5, 72076 Tübingen, Germany

[†] Corresponding author: b.ravi@uni-muenster.de

Abstract

Protein coding features can emerge *de novo* in non coding transcripts, resulting in emergence of new protein coding genes. Studies across many species show that a large fraction large fraction of evolutionarily novel non-coding RNAs have an antisense overlap with protein coding genes. The open reading frames (ORFs) in these antisense RNAs could also overlap with existing ORFs. In this study, we investigate how the evolution an ORF could be constrained by its overlap with an existing ORF in three different reading frames. Using a combination of mathematical modeling and genome/transcriptome data analysis in two different model organisms, we show that antisense overlap can increase the likelihood of ORF emergence and reduce the likelihood of ORF loss, especially in one of the three reading frames. In addition to rationalising the repeatedly reported prevalence of *de novo* emerged genes in antisense transcripts, our work also provides a generic modeling and an analytical framework that can be used to understand evolution of antisense genes.

Introduction

New protein coding genes often arise from existing protein coding genes. This process frequently involves duplication of an existing gene, and a subsequent divergence of one of the duplicated copies from the ancestral sequence (Long *et al.*, 2003; Rastogi

and Liberles, 2005; Näsvall *et al.*, 2012). Several studies have shown that protein coding genes can also emerge *de novo*, in DNA sequences that did not previously encode a protein (*de novo* gene emergence; Tautz and Domazet-Lošo, 2011; Zhao *et al.*, 2014; Schmitz and Bornberg-Bauer, 2017; Vakirlis *et al.*, 2017; Van Oss and Carvunis, 2019; Vakirlis *et al.*, 2020). A protein coding gene thus emerged does not inherit the DNA sequence features necessary for gene expression (transcription and translation), from an ancestral protein coding gene. It must therefore, acquire them through random mutations.

The most basic requirement for translation is an open reading frame (ORF), which is the region of an RNA that is translated into a protein sequence. Efficient translation often requires additional features such as Kozak consensus sequences (Kozak, 1986; Acevedo *et al.*, 2018; Noderer *et al.*, 2014), an optimal codon usage (Hanson and Coller, 2017), and other context dependent regulatory features present in the 5′ and 3′ untranslated regions of the RNA (Hinnebusch *et al.*, 2016; Mayr, 2017).

Because heritable (germline) mutations are rare in most organisms (less than 1 mutation in 100 million base pairs of DNA per generation; Schrider *et al.*, 2013; Zhu *et al.*, 2014; Jee *et al.*, 2016), it is unlikely for many features to emerge simultaneously. That is, features must evolve sequentially. This in turn means that emergence of a phenotype, such as gene expression, is more likely when some required features already exist, and the missing features emerge via mutations. For example, *de novo* emergence is more likely when an ORF is already present and transcriptional features emerge subsequently, or *vice versa*. In our recent work, we also show that *de novo* emergence is more likely via the trajectory where transcription emerges before the emergence of an ORF (Iyengar and Bornberg-Bauer, 2023). Thus stably synthesized RNAs that are not actively and specifically involved in protein synthesis (such as long non-coding RNAs or lncRNAs) can be good sources of new proteins.

Experimental analyses of the ribosome's footprint on RNAs (ribosome profiling) suggest that some ORFs present in lncRNAs are actively translated (Ruiz-Orera *et al.*, 2014; Ingolia *et al.*, 2014; Patraquim *et al.*, 2022; Blevins *et al.*, 2021; Wacholder *et al.*, 2023). Proteins synthesized from the translation of such ORFs can also be beneficial to the host organism (Patraquim *et al.*, 2022; Wacholder *et al.*, 2023). Many lncRNA genes share their genomic location with other genes, but are transcribed in the opposite direction (antisense overlap; Wu and Sharp, 2013; Jadaliha *et al.*, 2018; Tan-Wong *et al.*, 2019; Canzio *et al.*, 2019; Mattick *et al.*, 2023). A recent study has characterized previously unknown RNAs in different species of yeasts, and has shown that a large proportion of these RNA genes have an antisense overlap with existing genes (Blevins *et al.*, 2021). This study also shows that ORFs contained in these RNAs show signatures of translation. These translated ORFs also include those that have recently emerged in one specific species

of yeast. However, these species specific ORFs are less efficiently translated than the ORFs that are conserved between different species. Overall, this study lends support to a hypothesis that many new proteins arise from antisense RNAs. It is likely that the ORFs encoding such proteins are also antisense to existing genes.

In this study, we analyse the emergence of ORFs in antisense RNAs. We specifically focus on ORFs that have an antisense overlap with the coding region (canonical ORF) of an existing protein coding gene. We refer to these ORFs as antisense ORFs (asORFs). Evolution of asORFs is also interesting because it is constrained by the evolutionary selection pressure on the overlapping protein coding genes (Sabath *et al.*, 2012; Mir and Schober, 2014). A pair of mutually antisense ORFs can overlap with each other in three different reading frames. That is, the codon positions in the two ORFs can either perfectly overlap or be offset by one or two nucleotides. The constraints on the co-evolution of the two ORFs would be different in the different reading frames (Mir and Schober, 2014). Our study aims to explore the constraints that affect the evolution of asORFs. To this end, we employ a mathematical model to calculate the probabilities of asORF emergence and loss, in each of the three reading frames. Using the model, we predict that one of the reading frames has a higher propensity to harbor ORFs. We also predict that the likelihood of ORF emergence in this reading frame is higher, and that of ORF loss is lower, than in the other two reading frames. We support our model's predictions with genome analysis of two different organisms – *Saccharomyces cerevisiae* and *Drosophila melanogaster*. We also find that emergence of asORFs in reading frame 1 can be more likely than emergence of non-antisense (intergenic) ORFs.

Results

We developed a mathematical model to estimate the probabilities of ORF emergence and loss, in DNA regions antisense to existing protein coding ORFs. This model is defined by two kinds of probability. The first is the probability of finding a certain kind of DNA sequence, for example an ORF. This stationary probability depends on the nucleotide composition of the DNA region that can be roughly approximated by GC-content or by the frequencies of short DNA sequences (oligomers). The second kind of probability describes the mutational change of a sequence to a different kind of sequence. For example, gain or loss of an ORF. This transition probability depends on the mutation rate and mutation bias, in addition to nucleotide composition. We estimate these parameters primarily from the data on the yeast, *Saccharomyces cerevisiae* (Table 1; Zhu *et al.*, 2014). Our choice is motivated by the fact that the budding yeast is a convenient model organism for laboratory experimental studies that can be used to validate several of our theoretical predictions. We also performed analogous analyses

using data obtained from *Drosophila melanogaster* (Table S1, Schrider *et al.*, 2013).

We estimated the stationary and transition probabilities of antisense ORFs (asORFs, Equations 1 – 3) using the existing (sense) ORF as a reference. asORFs can overlap with the sense ORFs in three different reading frames (henceforth referred to as just "frames"). In frame 0, the codons in the asORF exactly overlap the codons in the sense ORF. In frames 1 and 2, the codons in the asORF are shifted towards the 5′ end of the sense ORF by one and two nucleotide positions, respectively. Thus in frames 1 and 2, the sequence of an antisense codon is determined by two partially overlapping sense codons (dicodons, Figure 1A). Due to this sequence overlap, the evolution of asORFs would be constrained by the evolutionary selection pressures on the sense ORF. Furthermore, these constraints would be different for asORFs located in the three different frames. We analysed the evolution of asORFs when the sense ORF is under three different levels of purifying selection, defined in our study as follows. The first level describes an absence of purifying selection, where any kind of mutation except a non-sense mutation (gain of stop codon) in the sense ORF is tolerated. The second level describes a weak purifying selection that allows synonymous mutations, as well as mutations where an amino acid is substituted by a chemically similar amino acid (for example, aspartic acid to glutamic acid; Table 4). Finally, the third level describes a strong purifying selection, where only synonymous mutations are tolerated in the sense ORF.

Antisense ORFs are more likely to exist in frame 1

For any stretch of DNA to be an ORF, its sequence should contain $3n$ nucleotides ($n \geq 3$), with a start codon that marks its beginning, and exactly one stop codon that marks its

Substitution	Probability(μ)
A:T→T:A	0.063
A:T→G:C	0.144
A:T→C:G	0.110
G:C→A:T	0.349
G:C→T:A	0.182
G:C→C:G	0.152

Table 1: Mutation bias probabilities for different nucleotide mutations in *Saccharomyces cerevisiae* (Zhu *et al.*, 2014). A:T denotes an A-T base pair in a double stranded DNA. Thus A→G mutation on one DNA strand would cause a T→C mutation on the complementary strand. We describe the other mutations in the same way. For our model, we used the reported mutation rate of 1.7×10^{-10} mutations per nucleotide position per generation, in diploid *Saccharomyces cerevisiae* cells (Zhu *et al.*, 2014). For mutation bias probabilities in *D. melanogaster*, see Table S1.

end. The absence of any stop codon within the DNA sequence is the most important factor in determining the existence of an ORF. That is because the likelihood of a premature stop codon increases exponentially with the ORF's length, whereas the likelihoods of a start codon and a terminal stop codon are independent of the ORF's length (Equations 1 – 3).

Based on these considerations, we determined the probability of finding an asORF. To this end, we first calculated the probability of finding a stop codon in the three antisense frames, given the condition that no stop codon exists within the overlapping sense DNA. A stop codon can exist in frame 0 wherever the three reverse complementary codons exist in the sense ORF. Because these codons are allowed in the sense ORF, the overlap does not affect stop codon's probability in frame 0. A stop codon can exist in frames 1 and 2, overlapping with 192 possible dicodons in the sense ORF. However, given the restriction that these dicodons should not contain a stop codon, the number of possible dicodons that overlap a stop codon in antisense frame 1 reduces to 128. In contrast, stop codons in antisense frame 2 can overlap with all possible 192 dicodons, and their probability is thus unaffected by the overlap (see Supplementary Section 2). The probability of finding a stop codon in frame 1, is equivalent to the probability of finding the allowed dicodons. Codon and dicodon probabilities depend on the nucleotide composition, which can be approximated by the GC-content of the locus (Iyengar and Bornberg-Bauer, 2023). We calculated the probability of a start codon without considering the effect of antisense overlap because this effect would be small in magnitude. Using the start and stop codon probabilities, we estimated the probability of finding an asORF of different lengths in each of the three frames. We did so for four different values of GC-content (30, 40, 50 and 60%). The probabilities of asORFs in frames 0 and 2 are identical for all lengths and GC-content because the overlap does not affect stop codon probability in these frames. This in turn, means that asORFs in these frames are equally probable as intergenic ORFs (igORFs) with identical length and GC-content. This is not the case for frame 1, where we found that asORFs are more likely to be found than in the other two frames and intergenic regions (Figure 1B). The only exceptions are ORFs shorter than 17, 21, 27 and 39 codons present in a DNA region with a GC-content of 30%, 40%, 50% and 60%, respectively. Even for these exceptional cases, the probability of an asORF in frame 1 is no less than 74% of the corresponding ORF probabilities in the other frames. We expect that igORFs can indeed be more numerous than asORFs if intergenic regions are long. Our results merely suggest that given that length and GC-content are identical, the probability of an ORF increases when it has an antisense overlap with an existing ORF in frame 1.

We also calculated the probability of asORFs using actual codon and dicodon frequencies in annotated yeast ORFs. Likewise, we calculated the probability of igORFs us-

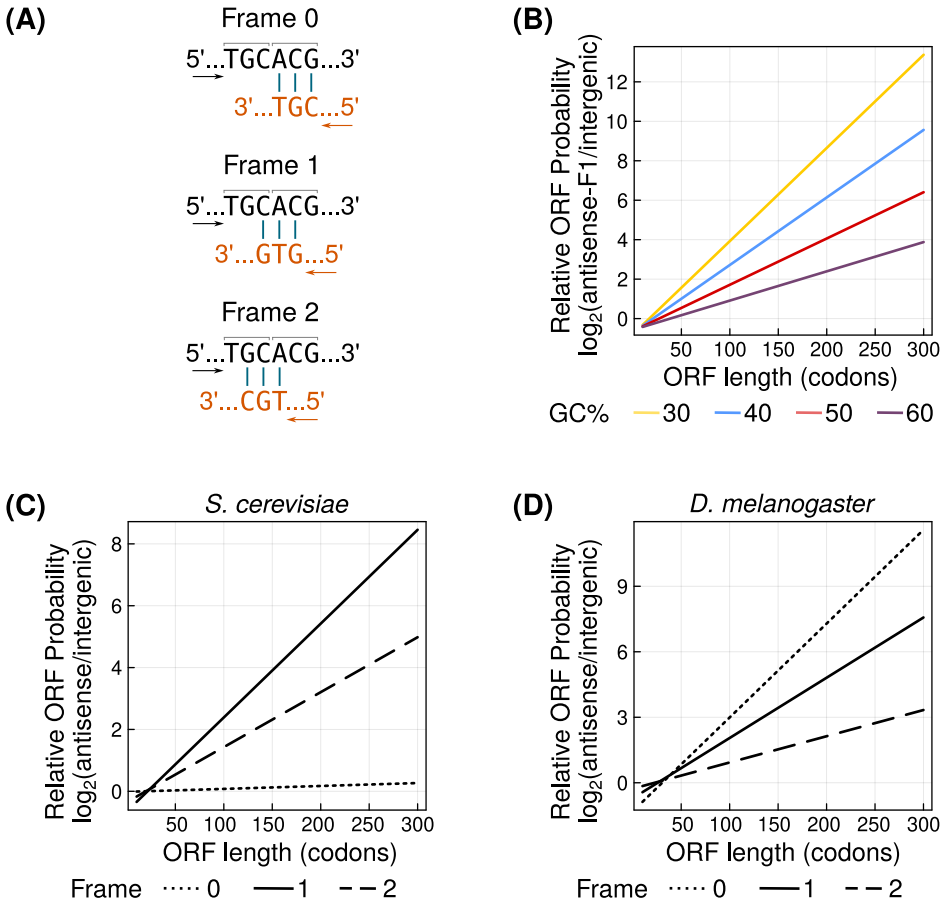

Figure 1: asORFs are more likely to exist than igORFs of identical lengths and composition. **(A)** A hypothetical antisense codon (bottom sequence, orange) can overlap with sense ORF (top sequence, black) in three different frames. Arrows indicate the direction of translation and vertical bars indicate base complementarity. Adjacent codons in the sense ORF are demarcated with horizontal square brackets. **(B)** The probability of asORFs in frame 1 relative to that of igORFs ($\log_2$ ratio, vertical axis), for different values of GC-content of the ORFs (line colors yellow = 30%, blue = 40%, red = 50%, purple = 60%). We do not show asORFs in frames 0 and 2 because their probabilities are identical to that of igORFs. The probability of asORFs relative to that of igORFs ($\log_2$ ratio, vertical axis), calculated using frequencies of short DNA sequences from **(C)** the yeast genome, and **(D)** the fruitfly genome. Frames 0, 1 and 2 are denoted by dotted, solid and dashed lines, respectively. Horizontal axes in panels **(B)** – **(D)** show the length of the ORFs. We only show asORFs that overlap completely with the sense ORF.

ing the frequencies of DNA trimers in yeast intergenic genome. With this analysis, we found that asORFs longer than 17, 21, and 19 codons, in frames 0, 1 and 2, respectively, are more likely to exist than igORFs of the same lengths (Figure 1C).

The probability of finding an ORF doesn't depend on mutation rate bias. Therefore, ORF probability calculations using GC-content (Figure 1B) is organism-independent. However, when we computed the ORF probabilities using the frequencies of codons, dicodons and intergenic trimers from *D. melanogaster*, we found that frame 0 was most likely to harbor long asORFs (>38 codons; Figure 1C). This difference between the predicted ORF probabilities of two organisms results because of differences in codon usage

between the two organisms. Specifically the codons that overlap stop codons (TTA, CTA, TCA) in antisense frame 0 encode serine and leucine. Both these amino acids are encoded by six codons each, and have similar frequencies in the coding regions of both the organisms. However, the usage of the codons – TTA, CTA, TCA, to encode the corresponding amino acids is relatively higher in *S. cerevisiae* than in *D. melanogaster* (Supplementary Section 3; Figure S1).

Antisense ORFs are frequently located in frame 1

Our mathematical model predicts that frame 1 is more likely to harbor asORFs than the other two frames. To verify this prediction, we analysed the genome of the budding yeast, *S. cerevisiae*. We specifically chose this yeast as a model because most of its genes lack introns. This in turn allows us to investigate asORFs whose overlap with the sense ORFs is not interrupted by intronic sequences. Our choice of yeast as a model was further motivated by the availability of data on novel antisense RNAs identified in a recently published study (Blevins *et al.*, 2021). This study further showed that new protein coding genes can emerge *de novo* from these antisense RNAs. We identified all asORFs located in the novel RNAs reported in this study, and calculated the frame in which they overlap with the annotated (sense) ORFs. We also included seven annotated yeast antisense RNAs for the identification of asORFs. Next, we calculated the number of asORFs in each of the three frames, that are at least 30nt long and are wholly con-

	Antisense Frame 0	**Antisense Frame 1**	**Antisense Frame 2**	**Intergenic**
Total loci	7985381	7985381	7985381	798843580
Expected number	592 (612)	657 (690)	632 (612)	49786 (49646)
Observed number	447	646	548	40647
Observed number + subORFs	494	903	623	48598
Expected frequency	7.4×10^{-5} (7.7×10^{-5})	8.2×10^{-5} (8.6×10^{-5})	7.9×10^{-5} (7.7×10^{-5})	6.2×10^{-5} (6.2×10^{-5})
Observed frequency (+ subORFs)	6.2×10^{-5}	1.1×10^{-4}	7.8×10^{-5}	6.1×10^{-5}

Table 2: Expected and observed numbers of antisense and igORFs. Expected numbers and frequencies of ORFs within parantheses were estimated using GC-content of each locus, whereas those outside the parantheses were estimated using DNA oligomer frequencies. For both expected and observed number of asORFs, we only consider ORFs that overlap completely with a sense ORF. Here "sub-ORFs" refers to smaller ORFs ($\geq$30nt) that exist within an ORF such both ORFs share the same stop codon.

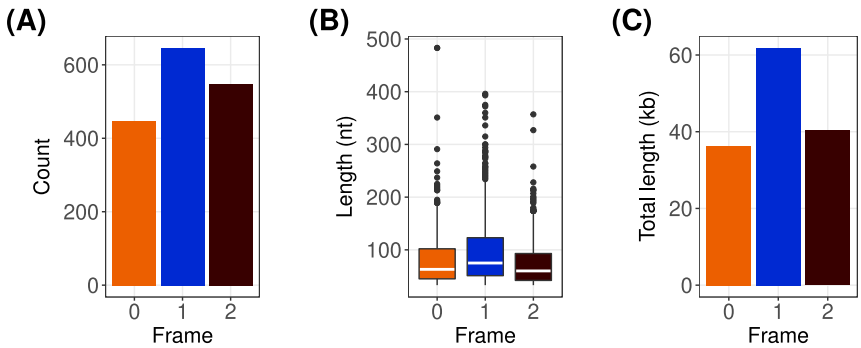

Figure 2: Yeast asORFs preferentially exist in frame 1 than in the other two frames. **(A)** Total number of asORFs (vertical axis). **(B)** asORF length distribution (vertical axis) denoted by a boxplot where the boxes extend from the first to the third quartile and the whiskers have a length equal to 1.5 $\times$ the interquartile range. We indicate the median length using a white horizontal bar. (**C**) Cumulative length of all asORFs (vertical axis). In all the panels, the horizontal axes denote the three different antisense frames. We only show asORFs that overlap 100% with the sense ORF.

tained within the boundaries of a sense ORF. We found that asORFs in frame 1 were significantly more numerous than those in the other two frames (one-tailed Fisher exact test, FDR corrected $P < 4 \times 10^{-5}$). Specifically, ∼39% of all asORFs were located in frame 1, while ∼33% and ∼28% asORFs were located in frames 2 and 0, respectively (Table 2, Figure 2). We also calculated the number of ORFs that have at least 50% of their sequences overlapping in antisense with a sense ORF. This relaxation of overlap percentage did not remarkably increase the number of identified asORFs. To understand if the observed number and proportion of asORFs are in agreement with the model, we calculated the expected number of asORFs in each frame (Equation 6). Specifically, we estimated the total number of expected ORFs that are at least 30nt long and are located in genomic region where antisense RNAs overlap with a known ORF. We found that the actual asORFs in the yeast genome were 1.6 – 24% fewer than expected (Table 2). The ORF identification tool we used (*getorf*; Rice *et al.*, 2000), reports the longest ORF. However, alternate start codons can exist within the ORF sequence wherever a methionine is encoded. Our model does not reject short ORFs (sub-ORFs) within a longer ORFs. When we included the sub-ORFs (≥30nt), the observed asORFs in frame 1 were significantly more numerous than expected (one-tailed Fisher exact test, $P = 5.2 \times 10^{-8}$ with locus specific GC-content, and $P = 2.5 \times 10^{-10}$ with average oligomer frequencies; Table 2). In contrast, observed asORFs in frame 0 were significantly fewer than expected (one-tailed Fisher exact test, $P < 1.7 \times 10^{-3}$). If the observed of ORFs are significantly fewer than expected then negative selection could be an explanation. We note that our calculation of expected number of asORFs (Equation 6) assumes that existence of ORFs in the three different frames is independent of each other. However, presence of an ORF in any one frame can reduce the probability of ORFs in overlapping alternate frames.

Probability of finding an ORF can not only determine the expected number of ORFs, but also the length of the ORFs. Therefore, we next asked if asORFs in frame 1 are generally longer than those in the other two frames. We found that asORFs in frame 1 (median length 75nt) were significantly longer than asORFs in frame 0 and frame 2 (median length 63nt and 60nt, respectively; one tailed Mann-Whitney U test, FDR adjusted $P < 10^{-4}$; Figure 2B). Furthermore, the cumulative length of all the asORFs in frame 1 (62kb) was higher than that of the asORFs in frames 0 and 2 (36kb and 40kb, respectively; Figure 2C).

Next, we analysed if the observed frequency of igORFs is different from that of asORFs. To this end, we calculated the observed number of igORFs including the sub-ORFs, in *S. cerevisiae* genome, using a procedure identical to that we used for identifying asORFs. We then compared the frequencies of igORFs (observed ORFs relative to total loci, Table 2) with that of each type of asORFs, and found that the frequencies of all the three types of asORFs were higher than that of igORFs (one-tailed Fisher exact test, $P < 10^{-8}$). We note again that this result does not indicate that igORFs are less likely to occur than asORFs, as we show that they are indeed more numerous than asORFs (Table 2).

We also performed a similar analysis of *D. melanogaster* genome. Specifically, we used genome and transcriptome data from inbred lines obtained from seven geographically distinct *D. melanogaster* populations (Grandchamp *et al.*, 2023). We used these datasets because they contain several novel RNAs that are not annotated in the reference genome. We found that among the three antisense frames, frame 1 harbored the most number of asORFs. The cumulative length of all the asORFs in the frame 1 was also higher than those in the other two frames (Figure S2). This was true for all the seven lines, and also for the set of unique orthologous sequences between all the lines (orthogroups). However, asORFs in frame 1 were not generally longer than those in the other two frames. Specifically, the median length of asORFs in frame 0, was the highest in all populations but this difference was not statistically significant in all populations (one tailed Mann-Whitney U test, 95% confidence interval). A possible reason for the larger median length of asORF in frame 0 could be the codon usage bias in *D. melanogaster* protein coding genes (Supplementary Section 3). We also analysed if igORFs have a higher frequency than asORFs in *D. melanogaster*. We restricted this analysis to asORFs that completely overlap with a coding exon. We also restricted our analysis to asORFs that do not have introns. That is because introns can change the overlap frame between the flanking exons, and one cannot attribute a specific frame to an asORF. Given these restrictions, we found that asORFs were significantly less frequent than igORFs. We speculate that this difference form *S. cerevisiae* exists because our search space for asORFs is much smaller than that of igORFs. This in turn, can cause many asORFs to not be detected.

ORFs that are more likely to exist may also evolve additional protein coding features. To verify if this is the case, we compared the translational efficiency of *S.cerevisiae* asORFs in different frames using ribosome profiling data (Wacholder *et al.*, 2023). We did not find any significant correlation between frame and translational efficiency of asORFs (Supplementary Section 5). However, igORFs in *S. cerevisiae* had significantly higher translational efficiency than asORFs. One possible reason is that the far more numerous igORFs can have a higher total rate of evolutionary adaptation than asORFs. We did not find any significant difference between the predicted translational efficiency (Kozak consensus sequence strength) for the different asORFs, and igORFs of *D. melanogaster*.

Overall, our genome data analyses from both organisms support our model's prediction that frame 1 offers the most optimal location for asORFs.

Antisense overlap can facilitate ORF emergence and reduce ORF loss

We next analysed how likely it is for asORFs to emerge, when they are not already present. To this end, we calculated gain probability of asORFs in each of the three frames, and under three different intensities of purifying selection. We also calculated the probability of ORF gain in the intergenic regions. We found that asORFs are less likely to emerge in frames 0 and 2, than ORFs in intergenic regions, for all ORF lengths and GC-content. In contrast, long asORF in frame 1 are more likely to emerge than identically sized igORFs (Figure 3A).

Increasing the intensity of purifying selection reduces the emergence likelihood of asORFs in all the three frames. However, long asORFs in frame 1 are still more likely to emerge than identically sized igORFs, even under strong purifying selection. Specifically, the minimum ORF length at which asORFs in frame 1 are more likely to emerge than igORFs, increases with GC-content and the intensity of selection. For example, in the absence of purifying selection, and at a GC-content of 40%, this length is 26 codons. At the same intensity of selection, this length is 46 codons when the GC-content is 60%. Under strong purifying selection and a GC-content of 60%, only the asORFs longer than 108 codons are more likely to emerge than identically sized igORFs (Figure 3A). Our analogous analysis with mutation bias parameters estimated from *D. melanogaster* produced similar results (Figure S4A).

Our analysis of ORF gain probabilities using the frequencies of DNA oligomers (codons, dicodons and intergenic trimers), also shows that asORFs are very likely to emerge in frame 1 (Figure 3B). ORFs longer than 29, 59 and 68 codons are more likely to emerge in antisense frame 1 than in intergenic regions, when the purifying selection is absent, weak and strong, respectively. Interestingly, this analysis revealed that, although

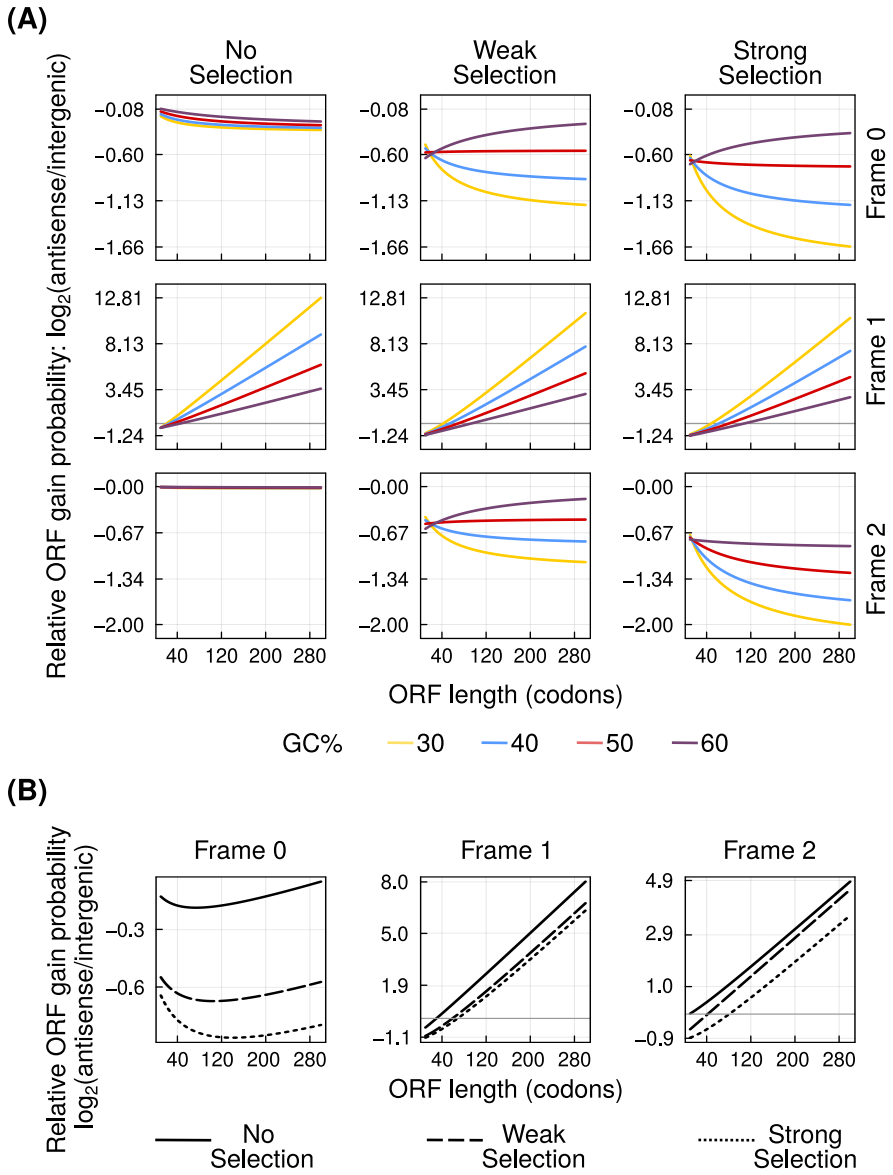

Figure 3: Antisense overlap can facilitate ORF emergence. **(A)** The probability of ORF emergence in the three antisense frames (left to right) relative to that in intergenic regions ($\log_2$ ratio, vertical axis), at different intensities of purifying selection (top to bottom). Line colors indicate the GC-content of the ORFs. **(B)** ORFs gain probability in the three antisense frames relative to that in intergenic regions ($\log_2$ ratio, vertical axis), calculated using frequencies of short DNA sequences from the yeast genome. Solid, dashed and dotted lines denote zero, weak and strong purifying selection, respectively. Horizontal axis in every plot shows the length of the ORFs. In every plot, we only show asORFs that overlap completely with the sense ORF. In plots where the log ratio spans both positive and negative values, we have highlighted the log ratio of zero using a grey horizontal gridline.

asORFs are less likely to emerge in frame 2 than in frame 1, they can emerge more frequently than igORFs. Specifically when the purifying selection is absent, weak and strong, ORFs that are more likely to emerge in antisense frame 2 than in intergenic regions, contain at least 10, 43 and 82 codons, respectively.

However, our analysis of ORF gain probabilities with DNA oligomers estimated from

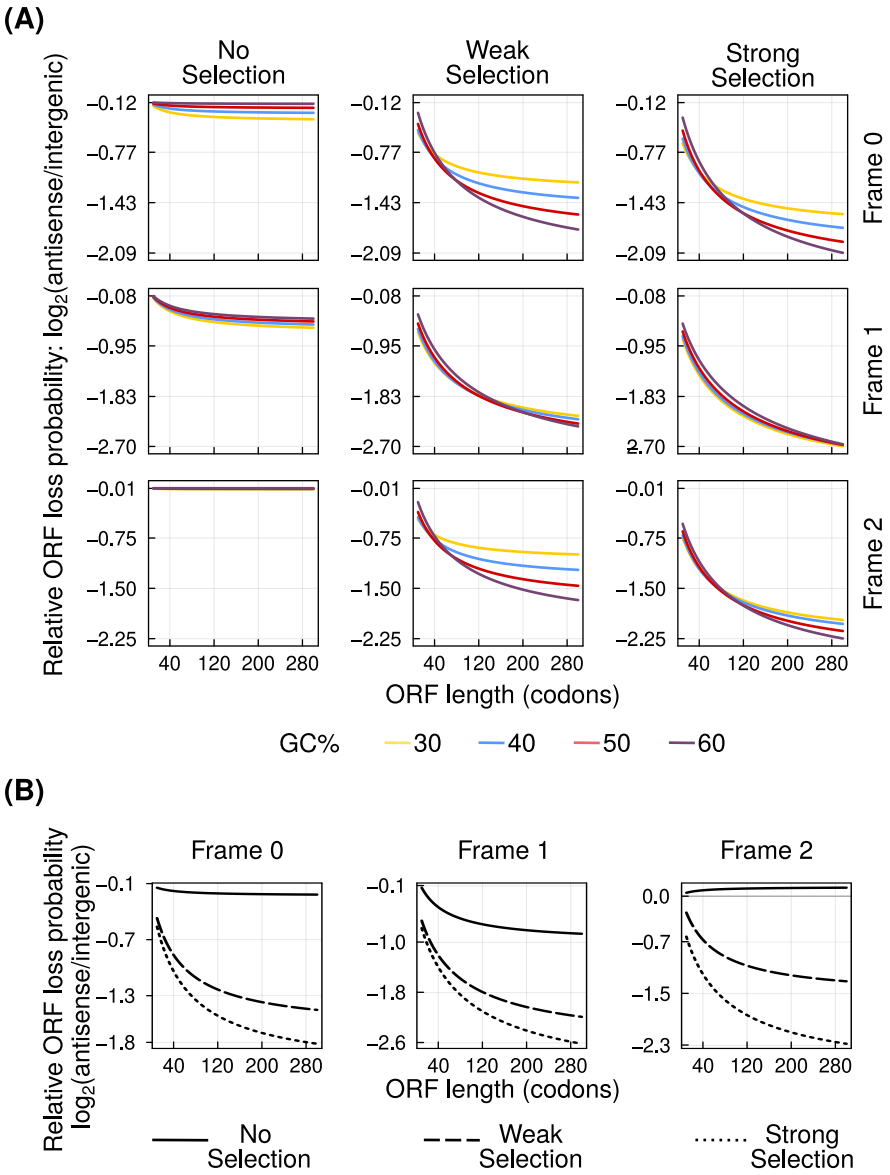

Figure 4: Antisense overlap can reduce ORF loss. **(A)** The probability of ORF loss in the three antisense frames (left to right) relative to that in intergenic regions ($\log_2$ ratio, vertical axis), at different intensities of purifying selection (top to bottom). Line colors indicate the GC-content of the ORFs. **(B)** The ORFs loss probability in the three antisense frames relative to that in intergenic regions ($\log_2$ ratio, vertical axis), calculated using frequencies of short DNA sequences from the yeast genome. Solid, dashed and dotted lines denote zero, weak and strong purifying selection, respectively. Horizontal axis in every plot shows the length of the ORFs. In every plot, we only show asORFs that overlap completely with the sense ORF.

D. melanogaster showed that frame 0 has the highest probability of asORF gain (Figure S4B). This finding is in agreement with the corresponding probabilities of finding the different asORFs (Figure 1C).

Purifying selection reduces the number of tolerated mutations in a DNA locus. We note again that even the lowest intensity of purifying selection according to our definition, disallows nonsense mutations from occurring in the sense ORFs. We thus hypothesized

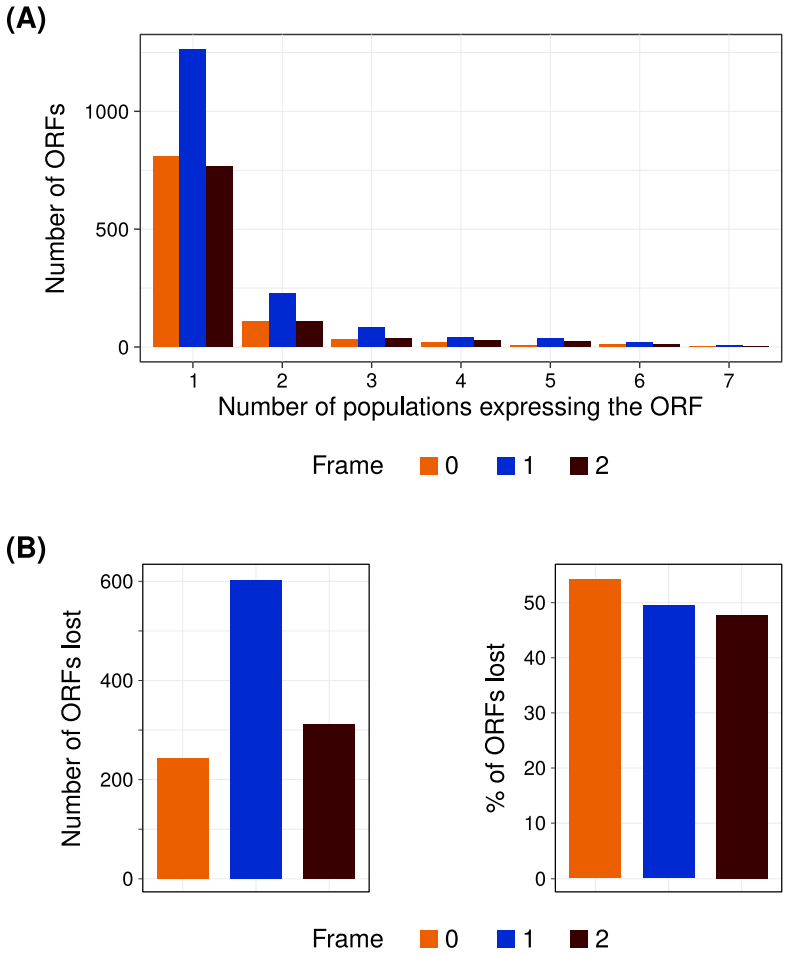

Figure 5: **(A)** Most recently gained asORFs in *D. melanogaster* are frequently located in frame 1. Horizontal axis denotes the number of *D. melanogaster* lines that contain an asORF in their transcriptome (express), and vertical axis denotes the number of such ORFs. **(B)** *D. melanogaster* asORFs in frame 0 have a higher rate of loss. First panel shows total number of lost ORFs (vertical axis) whereas the second panel shows the percentage of total asORFs that are lost. In all figure panels, the three frames are denoted by three different colors (0: orange, 1: blue, and 2: brown).

that overlap with a sense ORF may protect the asORFs from being lost. To this end, we calculated ORF loss probabilities for different ORF lengths, GC-content, and intensities of purifying selection (Figure 4A). In an analogous analysis, we used codon, dicodon, and intergenic trimer frequencies, instead of GC-content, to calculate ORF loss probabilities (Figure 4B). Our analyses show that asORFs are indeed protected from loss due to overlap with existing ORFs, especially when they exist in frame 1. This protection against loss increases with increasing intensity of purifying selection. Our analysis with parameters based on *D. melanogaster* was also in agreement with this result (Figure S5).

To corroborate some of our model's predictions, we analysed the genome and the transcriptome data from the seven different lines of *D. melanogaster*. Six of these lines were obtained from different locations in Europe, whereas one line, the outgroup, was obtained from Zambia (Grandchamp *et al.*, 2023). This data set allowed us to analyse gain

and loss of transcripts and ORFs in short evolutionary timescales (Supplementary section 6.2, Figure S6). If an asORF is found in at least one line, it is gained once in *D. melanogaster*. More specifically, the most recently emerged asORF would be detected in only one line, given the assumption that it is not independently lost in six other lines. We found that regardless of whether an asORF is present in one or many lines, they are more abundant in frame 1 than in the other two frames (Figure 5A). This corroborates our model's prediction (especially GC-content based calculation) that antisense overlap in frame 1 facilitates ORF gain (Figure 3A, Figure S4A).

Next, we analysed the rate of ORF loss in the *D. melanogaster* lines . The genetic variance (F_{ST}) between the European populations of *D. melanogaster* is low (Kapun *et al.*, 2020), suggesting that they are not significantly isolated (Whitlock and McCauley, 1999). As a consequence, we could not establish a clear phylogeny for them. Thus we used a very stringent identification of ORF loss. Specifically, if an ORF is present in the outgroup line (Zambian) and at least one European line, we assume that it was lost in the rest of the European lines. For this definition, we assumed that it is unlikely for an ORF to be gained multiple times independently, and that an ORF can be shared between a European line and the outgroup only if it was already present in their common ancestor. To understand the rate of ORF loss, we normalized the number of asORFs lost in any one frame with total number of asORFs present in the same frame. We found that the rate of ORF loss was highest in frame 0, followed by frames 1 and 2 respectively (Figure 5B). However, the magnitude of this difference was small (<5%) as qualitatively predicted by our model (Figure 4, Figure S5).

Although antisense overlap can protect ORFs from being lost, it can also constrain the evolution of their sequence. Furthermore, effect of mutations in the sense ORF can also affect different asORFs in the three frames differently. We found that when a sense ORF is under purifying selection (weak or strong), mutational effects are the strongest for asORFs located in frame 2, and the weakest for those in frame 0 (Figure S7).

Overall, our analyses suggest that antisense overlap with an existing ORF facilitates emergence of new ORFs, and protects the existing asORFs from being lost.

Discussion

To express a protein, a DNA sequence needs to be transcribed as well as translated. New protein coding genes can emerge *de novo* in non-genic sequences when both these requirements are met. Genomic regions that are already transcribed are thus more likely to evolve protein coding features (Iyengar and Bornberg-Bauer, 2023). Non-coding

RNAs indeed harbor ORFs, and some of these ORFs are also actively transcribed, albeit less efficiently than canonical ORFs present in mRNAs (Ruiz-Orera *et al.*, 2014; Ingolia *et al.*, 2014; Patraquim *et al.*, 2022; Wacholder *et al.*, 2023). Several long non-coding RNA genes overlap with other genes in an antisense orientation (Mattick *et al.*, 2023). This overlap can cause the evolution of asORFs to be constrained by the evolutionary pressures on the corresponding sense genes. The effect of ORF overlap is particularly important in viruses where novel genes frequently emerge overlapping with existing genes, in order to keep the genome compact (Sabath *et al.*, 2012). In this study, we investigate how likely it is for asORFs to exist in the three possible antisense frames, and how their evolution is constrained by the purifying selection on the sense ORFs. To answer these questions, we developed a mathematical model based on mutation probabilities, and analysed the genome sequence for validating some of the model's predictions.

Using the model, we show that asORF are most likely to be found in frame 1 than in the other two frames. This prediction is supported by our analysis of asORFs in *Saccharomyces cerevisiae* and *Drosophila melanogaster* genomes. Furthermore, asORFs in frame 1 are not only more likely to emerge, but may be also less likely to be lost than asORFs those in the other two frames. More interestingly, ORFs are generally more likely to emerge and to be found in antisense frame 1 than in intergenic regions. Conversely, these asORFs are less likely to be lost than igORFs, due to random mutations. This happens because presence of a sense ORF reduces the chances of premature stop codons to occur in the antisense frame 1.

A previous study has also investigated the effect of selection pressure on different frames, using information theory (Mir and Schober, 2014). Although this study also investigates antisense frames, its analytical approach is different from that of our model. Specifically, we calculate the probability of different kinds of mutations, and focus on the presence or absence of ORFs of different lengths, instead of measuring the fidelity of evolutionary information transfer based on relative rates of synonymous and nonsynonymous mutations. Despite these differences in the analytical approach, the findings of our study are in agreement with the previous study. That is, selection pressure on sense ORF (frame +1 in Mir and Schober, 2014) causes preservation of asORFs in frame 1 (frame −2 in Mir and Schober, 2014).

By limiting the number of tolerated mutations, an overlap with an existing ORF can affect the evolution of the protein sequence encoded in an asORF. We quantified mutational effects by estimating the average chemical difference between an original amino acid and a substituted amino acid that results due to random mutations. We found that mutational effects were the strongest in the asORFs in frame 2 (Figure S6). This means that the mutations tolerated in the sense ORFs under purifying selection produce ex-

treme non-synonymous changes in the asORFs in frame 2.

Like all computational models, our model is based on some assumptions and simplifications, that need to be considered. For example, we use GC-content as a measure of nucleotide composition which we use in turn to calculate different probability values. For these calculations, we also use codon, dicodon and DNA trimer frequencies, which are data based measures of nucleotide composition. Our results show that probability values calculated using GC-content can sometimes noticeably differ from the values calculated using DNA oligomer distributions, especially for *D. melanogaster*. For example, our estimated probability of finding a *D. melanogaster* asORF was highest in frame 1 when we used GC-content, whereas it was highest in frame 0 when we used oligomer distributions. Both our measures of nucleotide composition can vary significantly across the genome (with oligomer frequencies showing more variation; Supplementary Section 8, Figure S8). We used different values of GC-content for our calculations that can represent different genomic loci. In contrast, our DNA oligomer based calculations is based on the average frequency of oligomers from the whole genome. Thus they may not accurately represent any one specific locus. However, our computational framework can be adapted to analyse specific loci. Therefore, model predictions may not be 100% accurate. However, despite the possible inaccuracies, our models are able to produce results that qualitatively agree with real data. Our analyses of asORFs from *S. cerevisiae* and *D. melanogaster* support our model based finding that antisense frame 1 has higher likelihood to harbor asORFs. Our models are based on the assumptions of uniform mutation rate and independence of mutational events. These assumptions are not exactly accurate because mutation rates can vary across the genome (Monroe *et al.*, 2022), and multiple nucleotides can be mutated in a single mutational event (Harris and Nielsen, 2014). Furthermore, mutation rate bias can be different in different organisms (Cano *et al.*, 2022; Bergeron *et al.*, 2023, also compare Table 1 and Table S1). Our results show that despite the differences in the mutation rate and mutation rate bias, between yeast and *D. melanogaster*, the results qualitatively remain the same. Thus our predictions are robust to small changes in parameters.

We believe our work opens up interesting questions, and avenues for future research. For example, the cellular functions and biochemical properties of proteins encoded by asORFs would be worth investigating. This may be especially relevant for antisense lncRNAs, some of which are involved in regulation of gene expression. asORFs may possibly provide another dimension to the cellular function of these RNAs. Translation of ORFs in lncRNAs can indeed be spatiotemporally regulated (Patraquim *et al.*, 2022). asORFs may especially be relevant in organisms with compact genomes, such as viruses. Existing work indeed shows that new protein coding genes emerge in viruses, overlapping with existing genes (Sabath *et al.*, 2012; Schlub and Holmes, 2020; Romerio,

2023). This overlap couples the evolution of the two overlapping genes. Eventually, understanding viral evolution may help design better therapeutic strategies against viral diseases.

Materials and Methods

Probabilities of finding, gaining, and losing an ORF

We calculated the probabilities of finding, gaining and losing a ORF, using nucleotide composition, mutation rate and mutation rate bias, as described in our previous study (Iyengar and Bornberg-Bauer, 2023). Briefly, a reading frame is an ORF (P_{ORF}) when a start codon exists at its beginning (P_{ATG}), a stop codon exists at its end (P_{stop}), and no stop codon exists in the middle ($1 - P_{stop}$). An ORF emerges ($P_{ORF\text{-}gain}$) when two of the three required features are present and are not lost due to mutations, while the missing feature emerges due to mutations. Conversely, an ORF is lost ($P_{ORF\text{-}loss}$) when any one of the three required features is lost. The probabilities of finding, gaining and losing an ORF containing k codons, are described by the following equations (Equations 1 – 3). Table 3 describes the terms used in these equations.

$$P_{ORF}(k) = P_{ATG} \times P_{stop} \times (1 - P_{stop})^{k-2} \tag{1}$$

$$\begin{aligned} P_{ORF\text{-}gain}(k) = \quad & P_{ATG\text{-}gain} \times P_{stop\text{-}stay} \times (1 - P_{stop} - P_{stop\text{-}gain})^{k-2} \\ & + P_{ATG\text{-}stay} \times P_{stop\text{-}gain} \times (1 - P_{stop} - P_{stop\text{-}gain})^{k-2} \\ & + P_{ATG\text{-}stay} \times P_{stop\text{-}stay} \times P_{stop\text{-}loss} \times (k-2) \times (1 - P_{stop} - P_{stop\text{-}gain})^{k-3} \end{aligned} \tag{2}$$

$$P_{ORF\text{-}loss}(k) = P_{ATG\text{-}loss} + P_{stop\text{-}loss} + (k-2) \times \frac{P_{stop\text{-}gain}}{1 - P_{stop}} \tag{3}$$

Term	Description
P_{stop}	Probability of finding a stop codon
$P_{stop\text{-}gain}$	Probability of gaining a stop codon
$P_{stop\text{-}loss}$	Probability of losing a stop codon given that it already exists
$P_{stop\text{-}stay}$	Probability that a stop codon exists and is not lost due to mutations

Table 3: Description of the probability terms used in Equations 1 – 3. Here we describe the probabilities associated with stop codons. Analogous probability terms for a start codon are denoted by the subscript, *ATG* (instead of *stop*). For asORFs, P_{stop}, $P_{stop\text{-}gain}$, $P_{stop\text{-}loss}$ and $P_{stop\text{-}stay}$ will vary depending on the frame.

Modeling weak purifying selection

Both gain and loss probabilities of asORFs depend on the strength of selection on the sense ORF. That is, selection would limit the number of sense codons or dicodons that any of the existing codons and dicodons can mutate to. Under strong purifying selection only synonymous mutations are allowed, whereas weak purifying selection allows an amino acid to be substituted by a chemically similar amino acid. To determine chemically similar amino acids, we used an amino acid similarity matrix based on binding covariance of different short peptides to MHC (Major Histocompatibility Complex, Kim *et al.*, 2009). As noted by Kim *et al.* (2009), we identified chemically similar amino acids from pairs of amino acids whose covariance scores are more than 0.05 (Table 4).

Amino acid	Chemically similar amino acids
A	P, T, V
C	-
D	E
E	D
F	I, W, Y
G	-
H	K, R
I	F, L, M, V
K	H, R
L	I, M

Amino acid	Chemically similar amino acids
M	I, L
N	-
P	A
Q	-
R	H, K
S	T
T	A, S
V	A, I
W	F, Y
Y	F, W

Table 4: Chemically similar amino acids identified using the data from Kim *et al.* (2009)

Estimating trimer, codon, and dicodon frequencies

We used a sliding window of size 1nt, to calculate the frequency of all trimers in the annotated intergenic regions of *S. cerevisiae* (Engel *et al.*, 2014). We calculated codon and dicodon frequencies from a non-redundant list of annotated protein coding ORF sequences (CDS; Engel *et al.*, 2014).

We applied the same method for estimating DNA oligomer frequencies in *D. melanogaster*. To this end, we used the *D. melanogaster* reference genome (release 6.4.9; Gramates *et al.*, 2022).

Identification of asORFs in the genome

To identify asORFs in *Saccharomyces cerevisiae* genome, we first compiled a list of known antisense RNAs from the S288C reference genome (Engel *et al.*, 2014), and combined it with the list of novel RNAs identified in a recent study (Blevins *et al.*, 2021). Next, we identified all ORFs in the combined set of RNAs using the program *getorf* (Rice *et al.*, 2000). Specifically, we identified the longest sequence that starts with the canonical `ATG` start codon and ends with a stop codon. We used a minimum ORF length of 30nt (default value in *getorf*). We then mapped the genomic coordinates of all the identified ORFs, verified if they overlap with a known ORF in the opposite strand, and calculated the frame of antisense overlap. We used *awk* scripts for this analysis. To calculate the number of ORFs expected from the model, we first identified genomic regions where an antisense overlap exists between an annotated ORF and a RNA. For each such region A, with a length l_A, we caculated the number of loci (nLoci) where any asORF containing k codons could exist:

$$nLoci(A,k) = \frac{l_A - 3k + 1}{3} \tag{4}$$

$$nLoci\ (total) = \sum_{A} \sum_{\substack{k \geq 10 \\ 3k < l_A}} nLoci(A,k) \tag{5}$$

Total number of asORFs in any frame (f) would be defined as:

$$N_{asORF}(f) = \sum_{A} \sum_{\substack{k \geq 10 \\ 3k < l_A}} P_{ORF}(f,k)\ nLoci(A,k) \tag{6}$$

Where $P_{ORF}(f,k)$ is the probability of finding an ORF in a frame f (Figure 1).

We also identified igORFs from annoted *S. cerevisiae* intergenic regions (I; Engel *et al.*, 2014) using *getorf* (Rice *et al.*, 2000). We calculated the number of intergenic loci where an igORF could exist, and the total number of predicted igORFs as described by the following equations:

$$nLoci(I,k) = l_I - 3k + 1 \tag{7}$$

$$N_{igORF} = \sum_{A} \sum_{\substack{k \geq 10 \\ 3k < l_I}} P_{ORF}(k)\ nLoci(I,k) \tag{8}$$

We performed an analogous analysis for *D. melanogaster*. For details please see Supplementary Section 4.

Data availability

All scripts and necessary data files are freely available on GitHub: *BharatRaviIyengar/DeNovoEvolution*.

We implemented our model using Julia programming language using the following scripts:

- *antisenseGenes.jl* (main script)
- *antisenseGenes_supplement.jl* (calculations using codon, dicodon, and intergenic trimer frequencies)
- *nucleotidefuncts.jl* (dependency for basic functions)

The *awk* scripts for asORF identification from yeast and *D. melanogaster* genome are located in the folder *DataAnalysis*. A wrapper *bash* script implements the complete analysis pipeline in both cases. We also include some original data files for yeast but not for *D. melanogaster*.

Source data for the figures are provided with this paper.

Acknowledgments

References

Acevedo, J. M., Hoermann, B., Schlimbach, T., and Teleman, A. A. 2018. Changes in global translation elongation or initiation rates shape the proteome via the Kozak sequence. *Scientific Reports*, 8(1): 4018.

Bergeron, L. A., Besenbacher, S., Zheng, J., and others 2023. Evolution of the germline mutation rate across vertebrates. *Nature*, 615(7951): 285–291.

Blevins, W. R., Ruiz-Orera, J., Messeguer, X., and others 2021. Uncovering de novo gene birth in yeast using deep transcriptomics. *Nature Communications*, 12(1): 604.

Cano, A. V., Rozhoňová, H., Stoltzfus, A., McCandlish, D. M., and Payne, J. L. 2022. Mutation bias shapes the spectrum of adaptive substitutions. *Proceedings of the National Academy of Sciences*, 119(7): e2119720119.

Canzio, D., Nwakeze, C. L., Horta, A., and others 2019. Antisense lncRNA Transcription Mediates DNA Demethylation to Drive Stochastic Protocadherin α Promoter Choice. *Cell*, 177(3): 639–653.e15.

Engel, S. R., Dietrich, F. S., Fisk, D. G., and others 2014. The reference genome sequence of *Saccharomyces cerevisiae*: Then and now. *G3 Genes—Genomes—Genetics*, 4(3): 389–398.

Gramates, L. S., Agapite, J., Attrill, H., and others 2022. FlyBase: a guided tour of highlighted features. *Genetics*, 220(4).

Grandchamp, A., Kühl, L., Lebherz, M., and others 2023. Population genomics reveals mechanisms and dynamics of de novo expressed open reading frame emergence in drosophila melanogaster. *Genome Research*, 33(6): 872–890.

Hanson, G. and Coller, J. 2017. Codon optimality, bias and usage in translation and mRNA decay. *Nature Reviews Molecular Cell Biology*, 19(1): 20–30.

Harris, K. and Nielsen, R. 2014. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Research*, 24(9): 1445–1454.

Hinnebusch, A. G., Ivanov, I. P., and Sonenberg, N. 2016. Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science*, 352(6292): 1413–1416.

Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., and others 2014. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Reports*, 8(5): 1365–1379.

Iyengar, B. R. and Bornberg-Bauer, E. 2023. Neutral models of de novo gene emergence suggest that gene evolution has a preferred trajectory. *Molecular Biology and Evolution*, 40(4).

Jadaliha, M., Gholamalamdari, O., Tang, W., and others 2018. A natural antisense lncRNA controls breast cancer progression by promoting tumor suppressor gene mRNA stability. *PLOS Genetics*, 14(11): e1007802.

Jee, J., Rasouly, A., Shamovsky, I., and others 2016. Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. *Nature*, 534(7609): 693–696.

Kapun, M., Barrón, M. G., Staubach, F., and others 2020. Genomic analysis of european drosophila melanogaster populations reveals longitudinal structure, continent-wide selection, and previously unknown DNA viruses. *Molecular Biology and Evolution*, 37(9): 2661–2678.

Kim, Y., Sidney, J., Pinilla, C., Sette, A., and Peters, B. 2009. Derivation of an amino acid similarity matrix for peptide:MHC binding and its application as a bayesian prior. *BMC Bioinformatics*, 10(1).

Kozak, M. 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell*, 44(2): 283–292.

Long, M., Betrán, E., Thornton, K., and Wang, W. 2003. The origin of new genes: glimpses from the young and old. *Nature Reviews Genetics*, 4(11): 865–875.

Mattick, J. S., Amaral, P. P., Carninci, P., and others 2023. Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nature Reviews Molecular Cell Biology*.

Mayr, C. 2017. Regulation by 3′-untranslated regions. *Annual Review of Genetics*, 51(1): 171–194.

Mir, K. and Schober, S. 2014. Selection pressure in alternative reading frames. *PLoS ONE*, 9(10): e108768.

Monroe, J. G., Srikant, T., Carbonell-Bejerano, P., and others 2022. Mutation bias reflects natural selection in Arabidopsis thaliana. *Nature*, 602(7895): 101–105.

Näsvall, J., Sun, L., Roth, J. R., and Andersson, D. I. 2012. Real-time evolution of new genes by innovation, amplification, and divergence. *Science*, 338(6105): 384–387.

Noderer, W. L., Flockhart, R. J., Bhaduri, A., and others 2014. Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Molecular Systems Biology*, 10(8): 748.

Patraquim, P., Magny, E. G., Pueyo, J. I., Platero, A. I., and Couso, J. P. 2022. Translation and natural selection of micropeptides from long non-canonical RNAs. *Nature Communications*, 13(1).

Rastogi, S. and Liberles, D. A. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evolutionary Biology*, 5(1).

Rice, P., Longden, I., and Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(6): 276–277.

Romerio, F. 2023. Origin and functional role of antisense transcription in endogenous and exogenous retroviruses. *Retrovirology*, 20(1).

Ruiz-Orera, J., Messeguer, X., Subirana, J. A., and Alba, M. M. 2014. Long non-coding RNAs as a source of new peptides. *eLife*, 3.

Sabath, N., Wagner, A., and Karlin, D. 2012. Evolution of viral proteins originated de novo by overprinting. *Molecular Biology and Evolution*, 29(12): 3767–3780.

Schlub, T. E. and Holmes, E. C. 2020. Properties and abundance of overlapping genes in viruses. *Virus Evolution*, 6(1).

Schmitz, J. and Bornberg-Bauer, E. 2017. Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA . *F1000Research*, 6(57).

Schrider, D. R., Houle, D., Lynch, M., and Hahn, M. W. 2013. Rates and Genomic Consequences of Spontaneous Mutational Events in Drosophila melanogaster. *Genetics*, 194(4): 937–954.

Tan-Wong, S. M., Dhir, S., and Proudfoot, N. J. 2019. R-loops promote antisense transcription across the mammalian genome. *Molecular Cell*, 76(4): 600–616.e6.

Tautz, D. and Domazet-Lošo, T. 2011. The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12(10): 692–702.

Vakirlis, N., Hebert, A. S., Opulente, D. A., and others 2017. A molecular portrait of de novo genes in yeasts. *Molecular Biology and Evolution*, 35(3): 631–645.

Vakirlis, N., Acar, O., Hsu, B., and others 2020. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nature Communications*, 11(1).

Van Oss, S. B. and Carvunis, A.-R. 2019. De novo gene birth. *PLOS Genetics*, 15(5): 1–23.

Wacholder, A., Parikh, S. B., Coelho, N. C., and others 2023. A vast evolutionarily transient translatome contributes to phenotype and fitness. *bioRxiv*.

Whitlock, M. C. and McCauley, D. E. 1999. Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm + 1)$. *Heredity*, 82(2): 117–125.

Wu, X. and Sharp, P. A. 2013. Divergent transcription: A driving force for new gene origination? *Cell*, 155(5): 990–996.

Zhao, L., Saelao, P., Jones, C. D., and Begun, D. J. 2014. Origin and spread of *de Novo* genes in *Drosophila melanogaster* populations. *Science*, 343(6172): 769–772.

Zhu, Y. O., Siegal, M. L., Hall, D. W., and Petrov, D. A. 2014. Precise estimates of mutation rate and spectrum in yeast. *Proceedings of the National Academy of Sciences*, 111(22).

How antisense transcripts can evolve to encode novel proteins
Supplementary Material

Bharat Ravi Iyengar[1,†], Anna Grandchamp[1], Erich Bornberg-Bauer[1,2]

[1]Institute for Evolution and Biodiversity, University of Münster,
Hüfferstrasse 1, 48149 Münster, Germany

[2]Department of Protein Evolution, Max Planck Institute for Biology Tübingen, Max-Planck-Ring 5,
72076 Tübingen, Germany

[†] Corresponding author: b.ravi@uni-muenster.de

Contents

1 Mutation rate and mutation rate bias in in *Drosophila melanogaster* **2**

2 Probability of asORFs in frames 0 and 2 is identical to that of igORFs of same length and GC-content **2**

3 Why asORFs appear to be most probable in frame 0, in *D. melanogaster* but not in *S. cerevisiae* **3**

4 Distribution of antisense ORFs in *Drosophila melanogaster* genome **4**

5 Translational efficiency of asORFs **6**

6 Gain and loss probabilities of antisense ORFs in *Drosophila melanogaster* **8**
6.1 Model predictions 8
6.2 Analysis of asORF gain and loss using genomics data 9

7 Effect of mutations on asORFs **11**

8 Is GC-content a better parameter for asORF probability calculation than global DNA oligomer frequencies? **13**

1. Mutation rate and mutation rate bias in in *Drosophila melanogaster*

Substitution	Probability(μ)
A:T→T:A	0.056
A:T→G:C	0.243
A:T→C:G	0.074
G:C→A:T	0.483
G:C→T:A	0.075
G:C→C:G	0.069

Table S1: Mutation bias probabilities for different nucleotide mutations based on Schrider *et al.* (2013) and Zhang and Gerstein (2003). A:T denotes an A-T base pair in a double stranded DNA. Thus A→G mutation on one DNA strand would cause a T→C mutation on the complementary strand. We describe the other mutations in the same way. We used an average mutation rate of 7.8×10^{-9} mutations per nucleotide position per generation (Schrider *et al.*, 2013)

2. Probability of asORFs in frames 0 and 2 is identical to that of igORFs of same length and GC-content

The probability of finding an antisense stop codon in frame 0 is same as the probability of finding the three reverse complementary codons in the sense ORF (TTA, CTA and TCA). These three codons are allowed in the sense ORFs, and their probability would be simply determined by the GC-content of the sense ORF. These three codons have the same GC composition as the stop codons, and therefore, their probability is identical to that of stop codons (given identical GC-content of the locus). Therefore, given these considerations, the probability of a frame-0 antisense ORF (asORF) is identical to that of an intergenic ORF (igORF) of same length and GC-content.

Next, we explain why the probability of frame-2 asORFs is identical to that of igORFs of similar nucleotide composition and length. The probability of finding a frame-2 antisense stop codon is determined by the corresponding dicodons in the sense ORF. There are 64 possible overlapping dicodons for both frame 1 and frame 2 antisense codons ($4^3 = 64$; three out of six positions in a dicodon are determined by the overlapping antisense codon). Thus, there are $64 \times 3 = 192$ dicodons that overlap with any of the three antisense stop codons. By definition, the sense ORF should not contain a stop codon which means that no dicodon can contain a stop codon. For frame-1 antisense stop codons, 64 overlapping sense overlapping dicodons contain a stop codon (Table S2A), whereas for frame-2 antisense stop codons none of the overlapping dicodons contain a stop codon (Table S2B). Therefore, the probability of an antisense

(A)

TAA		TAG	
AAT TAA	AAT TAG	AAC TAA	AAC TAG
TAT TAA	TAT TAG	TAC TAA	TAC TAG
GAT TAA	GAT TAG	GAC TAA	GAC TAG
CAT TAA	CAT TAG	CAC TAA	CAC TAG
ATT TAA	ATT TAG	ATC TAA	ATC TAG
TTT TAA	TTT TAG	TTC TAA	TTC TAG
GTT TAA	GTT TAG	GTC TAA	GTC TAG
CTT TAA	CTT TAG	CTC TAA	CTC TAG
AGT TAA	AGT TAG	AGC TAA	AGC TAG
TGT TAA	TGT TAG	TGC TAA	TGC TAG
GGT TAA	GGT TAG	GGC TAA	GGC TAG
CGT TAA	CGT TAG	CGC TAA	CGC TAG
ACT TAA	ACT TAG	ACC TAA	ACC TAG
TCT TAA	TCT TAG	TCC TAA	TCC TAG
GCT TAA	GCT TAG	GCC TAA	GCC TAG
CCT TAA	CCT TAG	CCC TAA	CCC TAG

(B)

TAA				TAG				TGA			
ATT AAA	ATT AAT	ATT AAG	ATT AAC	ACT AAA	ACT AAT	ACT AAG	ACT AAC	ATC AAA	ATC AAT	ATC AAG	ATC AAC
TTT AAA	TTT AAT	TTT AAG	TTT AAC	TCT AAA	TCT AAT	TCT AAG	TCT AAC	TTC AAA	TTC AAT	TTC AAG	TTC AAC
GTT AAA	GTT AAT	GTT AAG	GTT AAC	GCT AAA	GCT AAT	GCT AAG	GCT AAC	GTC AAA	GTC AAT	GTC AAG	GTC AAC
CTT AAA	CTT AAT	CTT AAG	CTT AAC	CCT AAA	CCT AAT	CCT AAG	CCT AAC	CTC AAA	CTC AAT	CTC AAG	CTC AAC
ATT ATA	ATT ATT	ATT ATG	ATT ATC	ACT ATA	ACT ATT	ACT ATG	ACT ATC	ATC ATA	ATC ATT	ATC ATG	ATC ATC
TTT ATA	TTT ATT	TTT ATG	TTT ATC	TCT ATA	TCT ATT	TCT ATG	TCT ATC	TTC ATA	TTC ATT	TTC ATG	TTC ATC
GTT ATA	GTT ATT	GTT ATG	GTT ATC	GCT ATA	GCT ATT	GCT ATG	GCT ATC	GTC ATA	GTC ATT	GTC ATG	GTC ATC
CTT ATA	CTT ATT	CTT ATG	CTT ATC	CCT ATA	CCT ATT	CCT ATG	CCT ATC	CTC ATA	CTC ATT	CTC ATG	CTC ATC
ATT AGA	ATT AGT	ATT AGG	ATT AGC	ACT AGA	ACT AGT	ACT AGG	ACT AGC	ATC AGA	ATC AGT	ATC AGG	ATC AGC
TTT AGA	TTT AGT	TTT AGG	TTT AGC	TCT AGA	TCT AGT	TCT AGG	TCT AGC	TTC AGA	TTC AGT	TTC AGG	TTC AGC
GTT AGA	GTT AGT	GTT AGG	GTT AGC	GCT AGA	GCT AGT	GCT AGG	GCT AGC	GTC AGA	GTC AGT	GTC AGG	GTC AGC
CTT AGA	CTT AGT	CTT AGG	CTT AGC	CCT AGA	CCT AGT	CCT AGG	CCT AGC	CTC AGA	CTC AGT	CTC AGG	CTC AGC
ATT ACA	ATT ACT	ATT ACG	ATT ACC	ACT ACA	ACT ACT	ACT ACG	ACT ACC	ATC ACA	ATC ACT	ATC ACG	ATC ACC
TTT ACA	TTT ACT	TTT ACG	TTT ACC	TCT ACA	TCT ACT	TCT ACG	TCT ACC	TTC ACA	TTC ACT	TTC ACG	TTC ACC
GTT ACA	GTT ACT	GTT ACG	GTT ACC	GCT ACA	GCT ACT	GCT ACG	GCT ACC	GTC ACA	GTC ACT	GTC ACG	GTC ACC
CTT ACA	CTT ACT	CTT ACG	CTT ACC	CCT ACA	CCT ACT	CCT ACG	CCT ACC	CTC ACA	CTC ACT	CTC ACG	CTC ACC

Table S2: (A) The 64 sense dicodons that contain a stop codon, and that overlap with an antisense stop codon in frame-1. **(B)** The 192 sense dicodons overlapping an antisense stop codon in frame-2. We have highlighed in red font the reverse complementary sequence corresponding to an antisense stop codon.

frame-2 stop codon is identical to that of a stop codon in an intergenic locus with identical GC-content.

3. Why asORFs appear to be most probable in frame 0, in *D. melanogaster* but not in *S. cerevisiae*

We analysed the differences between the predictions from the two species more closely. The most salient difference exists in the probability of asORFs in frame 0. The reason is that stop codons in frame 0 are 2.7 times more likely in *S. cerevisiae* than in *D. melanogaster* (Table S3). Therefore we analysed the frequency of these codons and their specific usage to encode the corresponding amino acids.

Stop codons in frame 0 overlap with the codons – TTA, CTA (coding for leucine) and TCA (coding for serine). Both leucine and serine are encoded by six codons. We analysed the coding regions of *S. cerevisiae* and *D. melanogaster* to estimate the codon usage for leucine and serine in both these organisms. We found that the total frequencies of leucine and serine are similar betwen the two organisms. However, the codons that overlap with an antisense stop codon are more frequently used in *S. cerevisiae* than in *D. melanogaster* (Figure S1).

	S. cerevisiae	*D. melanogaster*
Start codon	0.0169	0.0172
Stop codon: Frame 0	**0.0592**	**0.0216**
Stop codon: Frame 1	0.0399	0.0319
Stop codon: Frame 2	0.0482	0.0423

Table S3: Probability of start and stop codons in the three different antisense frames, calculated using distribution of codons and dicodons in *S. cerevisiae* and *D. melanogaster* coding sequences.

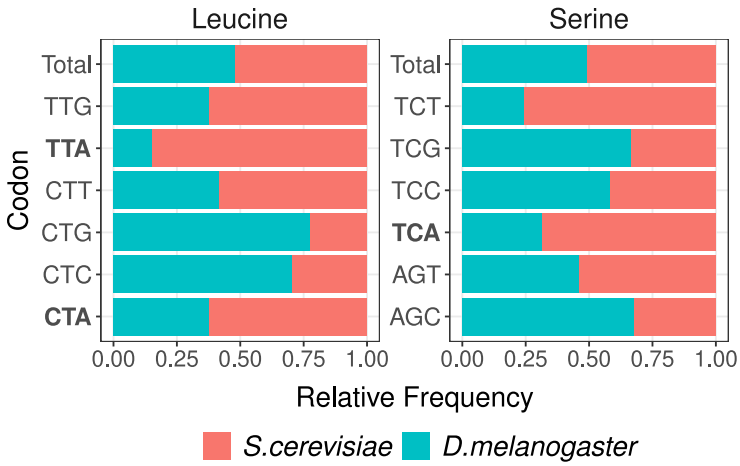

Figure S1: Codon usage of leucine and serine in *S. cerevisiae* and *D. melanogaster*. Codons highlighted in bold overlap with an antisense stop codon.

4. Distribution of antisense ORFs in *Drosophila melanogaster* genome

We identified antisense ORFs and intergenic ORFs using genome and transcriptome data from seven *D. melanogaster* lines (Grandchamp *et al.*, 2023b). We performed the same analysis for every *D. melanogaster* line. Specifically, we first obtained the genome assembly, genome annotations and transcriptome assembly for each line (Grandchamp *et al.*, 2023b). Next, we identified RNAs that overlap in antisense to any annotated protein coding gene. Next, we extracted ORFs in these antisense RNAs using *getorf* (Rice *et al.*, 2000). Next, we mapped the genomic coordinates of these ORFs using nucleotide BLAST (100% query coverage and sequence identity; Altschul *et al.*, 1990; Camacho *et al.*, 2009), and identified all asORFs and their frame of overlap using *awk* scripts (we note that not all ORFs in antisense RNAs are asORFs). Finally, we only analysed asORFs whose genomic sequences were uninterrupted by introns (Table S4).

	Denmark	**Finland**	**Spain**	**Sweden**	**Türkiye**	**Ukraine**	**Zambia**
Total antisense loci	1440501	1272815	1326269	1119960	1641755	1143721	1200121
Expected asORF0	314 (327)	276 (290)	291 (304)	250 (256)	361 (374)	252 (262)	269 (279)
Observed asORF0	276	144	253	178	194	179	175
Expected asORF1	371 (299)	325 (265)	343 (279)	296 (230)	428 (338)	297 (239)	319 (252)
Observed asORF1	483	300	391	391	469	377	430
Expected asORF2	397 (327)	348 (290)	367 (304)	318 (256)	459 (374)	318 (262)	342 (279)
Observed asORF2	251	150	179	201	181	138	226
Total intergenic loci	2147483647	2147483647	2147483647	2147483647	2147483647	2147483647	2147483647
Expected igORF	1707687 (1768465)	1776758 (1839181)	1840872 (1906004)	1761983 (1823809)	1808499 (1873396)	1760268 (1822705)	1690669 (1750696)
Observed igORF	1763975	1828152	1889493	1807274	1858731	1811161	1740461

Table S4: Summary of antisense and intergenic ORFs identified in *D. melanogaster* lines. Expected numbers of ORFs within parantheses were estimated using GC-content of each locus, whereas those outside the parantheses were estimated using DNA oligomer frequencies. The different asORFs reported here include sub-ORFs within longer ORFs detected by *getorf*. Here we only report asORFs that do not contain introns and that completely overlap with a protein coding exon (sense ORF).

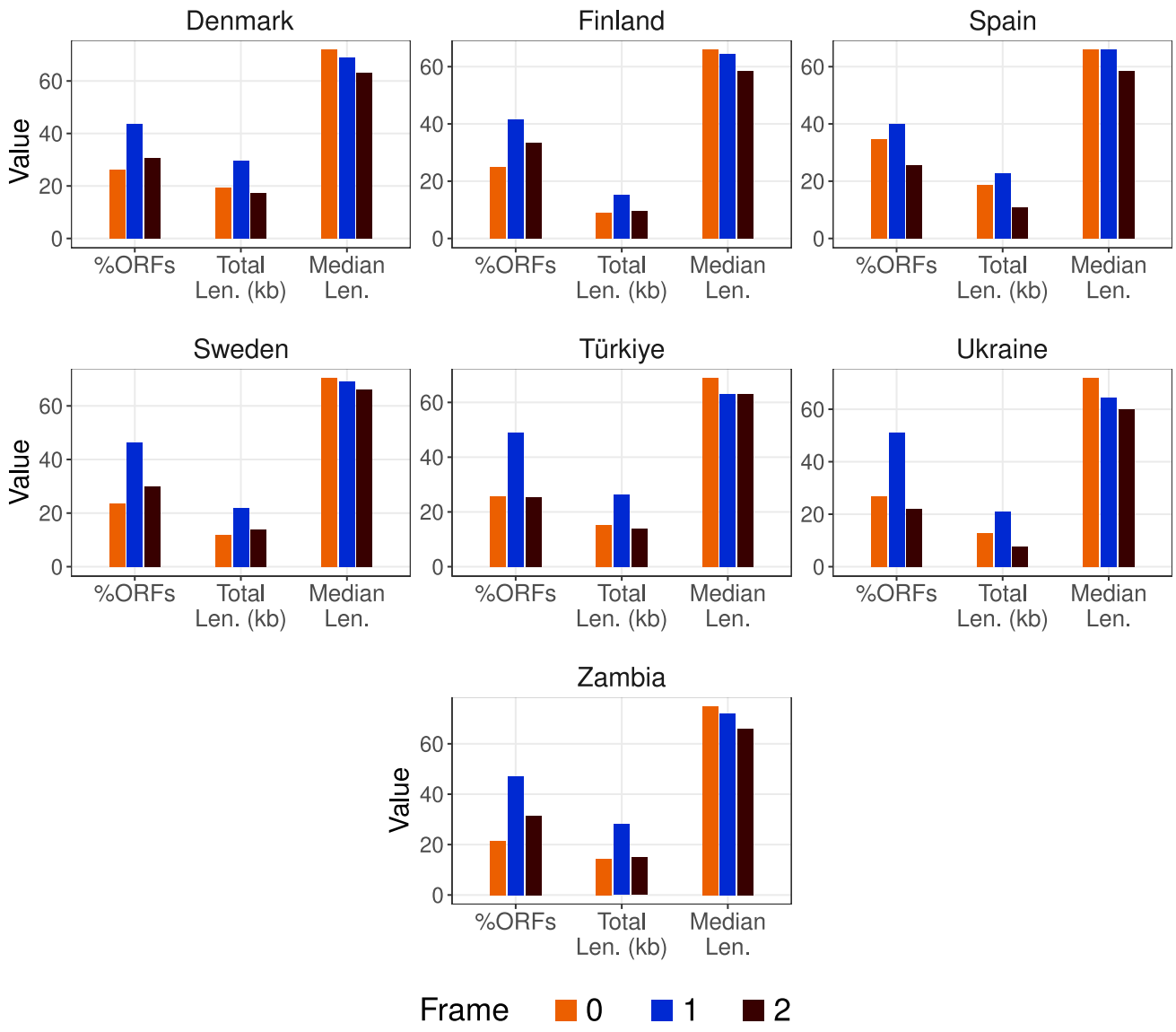

Figure S2: Frame preference of antisense ORFs in *D. melanogaster* genome that have 100% overlap with a codon exon. We show three metrics of frame preference as three bar groups – percentage of total ORFs (left), cumulative length of all antisense ORFs (middle), and median ORF length (right), in each of the three frames (bar colors). We calculated these metrics from the genomics and the transcriptomics data from the seven different *D. melanogaster* lines (Grandchamp *et al.*, 2023b,a).

5. Translational efficiency of asORFs

To estimate the translational efficiency of asORFs in *S. cerevisiae*, we used data from a recently published study (Wacholder *et al.*, 2023). This large dataset (iRibo) has been compiled from different published ribo-seq (sequencing of ribosomal footprint) experiments in *S. cerevisiae* such that every ORF (predicted or annotated) is assigned a number of reads that are in-frame with the ribosome's elongation periodicity. For every antisense-ORFs (as annotated by this study), we extracted the number of reads, and calculated the frame of overlap. We note that iRibo dataset is recent and was not available when we started our study. However, our analysis of asORFs from iRibo agrees with our model's predictions, and qualitatively agrees with the observed frequencies of asORFs shown in Table 2 and Figure 2 (Figure S3A/B). More specifically, the asORFs in frame 1 are significantly more numerous than those in the other two frames (Figure S3A; one tailed Fisher exact test, FDR corrected $P < 10^{-22}$). The asORFs in frame 1 are also significantly longer than those in the other two frames (Figure S3B; one tailed Mann-Whitney U test, FDR corrected $P < 10^{-22}$). Next, we analysed if asORFs in frame 1 have more riboseq reads than those in the other two frames. We found that asORFs in frame 1 have significantly more reads than asORFs in frame 0 (one tailed Mann-Whitney U test, FDR corrected $P = 7.5 \times 10^{-3}$) but not asORFs in frame 2 (one tailed Mann-Whitney U test, FDR corrected $P = 0.115$). This does not indicate that there is no significant difference in the total translational output for asORFs in the different frames. That is so because both the number of asORFs and the translational efficiency is responsible for translational output. We found that the total translational output is significantly higher for asORFs in frame 1 than those in the other two frames (Figure S3A; one tailed Fisher exact test, FDR corrected $P < 10^{-22}$). Next, we compared the number of riboseq reads of the different asORFs and igORFs. We found that igORFs had a significantly larger number of reads than all asORFs (Figure S3B; one tailed Fisher exact test, FDR corrected $P < 10^{-22}$). More interestingly, the riboseq read count distribution of igORFs was bimodal. Specifically, a subset of igORFs was expressed more than the other

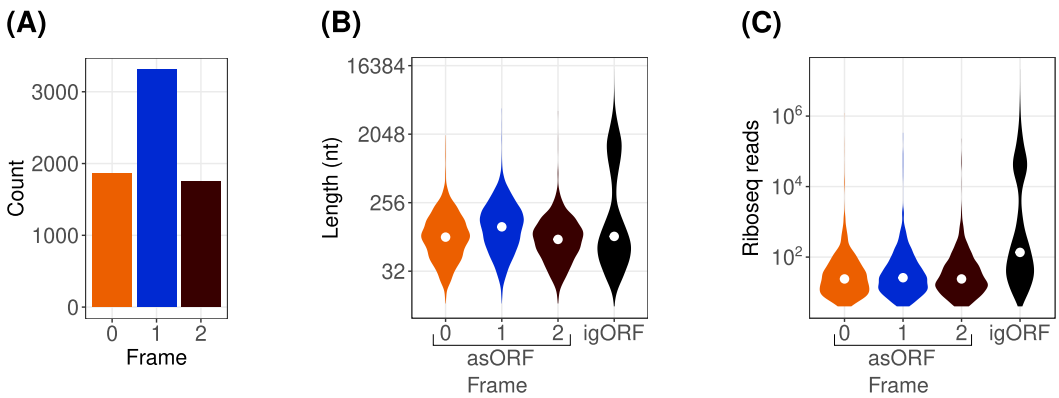

Figure S3: Yeast asORFs from iRibo (Wacholder *et al.*, 2023). Number of total asORFs (**A**, vertical axis), ORF length distribution of asORFs and igORFs (**B**, vertical axis), and riboseq reads distribution of asORFs and igORFs (**C** vertical axis), in each of the three frames (horizontal axis). We only show asORFs that overlap 100% with the sense ORF.

subset, by two orders of magnitude. Interestingly, the length distribution of igORFs was also bimodal. These observations suggest that there are two different kinds of igORFs. The longer and highly translated igORFs could have undergone adaptive evolution.

To perform an analogous analysis for *D. melanogaster* asORFs, we did not find a compiled resource like iRibo. Therefore we used Kozak consensus sequence (KCS) score (Acevedo *et al.*, 2018) and ORF position in the RNA as proxies of translational efficiency as shown in another study (Patraquim *et al.*, 2022). We did not find any statistically significant difference between the values of these parameters for the different frames, that is also consistent across the seven different *D. melanogaster* lines (Mann-Whitney U test, 95% confidence interval, FDR corrected). We also did not find any significant difference between the KCS scores of igORFs and any of the three kinds of asORFs (Mann-Whitney U test, 95% confidence interval, FDR corrected).

6. Gain and loss probabilities of antisense ORFs in *Drosophila melanogaster*

6.1 Model predictions

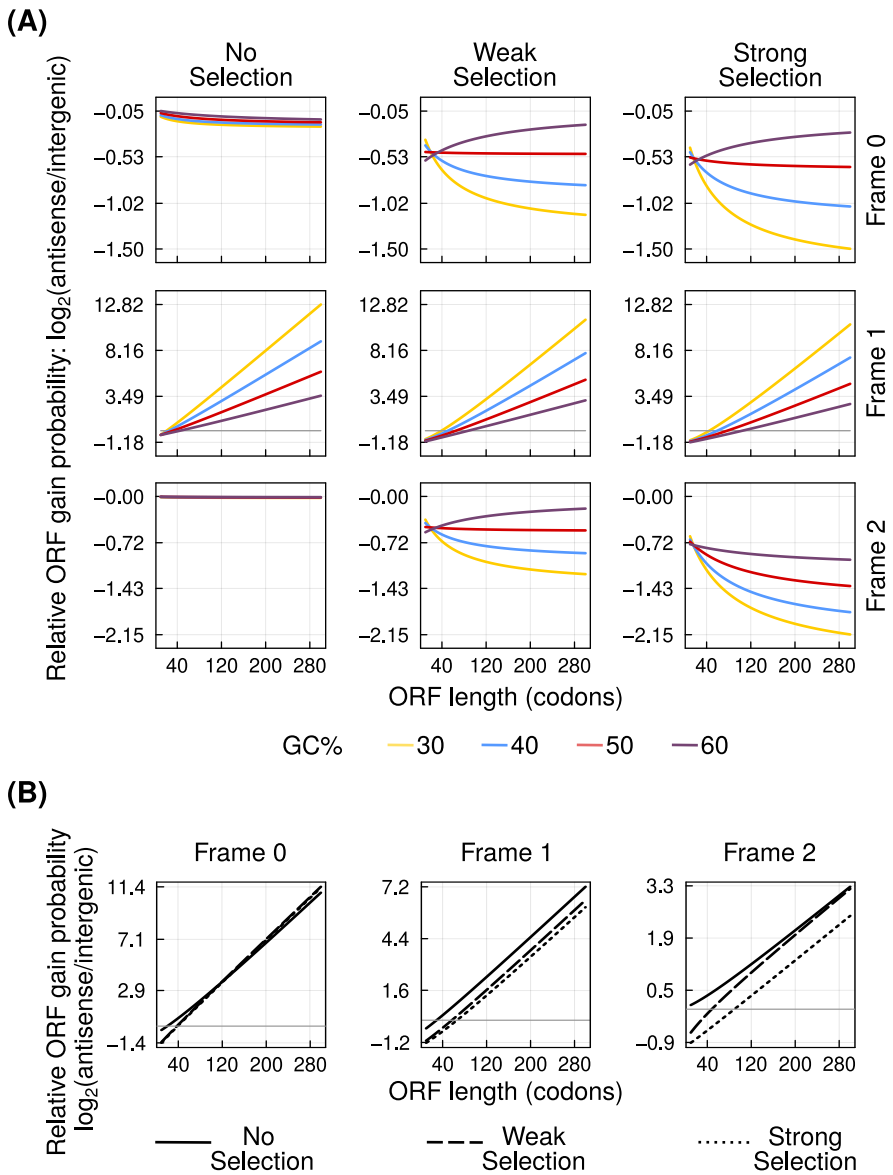

Figure S4: Antisense overlap can facilitate ORF emergence. Panel **(A)** shows the probability of ORF emergence in the three antisense frames (left to right) relative to that in intergenic regions ($\log_2$ ratio, vertical axis), at different intensities of purifying selection (top to bottom). Line colors indicate the GC-content of the ORFs. Panel **(B)** shows the ORFs gain probability in the three antisense frames relative to that in intergenic regions ($\log_2$ ratio, vertical axis), calculated using frequencies of short DNA sequences from *D. melanogaster* genome. Dotted, solid and dashed lines, denote the zero, weak and strong purifying selection, respectively. Horizontal axis in all panels shows the length of the ORFs. For data in both panels, we assume that antisense ORFs overlap completely with the sense ORF.

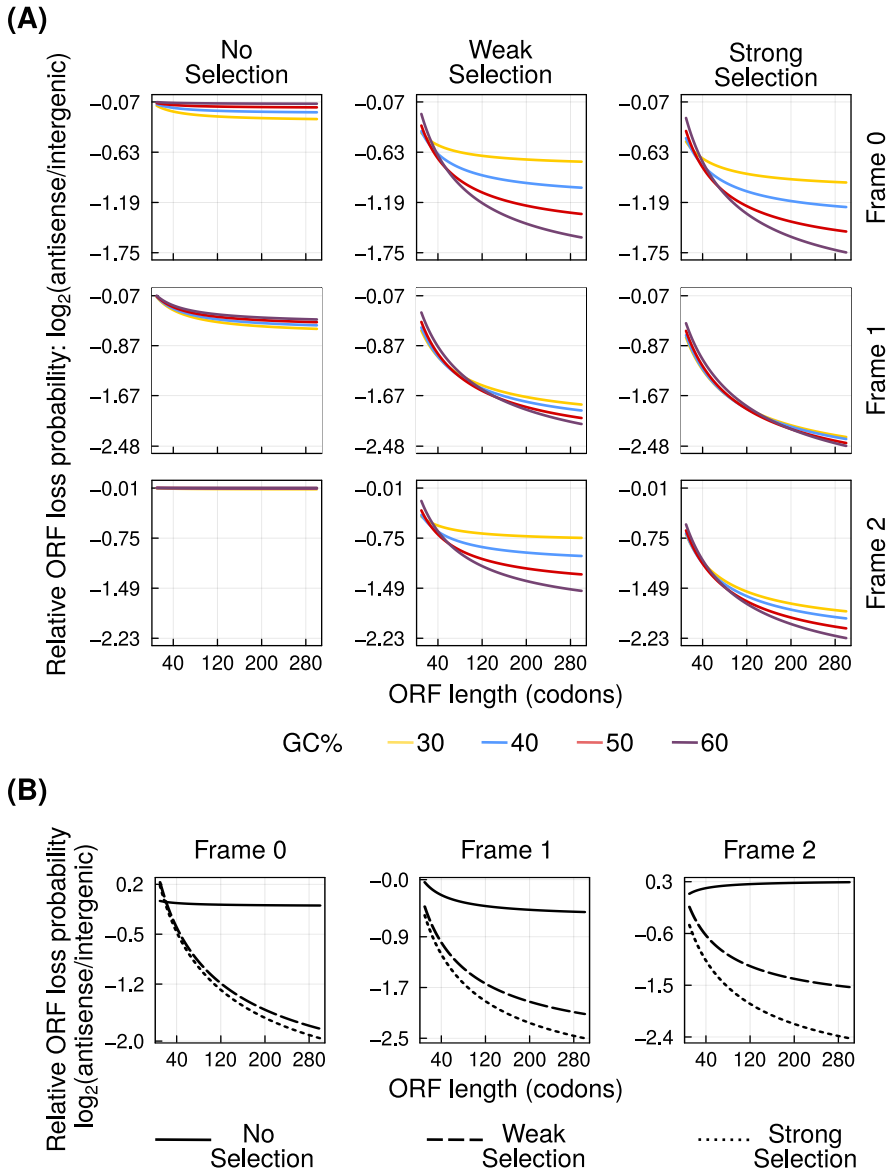

Figure S5: Antisense overlap can reduce ORF loss. Panel **(A)** shows the probability of ORF loss in the three antisense frames (left to right) relative to that in intergenic regions ($\log_2$ ratio, vertical axis), at different intensities of purifying selection (top to bottom). Line colors indicate the GC-content of the ORFs. Panel **(B)** shows the ORFs loss probability in the three antisense frames relative to that in intergenic regions ($\log_2$ ratio, vertical axis), calculated using frequencies of short DNA sequences from *D. melanogaster* genome. Dotted, solid and dashed lines, denote the zero, weak and strong purifying selection, respectively. Horizontal axis in all panels shows the length of the ORFs. For data in both panels, we assume that antisense ORFs overlap completely with the sense ORF.

6.2 Analysis of asORF gain and loss using genomics data

To estimate gain and loss of asORFs we compared their presence or absence in the transcriptome of the different *D. melanogaster* lines. We assume that an ORF emerges only once. That is, if an ORF is detected in five lines, we assume that it emerged once and spread in five lines.

In the first step, we identified ORFs that were shared by several lines. We call defined an

orthogroup as a group of query unique ORF sequences detected in any of the seven lines. Our definition of orthology in this case is very stringent. If an ORF duplicated in two lines, we classified the duplicated copies into two separate orthogroups. That is so because we were interested in the gain and loss of the original ORF and its duplicated copy separately. We also discarded orthogroups where the ORFs from the different lines were not located in the same

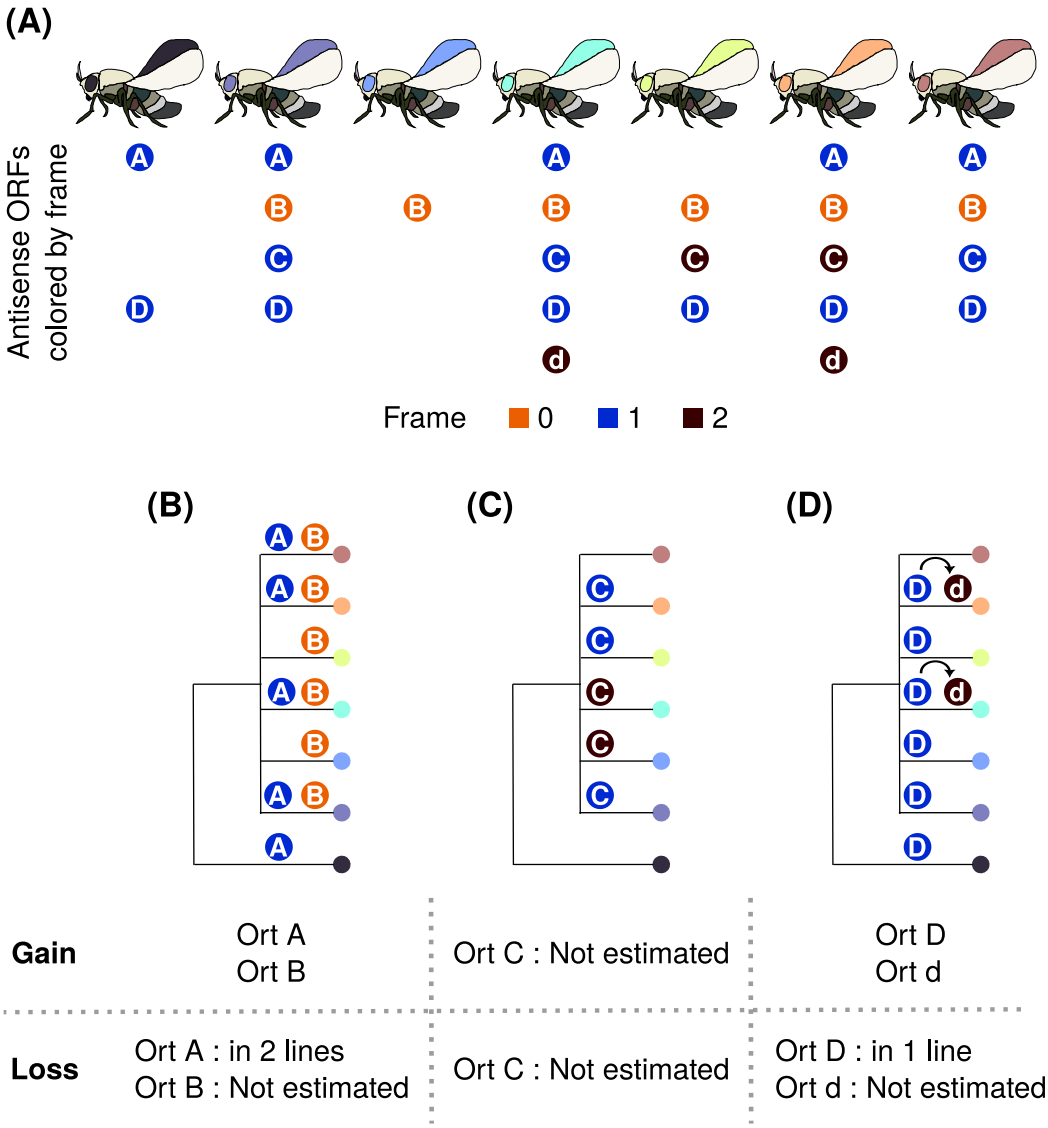

Figure S6: Summary of ORF gain and loss analysis in the seven *D. melanogaster* lines. **(A)** Hypothetical asORF orthogroups denoted by alphabets (A – D) with their frames denoted by the colors orange (0), blue (1) and brown (2). **(B)** The hypothetical example of the orthogroups A and B (containing ORFs A and B, respectively). In both the orthogroups, the ORFs are systematically located in the same frame in every line where they are present. For each of the two orthgroups, we count one gain event. ORF-A is detected in the Zambian outgroup line, but not in the European lines. Thus this ORF is lost in two lines. Because, ORF-B is not detected in the Zambian line, we do not analyse its loss. **(C)** ORF-C is detected in several lines but was located in different frames in the different lines. Thus we do not use this orthogroup for our analysis. **(D)** ORF-D is present in six lines, and has duplicated in two lines (denoted as ORF-D and ORF-d). The duplicated copy (ORF-d) is located in an different frame as ORF-D. Therefore, we classify them into consider 2 orthogroups – the orthogroup containing ORF-D, which is present in the Zambian and some European lines, so that we can estimate its loss. The orthogroup containing ORF-d is only present in two European lines, and therefore we cannot estimate if it was indeed lost in the other lines or only gained in these two lines.

frame. We did so because if would be difficult to infer in which frame (line) the ORF gain occurred first.

To identify orthogroups, we used nucleotide BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2009). We used nucleotide BLAST instead of protein BLAST for a specific reason – we wanted to identify orthologous asORFs that may be frameshifted. In case of a frameshift, BLASTp may not detect any homology. For the BLASTn analysis, we used an e-value cutoff of 10^{-2} and required a 100% query coverage. Furthermore, we verified that the orthologous asORFs antisense-overlapped with the same protein coding gene. Given these criteria, our algorithm picks the highest scoring hit if there are multiple hits. To keep the analysis focused and less complicated, we only analysed asORF orthologs in which the frame was conserved. Thus our BLAST analyis is overall quite stringent.

Most orthogroups contained only one ORF per line. However, some orthogroups contained several ORFs in a single line, due to tandem duplications. We split these orthogroups such that they contained only one ORF per line, and sorted them according to their frame and the overlapping "sense" ORF. Among the 3536 orthogroups we detected, 105 had several ORFs in several lines. 32 out of these 105 orthogroups contained more than four duplicates in some lines. We discarded these orthogroups because we could not reliably categorize them into sub-orthogroups after splitting them based on frame and position. We also discarded 147 orthogroups were from our analysis because the homologous ORFs were located in different frames.

To estimate the loss, we used the outgroup (Zambian) line. The Zambian populations separated from the European populations between 14000 – 30000 years ago (Li and Stephan, 2006; Laurent *et al.*, 2011). Therefore, if an ORF was found in the outgroup and at least one European line, we assume that it emerged in an ancestral *D. melanogaster* population and was lost in rest of the five European lines. We found 319 orthogroups where the ORF was present in the Zambian line and at least one European line but not all six of them.

7. Effect of mutations on asORFs

In the previous sections, we showed that purifying selection on the sense ORF can affect the emergence and loss of asORFs. We next asked if this purifying selection can also constrain the diversification of the proteins encoded by asORF sequences. To this end, we first calculated the "chemical distance" (δ) between any two amino acids. For this calculation we used a distance matrix that we derived from an experimentally estimated amino acid similarity matrix reported in a previous study (Kim *et al.*, 2009). Next, we calculated the average chemical difference ($\bar{\delta}$) introduced by a random mutation, weighted by the probability of different mutations

(Equation 1). To this end, we created an amino acid distance matrix by modifying the amino acid similarity matrix of Kim *et al.* (2009). Specifically, we subtracted the value of 0.3 from each element of the matrix, reversed the sign of each element, and set the diagonal to zero. By doing this, we set every distance value to be greater than 0. Next, we calculated the average chemical difference introduced by any mutation ($i \rightarrow j$) allowed under a selection regime. Specifically, if i denotes the original codon, j denotes the substituted codon, P_i denotes the probability of finding codon-i, μ_{ij} denotes the probability of codon-i mutating to codon-j, and δ_{ij} denotes the chemical difference between the amino acids encoded by these codons, then the average chemical difference is defined by the following equation:

$$\bar{\delta} = \frac{\sum_i P_i \sum_j \mu_{ij}\delta_{ij}}{\sum_i P_i \sum_j \mu_{ij}} \tag{1}$$

Using $\bar{\delta}$ as a measure of divergence, we estimated the extent to which asORFs in the three frames can diverge as a result of mutations, and due to purifying selection on the sense ORFs. Likewise, we also calculated the divergence of intergenic ORFs as a consequence of random mutations. We found that frame 2 allows maximum divergence of asORFs, under both weak and strong purifying selection on the sense ORF (Figure S7A). asORFs in frame 0 diverge the least. Interestingly, strong selection on sense ORFs increases the divergence of asORFs in frame 2. The reason could be that the few mutations that do occur under strong purifying selection, cause a relatively higher increase in divergence than the more numerous mutations that are allowed to occur under weak purifying selection. We also found that the divergence of asORFs in frame 2 was higher than than that of intergenic ORFs under both selection regimes. We note this result does not mean that intergenic ORFs can diverge less than asORFs. Evolution of intergenic ORFs is not constrained by another DNA sequence. However, as long as the mutants do not affect the organismal fitness, evolution would not be biased towards divergence increasing mutations. Thus random mutations in intergenic ORFs could also consist of many synonymous and chemistry preserving mutations, that are probably disallowed in frame 2 asORFs due to purifying selection on sense ORFs.

In contrast to frame 2, the divergence of asORFs in the other two frames decreased with increasing strength of purifying selection on the sense ORF (Figure S7A). For example, asORFs in frame 0, did not diverge at all when the sense ORF was under strong purifying selection. asORFs in frames 0 and 1 also diverged less than intergenic ORFs under both selection regimes.

We observed identical trends in divergence of asORFs from our analysis based on *D. melanogaster* parameters (Figure S7B).

These findings not negate the fact that intergenic ORFs have less constraints on their evo-

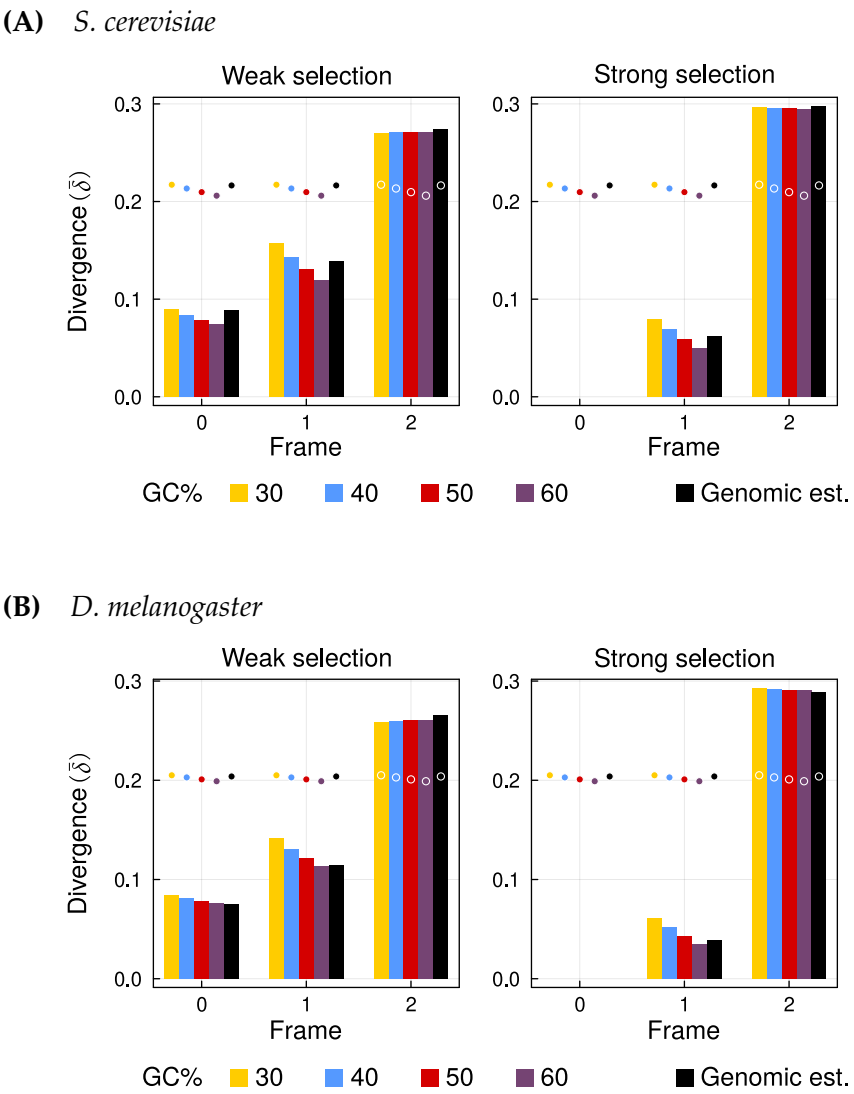

Figure S7: Antisense ORFs in **(A)** *S. cerevisiae* and **(B)** *D. melanogaster*, can diversify when sense ORFs are under purifying selection. Vertical axis denotes the divergence ($\bar{\delta}$) of asORFs due to a random mutation when the sense ORF is under weak (left) or strong (right) purifying selection. Horizontal axes denote the three antisense frames. Colored bars denote divergence values of asORFs with different GC-content, and black bars denote the diversity values calculated using frequencies of short DNA sequences from the yeast genome. Filled circles that are similarly color coded, denote the divergence of intergenic ORFs due to mutations.

lution. Even though chemical consequences of tolerated mutations may be larger in some asORFs than in intergenic ORFs, purifying selection on the sense ORF limits the total number of possible mutations. This would not be the case for intergenic ORFs.

8. Is GC-content a better parameter for asORF probability calculation than global DNA oligomer frequencies?

Any calculation made using an averaged nucleotide composition distribution is likely to be an approximation. It is true for both GC-content (for example, using the average genomic

GC-content) or average distribution of DNA oligomers across different genomic loci. Both GC-content and oligomer distribution can be calculated for specific loci, which can make the calculations more realistic. In our plots for of stationary, gain and loss probability based on GC-content (Figures 1B, 3A and 4A), we show four different values of GC-content. They are correct as long as our assumptions hold true. The plots based on DNA oligomer frequencies (Figures 1C, 3B and 4B) may be less realistic because they assume that the oligomer distribution is uniform across the genome (CDS or intergenic regions). Thus the GC-content based plots are more informative.

To understand how realistic averages can be, we performed an empirical analysis of variance of nucleotide composition. Specifically, we normalized the distribution such that the sum of frequencies of a trimer (or GC-fraction) across all loci is equal to one, and calculated the variance of this distribution. We found that GC-content has a smaller variance than that of any DNA trimer (Figure S8). However, this empirical analysis does not prove that GC-content is a better estimate of the real nucleotide distribution.

Ultimately, the most realistic analysis would estimate parameters from each locus separately, and estimate the ORF probabilities specific to that locus. We have indeed done so for calculating expected number of ORFs based on GC-content (main text Table 2). To this end, we calculated the GC-content of each contiguous intergenic or antisense overlapping region, and estimated the ORF probability as well as expected number of ORFs using this specific GC-content. We found that the expected number of ORF using global DNA trimer distribution

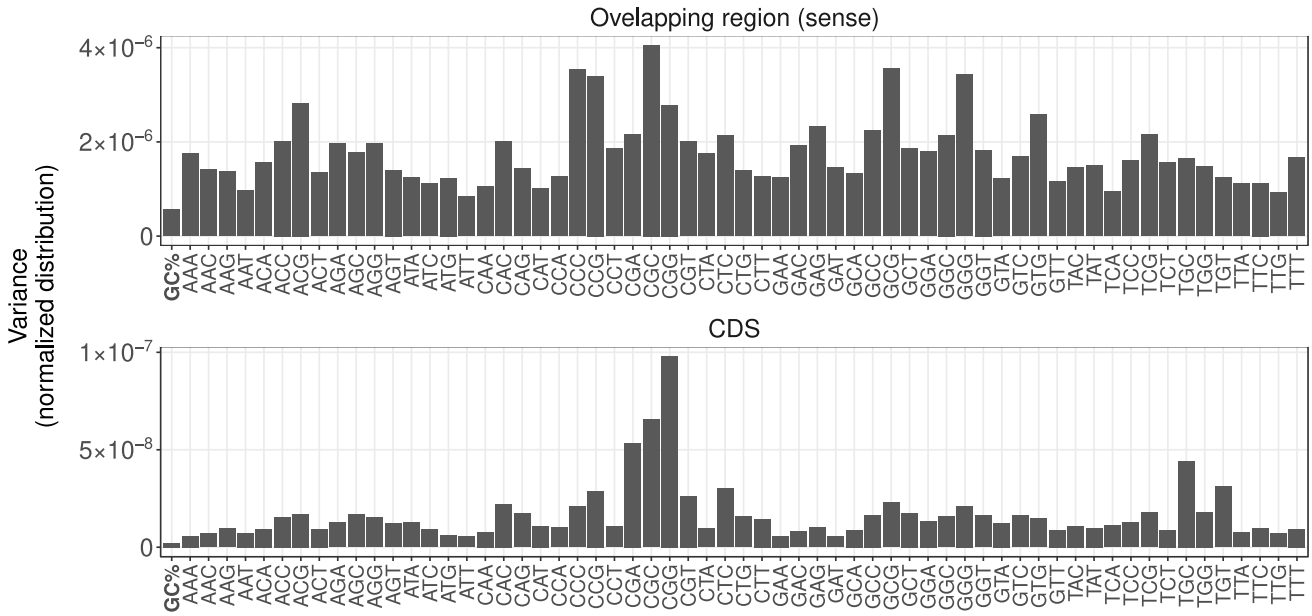

Figure S8: Variance of the normalized distribution of GC-content and of different DNA trimers in *S. cerevisiae*. For coding regions we calculated the frequencies of the different codons as they exist in annotated ORFs (top panel), whereas for regions overlapping with antisense ORFs, we calculated the distribution of DNA trimers using a sliding window (bottom panel). We have excluded stop codons from both the panels.

and locus specific GC-content do not differ significantly.

References

Acevedo, J. M., Hoermann, B., Schlimbach, T., and Teleman, A. A. 2018. Changes in global translation elongation or initiation rates shape the proteome via the Kozak sequence. *Scientific Reports*, 8(1): 4018.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3): 403–410.

Camacho, C., Coulouris, G., Avagyan, V., and others 2009. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1).

Grandchamp, A., Czuppon, P., and Bornberg-Bauer, E. 2023a. High turnover of *de novo* transcripts in *Drosophila melanogaster*. *bioRxiv*.

Grandchamp, A., Kühl, L., Lebherz, M., and others 2023b. Population genomics reveals mechanisms and dynamics of de novo expressed open reading frame emergence in drosophila melanogaster. *Genome Research*, 33(6): 872–890.

Kim, Y., Sidney, J., Pinilla, C., Sette, A., and Peters, B. 2009. Derivation of an amino acid similarity matrix for peptide:MHC binding and its application as a bayesian prior. *BMC Bioinformatics*, 10(1).

Laurent, S. J., Werzner, A., Excoffier, L., and Stephan, W. 2011. Approximate bayesian analysis of drosophila melanogaster polymorphism data reveals a recent colonization of southeast asia. *Molecular Biology and Evolution*, 28(7): 2041–2051.

Li, H. and Stephan, W. 2006. Inferring the demographic history and rate of adaptive substitution in drosophila. *PLoS Genetics*, 2(10): e166.

Patraquim, P., Magny, E. G., Pueyo, J. I., Platero, A. I., and Couso, J. P. 2022. Translation and natural selection of micropeptides from long non-canonical RNAs. *Nature Communications*, 13(1).

Rice, P., Longden, I., and Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(6): 276–277.

Schrider, D. R., Houle, D., Lynch, M., and Hahn, M. W. 2013. Rates and Genomic Consequences of Spontaneous Mutational Events in Drosophila melanogaster. *Genetics*, 194(4): 937–954.

Wacholder, A., Parikh, S. B., Coelho, N. C., and others 2023. A vast evolutionarily transient translatome contributes to phenotype and fitness. *bioRxiv*.

Zhang, Z. and Gerstein, M. 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Research*, 31(18): 5338–5348.