Modeling Length Changes in *De Novo* Open Reading Frames during Neutral Evolution

Marie Kristin Lebherz[1,†], Bharat Ravi Iyengar [1,*,†], Erich Bornberg-Bauer [1,2,*]

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

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

[†]Equal contribution.

*Corresponding authors: E-mails: b.ravi@uni-muenster.de; ebb@uni-muenster.de.

Accepted: June 06, 2024

Abstract

For protein coding genes to emerge *de novo* from a non-genic DNA, the DNA sequence must gain an open reading frame (ORF) and the ability to be transcribed. The newborn *de novo* gene can further evolve to accumulate changes in its sequence. Consequently, it can also elongate or shrink with time. Existing literature shows that older *de novo* genes have longer ORF, but it is not clear if they elongated with time or remained of the same length since their inception. To address this question we developed a mathematical model of ORF elongation as a Markov-jump process, and show that ORFs tend to keep their length in short evolutionary timescales. We also show that if change occurs it is likely to be a truncation. Our genomics and transcriptomics data analyses of seven *Drosophila melanogaster* populations are also in agreement with the model's prediction. We conclude that selection could facilitate ORF length extension that may explain why longer ORFs were observed in old *de novo* genes in studies analysing longer evolutionary time scales. Alternatively, shorter ORFs may be purged because they may be less likely to yield functional proteins.

Key words: *de novo* gene emergence, gene evolution, protein evolution, mathematical modeling, genomics.

Significance

New protein coding genes can emerge from non-genic DNA through a process called *de novo* gene emergence. Genes thus emerged usually have a small open reading frame (ORF). However, studies show that *de novo* genes with an older evolutionary origin have longer ORF than younger genes. To understand how ORF length evolves, we use a combination of mathematical modeling and population level genome data analysis. We find that in the absence of evolutionary selection, ORFs are likely to retain their length in short evolutionary timescales. However, when their length does change then it is likely to decrease than to increase. Therefore, *de novo* genes with long ORFs are probably selected by evolution to be retained in the genome.

Introduction

De novo gene birth is a phenomenon by which new protein coding genes can emerge from previously non-genic regions of the genome (Schmitz and Bornberg-Bauer 2017; Van Oss and Carvunis 2019). This process, that was long thought to be unlikely (Jacob 1977), is now being increasingly better documented (Carvunis et al. 2012; Neme and Tautz 2013, 2016; Zhao et al. 2014; Baalsrud et al. 2017; Gubala et al. 2017; Vakirlis et al. 2017, 2020; Prabh and Rödelsperger 2019; Witt et al. 2019; Blevins et al. 2021; Lange et al. 2021; Wacholder et al. 2023). For any stretch of DNA to qualify as a putative protein

coding gene, it needs to be transcribed, as well as contain an open reading frame (ORF). Further, for a gene to be considered a *bona fide* protein coding gene, its ORF also needs to be translated. While protein coding genes that arise from existing protein coding genes can inherit the sequence features necessary for transcription and translation from their ancestors, a prospective *de novo* gene must evolve them sequentially through random mutations.

A novel protein coding gene is more likely to be fixed in an organism's genome, if its products increase the fitness of the organism. Conversely, the gene is unlikely to be fixed if its products are toxic or are too expensive for the organism. Because a *de novo* gene does not inherit functional features from a pre-existing gene, its effect on organismal fitness is difficult to predict. While some *de novo* genes have been known to increase the fertility or survivability of an organism (Baalsrud et al. 2017; Gubala et al. 2017; Lange et al. 2021), the fitness effects of the vast majority of them is unknown. It is possible that a tiny fraction of newly emerged *de novo* genes are beneficial to the organism by chance. However, most of them may not affect organismal fitness at all, or can possibly even be detrimental by causing proteotoxicity (Bucciantini et al. 2002; Hartl 2017).

A protein's activity is closely linked to how it folds. An established study on protein folding suggests that globular proteins are most likely to fold if they contain more than 70 but less than 2,000 residues, and that there exists an optimal protein length that is the best for folding (Dill 1985). While our understanding of protein folding has advanced significantly, it is still not clear how protein length affects its folding. Unfolded proteins can misfold and cause proteotoxicity (Bucciantini et al. 2002; Hartl 2017). However, not all proteins need to be well folded to be functional (Dyson and Wright 2005; Wright and Dyson 2014). For example, many structurally disordered proteins are involved in cell signaling (Wright and Dyson 2014).

Several studies have shown that conserved proteins are typically longer than putative *de novo* proteins. Furthermore, evolutionarily older *de novo* genes encode longer proteins than younger genes (Carvunis et al. 2012; Neme and Tautz 2013; Zhao et al. 2014; Vakirlis et al. 2017; Dowling et al. 2020; Heames et al. 2020; Blevins et al. 2021; Papadopoulos et al. 2021; Middendorf and Eicholt 2024). There is a simple explanation for why the protein coding regions (ORFs) of young genes are short. That is because the likelihood of finding an ORF by chance, as well as the likelihood of ORF emergence, reduces exponentially with ORF length (Iyengar and Bornberg-Bauer 2023). Thus, it is not known whether it is more likely that ORFs of older genes become longer with age (Papadopoulos et al. 2021), or if the genes with long ORFs are more likely to be fixed, even though they are less likely to occur by chance.

We attempt to address this question in this study. To this end, we use a combination of mathematical modeling and analysis of genome sequencing data. Mathematical models are useful to understand processes that cannot be easily explained through intuition alone. For example, it had been suggested using empirical analyses that emergence of transcription is likely to precede the emergence of ORF during *de novo* gene emergence (Ruiz-Orera et al. 2014; Zhang et al. 2019; Blevins et al. 2021). In a recent work, we explained this observation with a more formal analysis using mathematical models (Iyengar and Bornberg-Bauer 2023). In this study, we developed a mathematical model of ORF length change in *de novo* genes. Because little is known about how *de novo* genes affect organismal fitness, our model is based on the assumption of evolutionary neutrality. Although this does not explain evolutionary dynamics of *de novo* genes through selection, it does provide a good null hypothesis against which observations could be tested. To validate some of our model's predictions, we analyse the genome data for seven *Drosophila melanogaster* populations, and identify how ORF length changes in a short evolutionary timescale. These datasets were generated in a previous study that created inbred lines from a sample of *D. melanogaster* populations from seven different geographical locations, sequenced their genome and transcriptome using deep sequencing, and identified several putative *de novo* protein coding genes in their genomes (Grandchamp et al. 2023a). A neutral gene can be lost due to drift, the likelihood of which increases with the gene's age. Thus more young genes are likely to be neutrally evolving than older genes. Our analyses of populations that have diverged over a much smaller evolutionary timescale than that of different species, allows us to study young *de novo* genes that may still be evolving neutrally.

Using our two-way approach, we found that ORF of young *de novo* genes are more likely to remain unchanged after many generations. However, when their length does change then it is more likely to reduce than to increase. This suggests that neutral evolutionary theory alone cannot explain why older *de novo* genes have longer ORF, and thus selection must be considered to explain this outcome.

Results

Development of the Mathematical Model

We developed a mathematical model to study the dynamics of ORF length evolution, under the assumption that no evolutionary selection occurs for any ORF (neutrality). Specifically, we modeled a Markov jump process, where an ORF can become longer or shorter, or remain of the same length, at any discrete generation. The likelihood of ORF length change is described by a "transition probability", which is a function of initial ORF length and final ORF length. ORF length can change via substitutions, insertions, deletions, transpositions, and even chromosomal

recombination. In this study, we focus on ORF length changes that arise due to nucleotide substitutions that are the most frequent kind of mutations (Schrider et al. 2013; Zhu et al. 2014). The likelihood of different nucleotide substitutions in a specific genomic locus, depends on mutation rate bias as well as nucleotide composition of the locus (Iyengar and Bornberg-Bauer 2023). Therefore the ORF length transition probabilities also depend on these two parameters. We used mutation rate bias data from two different organisms—the budding yeast, *Saccharomyces cerevisiae* and the fruitfly, *D. melanogaster* for calculating transition probabilities. We chose these two organisms because both of them have been frequently studied in the context of *de novo* gene emergence (Carvunis et al. 2012; Zhao et al. 2014; Gubala et al. 2017; Vakirlis et al. 2017, 2020; Witt et al. 2019; Blevins et al. 2021; Lange et al. 2021; Wacholder et al. 2023). We denoted nucleotide composition as GC-content or the frequency of trinucleotides in the intergenic regions from each of the two organisms. Although the ORF length transition probability is determined by nucleotide composition and biased mutation rate, it remains constant over time for any one specific locus. Therefore, the probability distribution of ORF length at a specific locus at any one generation depends only on the same distribution at the previous generation; hence length change is a Markov process.

An ORF can be extended or truncated from both ends through gain or loss of start and stop codons (Fig. 1). For example, an ORF extends from the 3′ end if it loses its stop codon, there exists another in-frame stop codon in its 3′ untranslated region (UTR), and no other stop codon exists between the old and the new stop codon positions. For the ORF to shorten from the 3′ end it only needs to gain a premature stop codon. The mechanisms of ORF extensions and truncations from the 5′ end are more diverse than that from the 3′ end. For example, a 5′ extension can occur if a new in-frame start codon emerges in the 5′UTR and there exist no stop codons between the new and the old start codons. A 5′ extension could also occur if an ORF fuses with an in-frame upstream ORF. This can happen if the stop codon of the upstream ORF is lost and there are no stop codons in the intervening sequence between the two ORFs. Finally, a 5′ extension can also occur if the RNA itself is extended from the 5′ (for example, through an alternative transcription start site), that gives rise to a new in-frame start codon in the RNA. Conversely, 5′ truncations can occur via loss of start codon either due to mutations or due to RNA truncation, and splitting of an ORF into two shorter ORFs by a gain of stop codons between two start codons. We note that an ORF can also fuse with a downstream ORF after loss of its stop codon, but this mechanism is in principle identical to that of ORF extension to the next in-frame stop codon in a UTR. Finally, ORF length can also change due to alternative splicing.

We modeled the different mechanisms of ORF length change that result due to mutations in start and stop codons, that is, excluding alternative transcription start/end and splicing. To this end we used gain, loss and stationary probabilities of start and stop codons (Iyengar and Bornberg-Bauer 2023). We analysed the length change of ORFs in the range with a minimum of 3 codons (theoretical minimum), and a maximum of 900 codons, which is an unusually large length for *de novo* ORFs. We defined length transition probabilities within this range. Specifically, we defined a transition matrix (M) where the rows and the columns denote ORF length and the elements (M_{ij}, Equation (7)) denote the transition probability ($i \rightarrow j$). Based on the properties of Markov processes, the length transitions over multiple generations (n) can be described by the n^{th} power of the transition matrix (M^n).

Truncation is the Most Likely Outcome of ORF Length Change

We used our Markov model to understand how the length of an existing ORF changes with time. Specifically using the transition matrix (M), we calculated the probability that at a future state F, an ORF remains of the same length (F^0), becomes longer (F^+, to any larger length) or becomes shorter (F^-, to any smaller length). These probabilities that depend

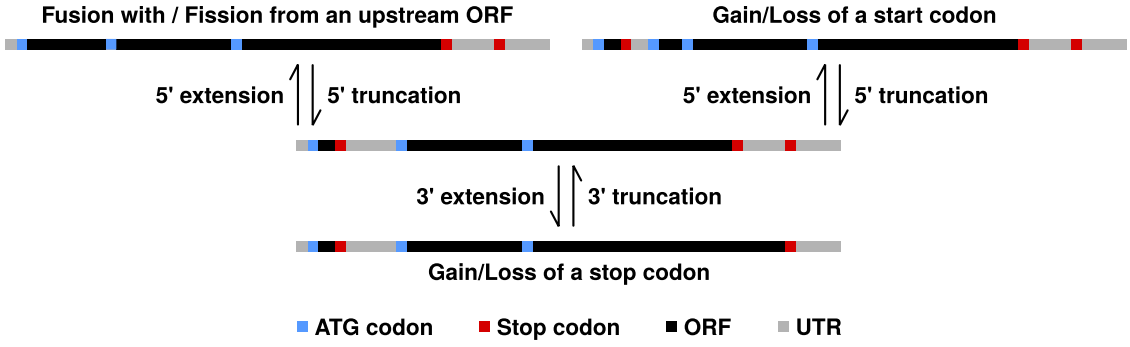

Fig. 1. Schematic of the different mechanisms of ORF length change, that result due to gain and loss of start and stop codons. We have depicted ATG (start) codons, stop codons, ORF body, and UTRs, using blue, red, black, and gray colors, respectively.

on the initial ORF length (i), and the number of generations (n), are described as:

$$F^0(i, n) = M^n_{i,i} \quad (1)$$

$$F^+(i, n) = \sum_{j>i} M^n_{i,j} \quad (2)$$

$$F^-(i, n) = \sum_{j<i} M^n_{i,j}. \quad (3)$$

We found that the probability that ORF length changes in one generation is several orders of magnitude (10^6–10^8) smaller than the probability that it does not. This is understandable because mutations are rare in most organisms (less than 1 mutation in 10^8 base pairs of DNA per generation; Schrider et al. 2013; Zhu et al. 2014; Jee et al. 2016). Accordingly, we found that the length of all ORFs, irrespective of their initial length and nucleotide composition, tends to remain constant even after several generations (data not displayed). This was the case for our probability estimates using the parameters from both *D. melanogaster* and *S. cerevisiae*. Next, we investigated whether truncation is more likely than extension (or *vice versa*), when the ORF length does change. We found that any ORF containing at least 21 codons is more likely to be truncated than extended (Fig. 2). This minimum ORF length is a function of both the number of generations and the nucleotide composition. For example, based on our *D. melanogaster* parameters, ORFs present in a locus with 30% GC-content, and containing at least 11 codons are likely to be truncated in one generation. After 2.7×10^8 generations all ORFs with at least 9 codons are likely to be truncated, irrespective of the nucleotide composition (Fig. 2a). We found similar trends using our *S. cerevisiae* parameters (Fig. 2b). For example, any ORF present in a locus with 30% GC-content and containing at least 11 codons is more likely to be truncated than extended up to $\sim 5.3 \times 10^6$ generations. After 8.6×10^9 generations, all ORFs with at least 9 codons are likely to be truncated, irrespective of the nucleotide composition.

ORFs are More Likely to be a Product of Extension Than of Truncation

We next analysed whether an ORF of a given length originated as an ORF of a different length (shorter or longer) or the same length. To this end, we calculated three probabilities that describe the current state of the ORF (C). First, the probability $C^0(i, n)$ that an ORF of a given length (i) remains of the same length after n generations. Second, the probability $C^+(i, n)$ that any ORF with length $j < i$ extends to an ORF of length i in n generations. Finally, the probability $C^-(i, n)$ that any ORF with length $j > i$ truncates to an ORF of length i in n generations. If $P_{ORF}(i)$ denotes the probability of finding an ORF containing i codons (Iyengar and Bornberg-Bauer 2023), then the three current state ORF probabilities are described as:

$$C^0(i, n) = P_{ORF}(i) M^n_{i,i} \quad (4)$$

$$C^+(i, n) = \sum_{j>i} P_{ORF}(j) M^n_{j,i} \quad (5)$$

$$C^-(i, n) = \sum_{j<i} P_{ORF}(j) M^n_{j,i}. \quad (6)$$

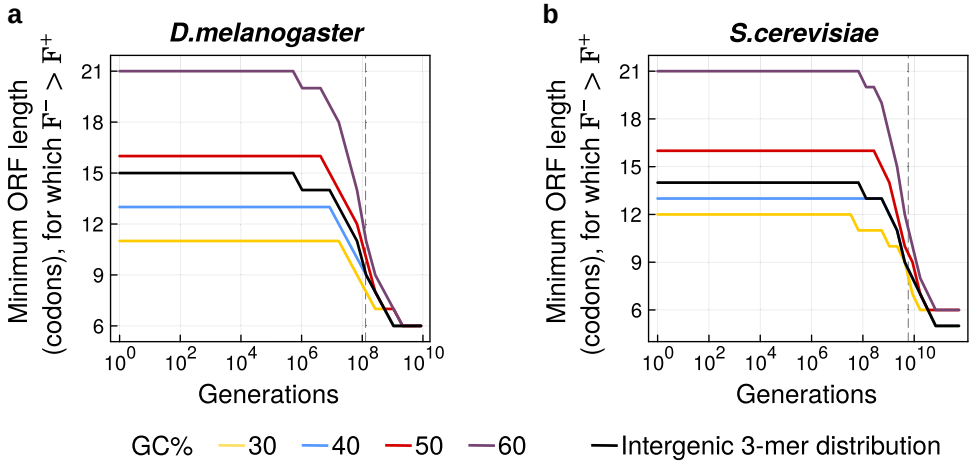

Fig. 2. ORFs tend to become shorter with time than to become longer. Vertical axes show the length of the shortest ORF (in codons) that is more likely to become shorter than to elongate ($F^- > F^+$; Equations (2) and (3)). Horizontal axes show the number of simulated generations of a) *D. melanogaster* and b) *S. cerevisiae*, in log scale. Line colors denote the nucleotide composition of the locus: yellow, 30% GC; blue, 40% GC; red, 50% GC; purple, 60% GC; black, intergenic trimer frequencies. The dashed vertical line in each panel denotes the inverse of the mutation rate of the corresponding organism.

The likelihood of finding an ORF decreases exponentially with its length (Iyengar and Bornberg-Bauer 2023). Therefore, we first asked if it is possible that the scenario where an ORF is extended from smaller ORFs in n generations (C^+) is more likely than the scenario where an ORF's current length is same as what it was n generations ago (C^0). From our previous analyses we know that the probabilities of finding an ORF and of its length change depend on its initial length.

Accordingly, we found that longer ORFs are less likely to have emerged from an ORF of the same length. In contrast, the longer an ORF is, the higher the chances are that it emerged from a shorter ancestor. To illustrate this finding better, we identified the smallest ORF (i) for which $C^+(i, n)$ (Equation (5)) is greater than $C^0(i, n)$ (Equation (4)), at different generations (n). We found that C^0 is greater than C^+ for all the analysed ORF lengths until $\sim 1.4 \times 10^5$ simulated *D. melanogaster* generations (Fig. 3a). From this generation onward the minimum ORF length for which C^+ is greater than C^0, decreases with increasing generations. For example, it decreases from 813 codons at $\sim 1.4 \times 10^5$ generations to 4 codons at $\sim 2 \times 10^9$ generations, for a locus with a GC-content of 30%. We further found that the likelihood of ORF extension relative to that of its length remaining unchanged, decreases with the GC content. For example, after $\sim 8.4 \times 10^6$ generations any ORF with a GC content of 30% and containing at least 92 codons is likely to have been extended. Whereas, this minimum number of codons is 156 for an ORF present in a region with a 60% GC content. We made similar observations using our parameter estimates from *S. cerevisiae* (Fig. 3b). We note that for both the organisms, C^+ overtakes C^0 when the number of generations exceeds 1/1,000 times the inverse of the mutation rate (for example, the inverse of *D. melanogaster* mutation rate is $\sim 1.3 \times 10^8$).

Next, we asked if it is possible that at least some ORFs are more likely to arise due to truncation of larger ORFs than due to extension of smaller ORFs. To this end, we compared the probabilities C^- and C^+ and calculated the smallest ORF length for which C^+ is greater than C^-. We found that only short ORFs are likely to be a product of truncation (Fig. 3c). For example, after one simulated *D. melanogaster* generation, any ORF present in a locus containing 30% GC and a maximum size of 25 codons is likely to have been truncated from larger ORFs. This remains so until $\sim 4.2 \times 10^6$ generations after which even smaller ORFs are more likely to be products of extension. Our analysis using *S. cerevisiae* parameters also revealed similar findings (Fig. 3d). Like in case of *D. melanogaster*, an ORF present in a locus with 30% GC and with a minimum length of 25 codons, is more likely to have been extended from a smaller ORF than truncated from a larger ORF. This minimum length starts decreasing further after $\sim 1.34 \times 10^8$ generations. In both the organisms, the likelihood of an ORF being a product of extension relative to that of truncation, reduces with increasing GC-content.

Finally, we asked if the probability of an ORF originating from a larger ORF (C^-) can be greater than the probability of it originating from an ORF of the same size (C^0). Although less likely, this scenario is indeed possible if the number of generations is large enough. For example, at $\sim 1.3 \times 10^8$ simulated *D. melanogaster* generations, any ORF present in a locus with 30% GC and is longer than 22 codons, is most likely to have been truncated from larger ORFs than to have originated from an ORF of the same length (Fig. 3e). Our analysis with *S. cerevisiae* as a model also produced similar results (Fig. 3f). The number of generations where C^- exceeds C^0 is inversely proportional to the mutation rate. We also found that the likelihood of an ORF to be truncated (C^-), relative to that of it remaining of the same length (C^0), increases with increasing GC content.

Length Changes in *D. melanogaster de novo* ORFs are More Frequent Than Expected by Neutral Evolution

To understand how ORF length changes in actual organisms, we analysed a recently published dataset on *de novo* transcripts in seven inbred *D. melanogaster* lines obtained from seven geographically distinct populations (Grandchamp et al. 2023a, 2023b). These seven lines were predicted to have diverged from a common ancestor ca. 13,000 years ago (Li and Stephan 2006; Laurent et al. 2011; Grandchamp et al. 2023b). This corresponds to approximately 3,33,400 generations (generation time of 2 weeks; Fernández-Moreno et al. 2007). Next, we identified all possible ORFs (≥30nt) in the novel transcripts from each of the seven lines, and sorted them into groups of orthologous ORF sequences (orthogroups) based on sequence homology and synteny (that is, identical flanking genes). We also identified untranscribed ORF orthologs by analysing the genomic regions syntenic to those of transcribed orthologs, and included them in the orthogroups. We thus constructed 758 orthogroups of which 48% had orthologous ORFs from all seven lines (Fig. 4a). Next, we analysed the lengths of the ORFs within each orthogroup. Specifically, we identified orthogroups in which all the constituent ORFs had identical lengths, and those that had ORFs with different lengths. Most orthogroups (77%) did not have any length variation between their constituent ORFs (Fig. 4a). This observation only qualitatively agrees with our model, because the percentage of orthogroups with length variation (23%) between their constituent ORFs, was significantly higher than expected (~4.8%; $P < 10^{-6}$, Monte–Carlo sampling). Possible reasons for frequent ORF length changes could be variability of

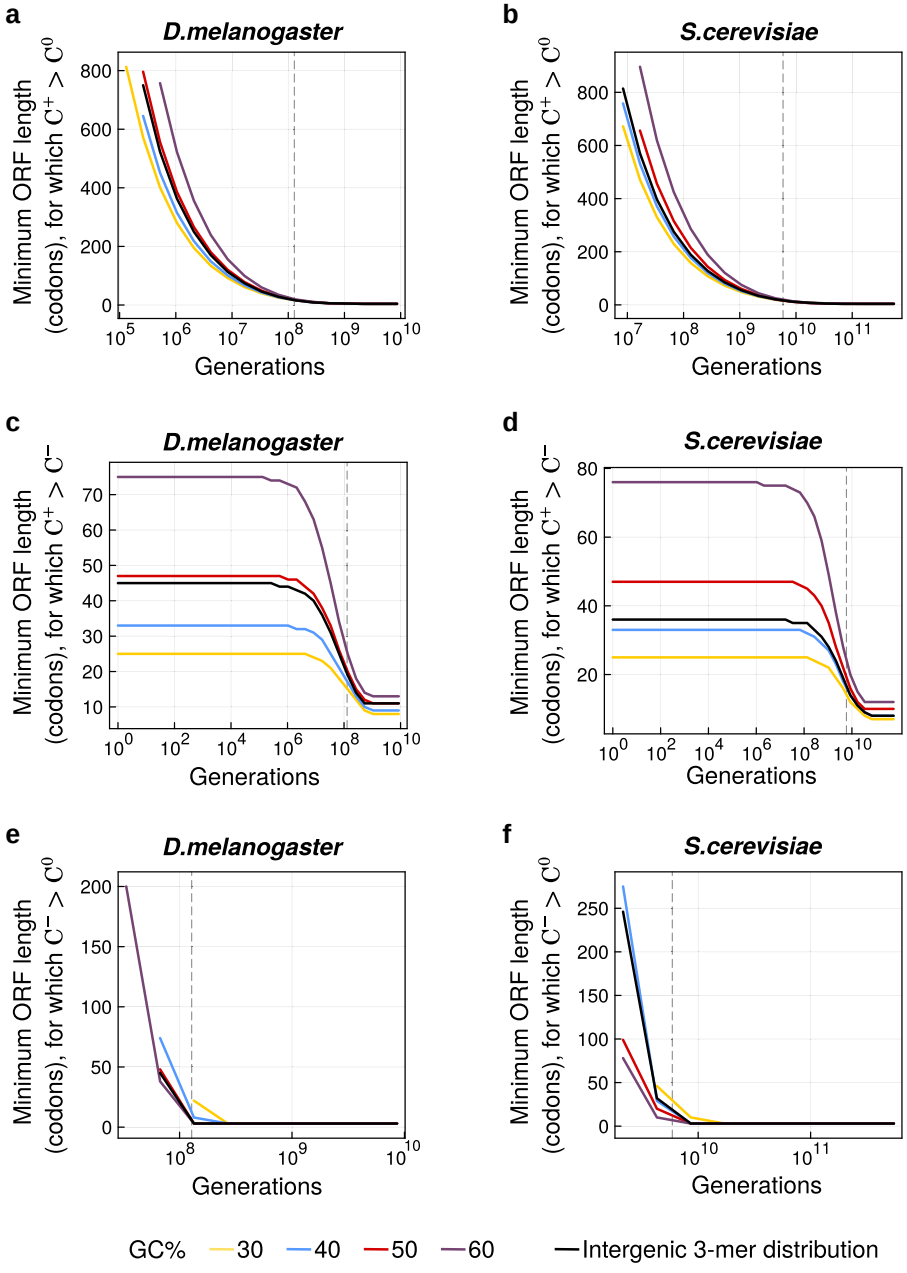

Fig. 3. Likelihoods of ORF length change when the ancestral length of the ORF is unknown. The shortest ORF (vertical axis) in a) *D. melanogaster* and b) *S. cerevisiae* that is more likely to have descended from any shorter ORF than from an ORF of the same length ($C^+ > C^0$). The shortest ORF (vertical axis) in c) *D. melanogaster* and d) *S. cerevisiae* that is more likely to have descended from any shorter ORF than from any longer ORF ($C^+ > C^-$). The shortest ORF (vertical axis) in e) *D. melanogaster* and f) *S. cerevisiae* that is more likely to have descended from any longer ORF than from an ORF of the same length ($C^- > C^0$). See Equations (4)–(6) for details. Horizontal axes in all panels show the number of simulated generations in log scale. Line colors denote the nucleotide composition of the locus: yellow, 30% GC; blue, 40% GC; red, 50% GC; purple, 60% GC; black, intergenic trimer frequencies. The dashed vertical line in each panel denotes the inverse of the mutation rate of the corresponding organism.

transcription start sites, which we did not incorporate in the model (but have explored in more detail in the following sections).

A few studies analysing *de novo* genes across long evolutionary time scales comprising speciation events, report that *de novo* ORFs become longer with evolutionary age (Carvunis et al. 2012; Neme and Tautz 2013; Zhao et al. 2014; Vakirlis et al. 2017; Dowling et al. 2020; Heames et al. 2020; Blevins et al. 2021; Middendorf and Eicholt 2024). In contrast, our model suggests that an ORF is more likely to become shorter than to become longer with age (Fig. 2). To further understand whether truncation

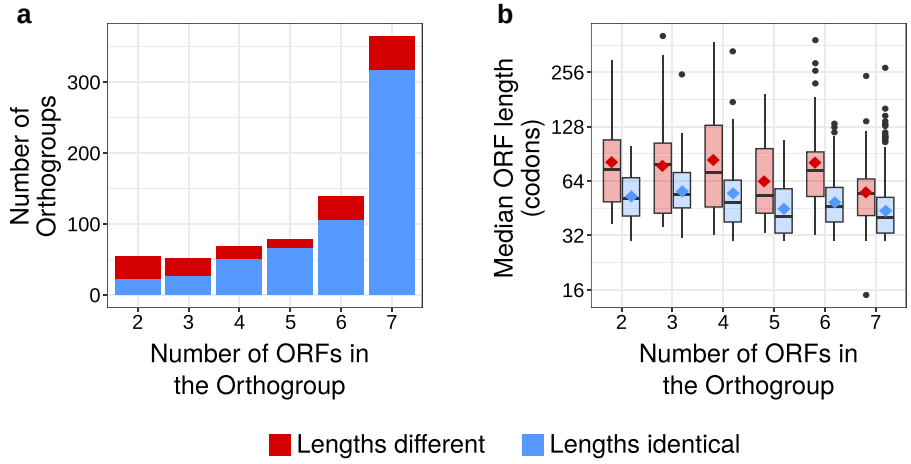

Fig. 4. Distribution of *de novo* ORFs in the seven different *D. melanogaster* populations. a) Vertical axis shows the number of orthogroups and horizontal axis shows the size of the orthogroups in the number of orthologous ORFs they contain (not more than one per population; vertical axis). b) Standard boxplots depicting the median length of the ORFs (vertical axis, $\log_2$ scaled) in the orthogroups of each size class (horizontal axis). In both the panels, orthogroups containing ORFs of identical length and those with at least two ORFs of different lengths are denoted using red and blue colors, respectively.

is more common than extension, we analysed our orthogroups that contain ORFs with different lengths. Orthogroups that contain more orthologous ORFs, may be evolutionarily older than those that contain fewer ORFs. This is especially evident for orthogroups that contain ORFs from all seven lines including the outgroup. Therefore, we analysed the correlation between the number of orthologous ORFs within an orthogroup and the median length of these ORFs. We did so for both the orthogroups with length variation and those without length variation. We found a significant negative correlation between median ORF length and the number of lines harboring an ORF ortholog (Fig. 4b; orthogroups with length variation: Spearman's $\rho = -0.183$, $P = 7.8 \times 10^{-3}$; orthogroups without length variation: Spearman's $\rho = -0.186$, $P = 2.89 \times 10^{-6}$). This analysis does not answer the question whether older orthogroups have shorter ORFs. Rather, it suggests that shorter ORFs are retained more frequently. We also performed a correlation analysis between median ORF length and the age of the orthogroup estimated from a dated phylogenetic tree of the *D. melanogaster* populations (Grandchamp et al. 2023b). We found that orthogroup age is significantly negatively correlated with median ORF length (Spearman $\rho = -0.19$, $P = 7 \times 10^{-8}$). However, this correlation is less statistically significant if we restrict our analysis to the orthogroups with ORFs of different lengths (Spearman $\rho = -0.11$, $P = 0.076$).

To more precisely understand if ORFs become shorter with time, we estimated the length of the ancestral ORF for each orthogroup that contains ORFs of different lengths. Specifically, we used the dated phylogenetic tree (Grandchamp et al. 2023b) to find out the most recent common ancestor that would have harbored an ORF belonging to a specific orthogroup. For example, if an orthogroup contains ORFs from the Swedish line and the Danish line, then the last common ancestor of both these lines must also contain a homologous ORF. We assign this ORF as the ancestor of the orthogroup. We note that we do not perform ancestral sequence reconstruction but simply assume that an ancestral ORF exists and it can have any possible length. Next, we calculate the probability that the ORFs in an orthogroup could have the length they have, given the phylogeny between the populations and that the ancestral ORF has a specific length. Specifically, we calculate the transition probability than an ancestral ORF of a length (i) gives rise to an extant ORF of length (j), in the number of generations estimated from the length of the evolutionary path that connects the ancestral population and the extant population (that contains the ORF). We perform the same calculation for every other ORF in the orthogroup, while excluding the branches of the tree that have been already counted. The multiplicative product of these different transition probability values indicates the likelihood of an ancestral ORF length, such that the most likely ancestral ORF length will produce the largest value. Using this technique, we predicted the most likely ancestral ORF length for every orthogroup, and in turn, the frequencies of truncations and extensions. We found that 70.5% ORFs had the same length as their ancestor, 25% had been truncated from a longer ancestor, and 4.5% had been extended from a shorter ancestor. We emphasize that these results are based on the assumption of evolutionary neutrality. This analysis helps quantify the expected frequencies of length changes based on the model. It is not an independent assessment of the model's prediction. Therefore, we also used a maximum parsimony approach to predict ancestral ORF length such that the ancestral length is same as the most frequent length within the

orthogroup. This analysis is only possible for orthogroups where at least two ORFs have identical lengths (135 out of 141 orthogroups). We found that the ancestral ORF length predicted by the maximum parsimony approach agrees with our model based "maximum-likelihood" estimate in 106/135 cases. In 24 out of the remaining 29 cases, the maximum parsimony approach predicted a longer ancestor than the maximum-likelihood estimate. Overall, the maximum parsimony method predicted a higher proportion of truncations, than the maximum-likelihood method. Specifically, in the 135 orthogroups where maximum parsimony analysis was possible, the method predicted that 32.2% of the ORFs were truncated from a longer ancestor, 1.9% were extended from a shorter ancestor, and 65.9% had the same length as the ancestor. The corresponding estimates from the maximum-likelihood method were 21.8%, 5%, and 73.2%, respectively. Thus, an orthogonal analysis of ancestral ORF length supports our model based estimates that truncations are more likely than extensions.

Overall, our analyses of *D. melanogaster de novo* ORFs suggest that truncation is more likely than extension.

Length Changes in *D. melanogaster de novo* ORFs are Larger When They Occur at the 5′ End Than from the 3′ end

We next focused on the mechanism of ORF length changes. We could expect 3′ extensions to be smaller in magnitude than 5′ extensions because it is more likely to find one of the three stop codons by chance than to find a start codon. However, unlike the loss of stop codons in 3′ extensions, a gain of a new start codon during 5′ extension does not require the loss of the downstream (old) start codon. Furthermore, the mechanisms and probabilities of truncation are different from that of extension (Fig. 2, Equations (2) and (3)). Because the differences in the mechanisms of 5′ and 3′ length changes are not trivial, we analysed both our model and the *D. melanogaster* data, to understand which change is more frequent and higher in magnitude.

To this end, we first identified the longest ORF in an orthogroup that has ORFs of different lengths, and aligned it to all the shorter ORFs in the same orthogroup, using protein BLAST (Altschul et al. 1990; Camacho et al. 2009). Using the alignment, we determined whether the longest ORF is extended from the 5′ end or the 3′ end (or both), relative to the shorter ORFs. We analysed 229 ORF pairs, out of which 142 pairs shared the same start position (62%, 3′ change), 81 shared the same stop position (35.4%, 5′ change), and 6 shared neither of the two termini (2.6%, changed form both the ends). Next, we compared the extent of length changes from the 5′ and the 3′ ends, and found that 5′ changes (median 48 codons) were larger than the 3′ changes (median 21 codons; one-sided Mann–Whitney U-test, $P = 4.6 \times 10^{-4}$; (Fig. 5a). This is in qualitative agreement with our model's prediction that also shows that large changes are more likely to occur in the 5′ end than from the 3′ end (Fig. 5b). Overall, we found that 3′ changes were more frequent but produced smaller length difference between the ORFs.

ORF length could also be altered due to changes in transcription start site. For example, a more upstream transcription start site could cause an inclusion of an in frame start codon, that in turn could extend the ORF from the 5′ end. Because we constructed our orthogroups by first identifying transcribed ORFs, ORF length differences between two orthologous ORFs could exist due to differences in their transcription start sites (TSS). Therefore, we repeated our previous analyses of ORF start and end sites (previous para), with only the transcribed ORFs. We now analysed 130 ORF pairs, and found that 44 were altered from the 3′ end, 80 were altered from the 5′ end and 6 were altered from both the ends. Furthermore, we found that 5′ changes, that are now more frequent, also produced a greater extent of ORF length change (one-sided Mann–Whitney test, $P = 2.4 \times 10^{-6}$ (Fig. 5a).

Next, we analysed the mechanism underlying the ORF length changes. Length changes at the 3′ end could result from gain or loss of a stop codon. To verify if this is the case, we analysed the codon in the longest ORF that overlaps with the stop codons of smaller ORFs in the same orthogroup. Specifically, we asked if this codon is a single-nucleotide mutation away from a stop codon (stop neighbor). We performed this analysis with both our full dataset containing all ORFs and the dataset containing only the transcribed ORFs. For both datasets we found that a stop neighbor was present in the longest ORF in ∼52% of cases (52.1%: all, 52.3%: transcribed only; Fig. 5c). Because it is unlikely for a stop codon to mutate into a non-stop codon with more than 1 nucleotide substitutions ($\sim 10^{-17}$), one would expect nearly every codon in the long ORF that overlaps with a stop codon in the short ORF, to be a stop neighbor. Our data analysis suggests that a large proportion of 3′ changes do not occur due to a simple gain or loss of a stop codon. We performed an analogous analysis where we identified the codon in the longest ORF that overlaps with the start codon in the shorter ORFs. We found that in every case (and in both the datasets), the longest ORFs indeed had an ATG (∼99%) or a 1nt-neighbor of ATG (start neighbor, ∼1%) at the position overlapping the start codon in the smaller ORFs (Fig. 5c), in accordance with our expectation.

A truncation from the 5′ end could result from either the loss of a start codon (with the availability of a downstream in-frame start codon) or the gain of a stop codon in between two in frame start codons. Furthermore, transcript truncation from the 5′ end (alternative TSS) could also cause ORF truncation. To assess these diverse possibilities, we analysed the dataset containing only the transcribed

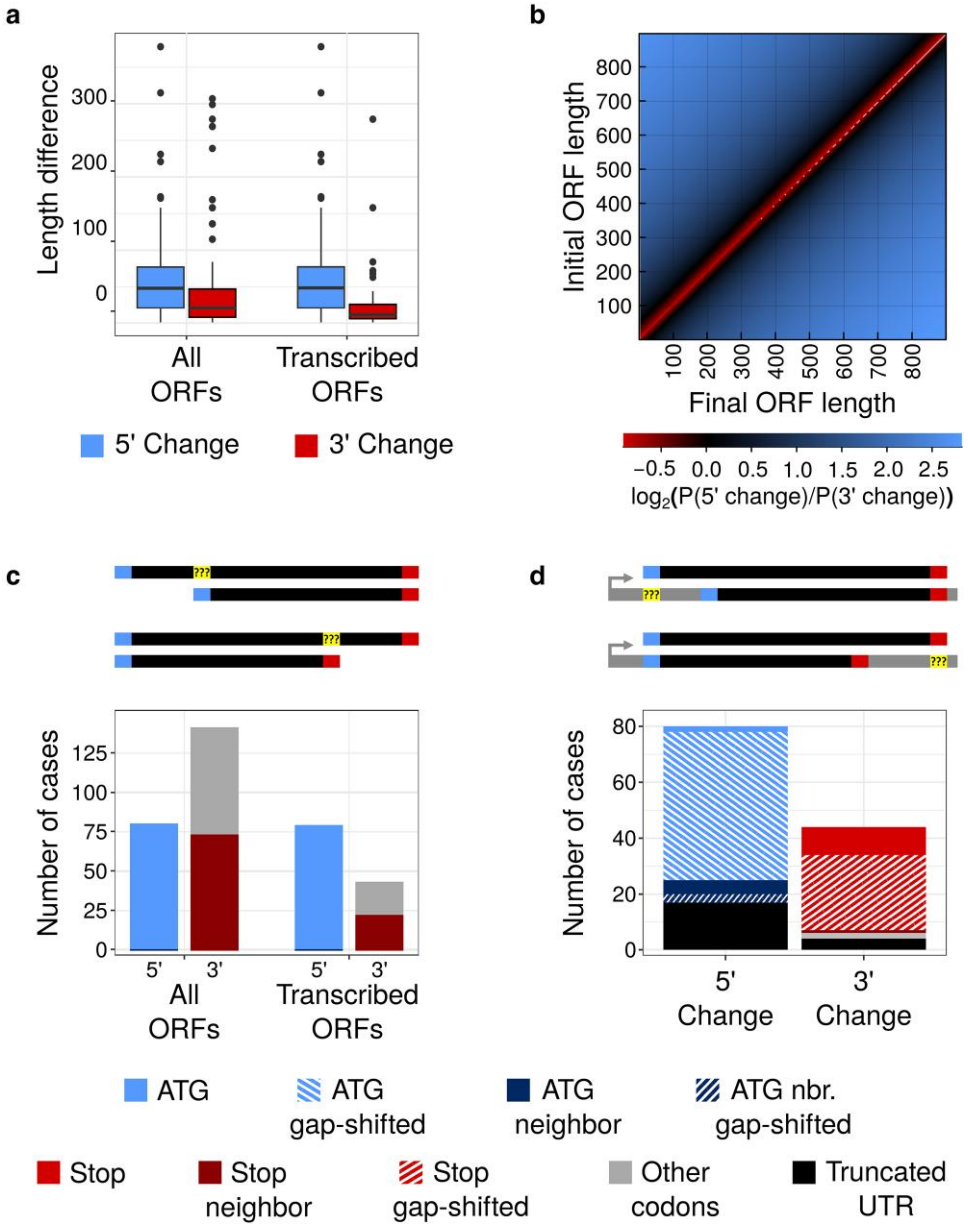

Fig. 5. Analyses of ORF length changes from the 5′ end and the 3′ end. a) Standard boxplots depicting ORF length changes (vertical axis) from the 5′ end (blue boxes) and the 3′ end (red boxes) in the orthogroup dataset constructed using all ORFs as well as the orthogroup dataset containing only the transcribed ORFs (horizontal axis) b) Heatmap depicting the $\log_2$ transformed ratio of the probability of a 5′ length change relative to that of a 3′ length change (color scale) for different values of initial (vertical axis) and final (horizontal axis) ORF lengths. Blue shades denote probability of 5′ change being greater than that of 3′ change, while red shades denote *vice versa*. See Equations (8)–(9). c) Number of cases in both the datasets (all ORFs and transcribed ORFs, groups in horizontal axis) where the codons of each category in the longest ORF (colors: blue, ATG; dark blue, start neighbor; dark red, stop neighbor; grey, other codons), map to the terminal codons (5′ or 3′, horizontal axis) of the shorter ORFs. d) Number of cases in the datasets with transcribed ORFs where the codons of each category in the UTR of shorter ORFs (same color scheme as c) that map to the terminal codons (5′ or 3′, horizontal axis) of the longest ORFs. Black bar denotes the number of cases where the UTR of the short ORF is truncated short of the terminal codons of the longest ORF. Hatched bars denote the cases where a codon in the UTR of the short ORF aligns with the terminal codons of the longest ORF, but with gaps in the alignment (colors: blue, ATG; dark blue, start neighbor; red, stop).

ORFs, and identified the codon in the 5′UTRs of the short ORFs in an orthogroup that map to the start codon in the longest ORF of the same orthogroup. In 2 cases out of 80 (2.5%), an ATG in the 5′ UTR of the short ORFs mapped to the position of the start codon in the longest ORF (Fig. 5d). This suggests that there could have been a gain of stop codon between this upstream ATG and the actual start codon. In 5 cases (6.25%), a start neighbor in the

5′UTR mapped to the position of the start codon of the longest ORF, suggesting that a loss of a start codon could have caused the length change (Fig. 5d). We further explored the cases where no start codons or neighbors could be mapped to the start codon of the longest ORF. In 17 cases (21.25%), we found the transcript of the short ORF to be truncated from the 5′ end such that no 5′UTR positions existed that could overlap with the start codon of the longest ORF. Still, in a large majority of the cases (71.25%), the overlapping codon in the 5′UTR was not a stop codon or a stop neighbor. We suspected that this could occur because of insertions/deletions (indels). To investigate this, we aligned the nucleotide sequences of the 5′UTRs of the short ORFs with the longest ORF using MAFFT (Katoh et al. 2002). We identified that in all these 57 cases, either an ATG (53 cases, 66.25%) or a start neighbor (3 cases, 3.75%) in the 5′UTR aligned with the start codon of the longest ORF but with gaps in the alignment. This shows that a large majority of 5′ length changes occurred due to indels.

We analysed the 3′ changes (44 cases) analogously, and found that a stop codon in the 3′UTR of the short ORF aligned (MAFFT) with the stop codon of the long ORF, in 37 cases (84.1%). Of these, 10 cases (22.7%) contained no gaps in the alignment. In 1 case, the overlapping codon was a start neighbor and in 2 cases it was a different codon. In the remaining 4 cases (9.1%), the stop codon in the longest ORF was located farther than the end of the transcript containing the short ORF (Fig. 5d).

Overall, our analyses show that ORF length tends to change more from the 5′ end than the 3′ end, and that a large proportion these changes from occur due to indels and alternative transcript termini.

Discussion

In this study, we aimed to understand how ORF length changes during the course of evolution. To this end, we first developed a mathematical model that predicts ORF length changes under the assumption of evolutionary neutrality (Equation (7)). We used the model to ask two questions. First we asked, if given that an ORF of a certain length exists, what the chances are that it retains its length, becomes shorter or becomes longer (Equations (1)–(3)). ORFs are more likely to retain their length than to become longer or shorter. This is an expected outcome because mutations are rare in most organisms including the two used in our modeling analyses (Schrider et al. 2013; Zhu et al. 2014). However, we found that when ORF length does change then it is more likely to decrease than increase (Fig. 2). Our second question pertains to the evolutionary past of the ORF. Specifically, we asked if ORFs originate from ancestral sequences that are shorter, longer, or of the same size (Equations (4)–(6)). Our model shows that ORFs are most likely to arise from ancestral ORFs of identical length in short evolutionary timescales. However, the likelihood of an ORF originating from an ORF of a different length becomes higher with time (generations). ORFs are especially likely to elongate from shorter ancestors than from longer ancestors (Fig. 3). This finding may seem contradictory to that of our first question where we find that ORFs are more likely to become shorter with time. The difference hinges mainly on what we know about an ORF. When we predict the future of the ORF, we are sure about its current length and estimate an average of all possible future outcomes (extensions or truncations). However, when we predict the past states of an ORF, we are unsure about the initial ORF length and assume that they are geometrically distributed such that longer ORFs are exponentially less likely to exist (Iyengar and Bornberg-Bauer 2023). Longer ORFs that are more likely to be truncated, are less likely to exist by chance. Conversely, shorter ORFs, that are more likely to extend, are more likely to exist by chance. Thus on an average, ORFs are more likely to extend from shorter ancestors.

To validate some of these predictions, we analysed length changes in orthologous *de novo* ORFs from seven *D. melanogaster* lines obtained from seven different populations. We found that most orthogroups contained ORFs of identical length, which is qualitatively in line with our model's predictions. However, we found more orthogroups with length changes than expected. One possible reason could be that our model ignores mutational events that can change transcription start sites (TSS) and splice sites, that can be important determinants of ORF length change. Indeed, our subsequent analysis shows that in 20% of ORFs with a length change from the 5′, occurs due changes in the TSS in the corresponding transcripts (Fig. 5d). The variations in TSS could result from biological mechanisms such as noisy transcription initiation, or from technical artifacts of transcriptome assembly that could arise due to low sequencing depth. We also analysed the orthogroups to ask if ORFs indeed become truncated with time. To address this question we first used a simple approach where we compared the median ORF length of an orthogroup to the number of ORFs contained in it. We found that orthogroups with more ORFs had a smaller median length suggesting that short ORFs are more likely to survive (Fig. 4b). The orthogroups with fewer ORFs need not necessarily have a more recent ancestor than the orthogroups containing ORFs from all the seven lines, possibly because of widespread ORF loss. Thus widespread survival of short ORFs alone does not suffice to answer if they originated from a longer ancestor. Therefore, we used a more refined approach where we inferred the most likely ancestral length of an orthogroup using our mathematical model and a dated phylogenetic tree of the seven *D. melanogaster* populations (Grandchamp et al. 2023b). Using this approach, we quantified truncations that are more likely than extensions.

Our findings stand in contrast to those of several studies that report that ORFs of older *de novo* protein coding genes are longer (Carvunis et al. 2012; Neme and Tautz 2013; Zhao et al. 2014; Vakirlis et al. 2017; Dowling et al. 2020; Heames et al. 2020; Blevins et al. 2021; Middendorf and Eicholt 2024). Our model is based on the assumption of neutral evolution, and the *de novo* ORFs we analysed are very young (most of them not even fixed in the species). In contrast, the above mentioned studies consider single genomes from many species that have diverged several millions of years ago. Therefore, evolutionary selection is a very likely explanation for the larger length of old *de novo* ORFs found in these studies, that our analyses do not take into account. Some *de novo* genes may become fixed in a species (or a clade) due to positive selection. Indeed, some *de novo* genes have been experimentally shown to increase organismal fitness (Baalsrud et al. 2017; Gubala et al. 2017; Lange et al. 2021; Wacholder et al. 2023). Purifying selection on functional *de novo* genes may prevent the truncation of their ORFs. On the other hand, ORF extension may increase the beneficial fitness effects of a gene which can lead to its fixation. For example, C-terminal (3′) extensions that are structurally disordered could protect proteins from deleterious effects of stop codon readthrough (Kleppe and Bornberg-Bauer 2018). An evolutionary stable gene, especially one that produces a thermodynamically stable protein, can also facilitate ORF extension (Bornberg-Bauer and Chan 1999; Bloom et al. 2006; Zheng et al. 2020). This is possible because protein stability can facilitate evolutionary innovation by dampening the destabilizing effects of mutations including ORF extensions. In the absence of this buffering, these mutants (or extended proteins) can misfold, and can be eliminated by negative selection.

Our study, like every other study, is based on certain approximations and is limited by availability of data. Therefore, one must consider these limitations before applying the findings to a larger scale, and for making broad generalizations. For example, we use GC-content and trimer frequencies to approximate nucleotide distributions. Both these parameters may vary significantly throughout the genome. However, using locus-specific nucleotide distributions can predict evolution of the corresponding loci more realistically. Our model also does not take into account all the possible mechanisms that lead to ORF length change. For example, we do not incorporate changes in transcription start site (TSS), primarily because we do not have enough data to model the evolution of TSS. When such data become available the model can be updated. We also do not model mutational mechanisms other than substitutions, such as indels and transpositions. Inclusion of small indels in the model (based on indel rates from Schrider et al., 2014 and Zhu et al., 2013) only produces modest quantitative changes in the results (data not shown). Our analysis of ORF sequences is also limited by the amount of available information. For example, we cannot estimate the exact divergence time between the *D. melanogaster* populations. Moreover, the generation time is unlikely to be similar between populations living in vastly different climatic regions. Furthermore, the populations may not be strictly geographically isolated, as suggested by small genetic variation (F_{ST}) between the European populations (Kapun et al. 2020).

Despite many simplifications in our analyses, our study provides many qualitative answers, and also a formal basis to test hypotheses through more focused experiments. It also opens up several questions that may be a topic of future research. For example, it may be worth investigating if terminal extensions can indeed improve the fitness effect of a protein or if they are mostly deleterious or neutral. Ultimately, studying the ancestry of different extant *de novo* ORFs could help understand if they were extended from smaller ancestors or if they were born with the same length, and could help dissect the different mechanisms that lead to fixation of *de novo* genes with long ORFs.

Methods

Calculation of Transition Probability

We calculated the probability of ORF length change due to gain and loss of stop codons (Fig. 1), in the form of a transition matrix (M) whose rows denote the initial ORF length and the columns denote the final ORF length. We generated transition matrices for two organisms—*D. melanogaster* and *S. cerevisiae*. For each organism we calculated the transition probabilities using the values of biased mutation rate (Table 1), and nucleotide composition described by four different values of GC-content (30%, 40%, 50%, and 60%), as well as the distribution of DNA trimers in the intergenic regions of the organism. We

Table 1 Mutation bias probability values for **(A)** *D. melanogaster* (mutation rate = 7.8×10^{-8}, Schrider et al. 2013) and **(B)** *S. cerevisiae* (mutation rate = 1.7×10^{-10}, Zhu et al. 2014)

(A) *D. melanogaster*

	A	T	G	C
A	…	0.056	0.243	0.075
T	0.056	…	0.075	0.243
G	0.438	0.075	…	0.113
C	0.075	0.438	0.113	…

(B) *S. cerevisiae*

	A	T	G	C
A	…	0.063	0.144	0.110
T	0.063	…	0.110	0.144
G	0.349	0.182	…	0.152
C	0.182	0.349	0.152	…

Table 2 Description of the probability terms used in Equation (7). Here we describe the probabilities associated with stop codons (based on Iyengar and Bornberg-Bauer 2023). Analogous probability terms for a start codon are denoted by the subscript, *ATG* (instead of *stop*)

Term	Description	Derivation
P_{stop}	Finding a stop codon	…
$P_{stop-gain}$	Gaining a stop codon	…
$P^*_{stop-gain}$	Gaining a stop codon given it does not exist	$P^*_{stop-gain} = P_{stop-gain}/(1 - P_{stop})$
$P_{stop-loss}$	Losing a stop codon given it exists	…
$P_{stop-stay}$	Not losing an existing stop codon	$1 - P_{stop-loss}$
$P_{no-stop-stay}$	No stop codon exists and none is gained	$1 - P_{stop-gain} - P_{stop}$

calculated the values of the elements of the transition matrix (M_{ij}, Equation (7)) using probabilities of finding, gaining and losing start and stop codons (Table 2; Iyengar and Bornberg-Bauer 2023). Specifically, in any one generation an ORF's length remains unchanged if there is no loss of start ($P_{ATG-stay}$) and stop ($P_{stop-stay}$) codons, and no gain of premature stop codons ($1 - P^*_{stop-gain}$).

An ORF's length increases from the 3′ end when there is a loss the stop codon ($P_{stop-loss}$), a new stop codon is located in the 3′ UTR (P_{stop}). The length can increase from the 5′ end when a new upstream start codon is gained ($P_{ATG-gain}$). A 5′ extension can also occur when the ORF fuses with an in-frame upstream ORF when the latter loses its stop codon. In all these three mechanisms of extension there should be no gain of a premature stop codon in the coding region of the original ORF ($(1 - P^*_{stop-gain})^{i-2}$), and no stop codons should be found or gained in the extended part of the ORF ($(P_{no-stop-stay})^{j-i-1}$).

An ORF truncation occurs when there is a gain of a premature stop codon ($P^*_{stop-gain}$). This produces a 3′ truncation by default, but can also cause a 5′ truncation if a downstream start codon can be found (P_{ATG}). A 5′ truncation can also result due to loss of the start codon ($P_{ATG-loss}$) and if the ORF sequence contains another downstream start codon (P_{ATG}). In all these three mechanisms, no premature stop codon must emerge in the truncated ORF ($(1 - P^*_{stop-gain})^{j-2}$). The length transition probabilities due to all these mechanisms are given by the following equation (Equation (7)).

$$M_{ij} = \begin{cases} P_{ATG-stay}P_{stop-stay}(1 - P^*_{stop-gain})^{i-2} & \text{if } i = j \\ (P_{no-stop-stay})^{j-i-1}(1 - P^*_{stop-gain})^{i-2} \times \begin{pmatrix} P_{ATG-stay}P_{stop-loss}P_{stop} \\ +P_{ATG-gain}P_{stop-stay} \\ +(j-i-2)\dfrac{P_{stop-loss}P_{stop}P_{ATG}}{P_{no-stop-stay}} \end{pmatrix} & \text{if } i < j \\ (1 - P^*_{stop-gain})^{j-2} \times \begin{pmatrix} P_{ATG-stay}P^*_{stop-gain} \\ +P_{ATG}P_{ATG-loss}(P_{no-ATG-stay})^{i-j-1} \\ +(i-j-1)P^*_{stop-gain}P_{stop-stay}P_{ATG} \end{pmatrix} & \text{if } i > j \end{cases} \quad (7)$$

By excluding some terms from Equation (7), we calculated the probability of length transitions from only the 5′ end (Equation (8)) or only the 3′ end (Equation (9)).

$$M_{5:ij} = \begin{cases} (P_{no-stop-stay})^{j-i-1}(1 - P^*_{stop-gain})^{i-2} \times \begin{pmatrix} P_{ATG-gain}P_{stop-stay} \\ +(j-i-2)\dfrac{P_{stop-loss}P_{stop}P_{ATG}}{P_{no-stop-stay}} \end{pmatrix} & \text{if } i < j \\ (1 - P^*_{stop-gain})^{j-2} \times \begin{pmatrix} P_{ATG}P_{ATG-loss}(P_{no-ATG-stay})^{i-j-1} \\ +(i-j-1)P_{stop-gain}P_{stop-stay}P_{ATG} \end{pmatrix} & \text{if } i > j \end{cases} \quad (8)$$

$$M_{3:ij} = \begin{cases} (P_{no-stop-stay})^{j-i-1}(1 - P^*_{stop-gain})^{i-2}\ P_{ATG-stay}P_{stop-loss}P_{stop} & \text{if } i < j \\ (1 - P^*_{stop-gain})^{j-2}\ P_{ATG-stay}P^*_{stop-gain} & \text{if } i > j \end{cases} \quad (9)$$

Construction of ORF Orthogroups Containing Transcribed ORFs

We identified ORFs in the novel intergenic transcripts originating in seven *D. melanogaster* inbred lines obtained from seven different geographically distinct populations (Lebherz et al. 2024). Specifically, we identified all ORFs starting with the canonical start codon (ATG), and containing a total of at least 10 codons (30nt), using the program *getorf* (Rice et al. 2000).

Next, we grouped these ORFs into orthogroups. To this end, we used protein BLAST (Altschul et al. 1990; Camacho et al. 2009) to align ORFs from all the seven lines with each other, at 100% query coverage, 90% sequence identity and an E-value cutoff of 10^{-4}. We grouped all the BLAST hits into orthogroups such that every ORF in each orthogroup aligned with at least one another ORF in the same orthogroup (based on our hit cutoffs). We discarded orthogroups that contained more than one ORF per line. Next, we analysed the genes flanking the ORF of an orthogroup (synteny), using genome annotations for the seven lines (Grandchamp et al. 2023a). If the neighboring genes were identical for all ORFs, we kept them in the same orthogroup, and split them into multiple orthogroups, otherwise. We applied this method to every orthogroup. We further analysed only the orthogroups that contained at least two ORFs.

Identification of Non Transcribed Homologous ORFs

Because our study focuses primarily on ORF length extension but not gain and loss of transcription, we included non-transcribed ORFs in our orthogroups, that were homologous and syntenic to the transcribed ORFs the same orthogroup. To this end we adapted the analytical pipeline from Grandchamp et al. (2023a). Specifically, we first identified the flanking genes of an orthogroup. If an orthogroup did not contain an ORF from a certain line, we extracted the genomic region of the line that lay between the same two flanking genes as the ORFs in the orthogroup. Next, we aligned this region (subject) to the longest ORF in the orthogroup (query) using nucleotide BLAST with a cutoff of 60% query coverage per highest scoring pair (hsp). Next, we identified the highest scoring hit, verified if it is an ORF, translated it into a protein sequence, and mapped it with the longest ORF in the orthogroup using BLASTP (with the same criteria as that we used for transcribed ORFs). If the sequence was successfully mapped with BLASTP, we included it in the orthogroup. Using this method we identified the missing non-transcribed ORFs in every orthogroup that contained less than seven transcribed orthologous ORFs.

Predicting the Ancestral ORF Length

To estimate the length of the ancestral ORF for an orthogroup, we used a dated phylogenetic tree of the seven *D. melanogaster* populations (Grandchamp et al. 2023b), and the transition matrix estimated using the mutation rate and the intergenic trimer distribution of *D. melanogaster*. We estimate the evolutionary distances in generations from the length of the phylogenetic tree's branches, assuming that one year has 26 generations (Fernández-Moreno et al. 2007). To explain our method better, we use the example of a hypothetical orthogroup that contains ORFs from the Swedish, the Danish and the Spanish populations (leaf nodes), with ORF lengths 50, 50, and 70, respectively. We call the common ancestor of all the three populations "European", and the a common ancestor of the last two populations "Scandinavian". Next, we calculate the number of generations between the European node and each of the three leaf nodes, and use this number to calculate multi-generational transition matrices (M^n). For instance, we start this calculation with the Swedish node (with 254,462 generations). Next, we repeat the calculation for the Danish node, but we only count number of generations between the Scandinavian node and the Danish node (161,304 generations) because the evolutionary timespan between the European and the Scandinavian nodes (93,158 generations) is already calculated for the Swedish node. Finally, we calculate the transition matrix for the evolutionary timespan between the European and the Spanish nodes (254,462 generations).

An ancestral ORF can theoretically have any possible length. However, we do not assume a length for the ancestral ORF that is not identical to that of any ORF in the orthogroup. That is because probabilities of extension and truncation are less likely than the probability that ORF length is remains the same, in the evolutionary divergence times between the different *D. melanogaster* populations(Equations (4)–(6)). Our consideration minimizes the total number of length changes and excludes unlikely ancestral ORF lengths. Therefore in our hypothetical example, we test ancestral lengths of 50 and 70 and ask which one of them is more likely. First, we calculate the probability that the lengths of the Swedish, the Danish and the Spanish ORFs are 50, 50, and 70, respectively, given that the ancestral length is 50. This would be the product of the corresponding elements of the three transition matrices (as explained in the previous para). We repeat this calculation with now assuming an ancestral ORF length of 70. The ancestral ORF length that gives the largest probability is the most likely ancestor (in this case 50).

We apply the same technique to all the orthogroups that have ORFs with different lengths. Based on the predicted ancestral lengths, we calculate the expected number of ORF length changes (truncation, extension, or no change) relative to the ancestor.

We also applied a maximum parsimony approach to support the model-based maximum-likelihood estimates. We assigned the most frequent ORF length within an orthogroup as its ancestral length whenever that was possible (141 out

of 172). This rule does not work when the orthogroup has only two ORFs or if more than one distinct ORF length is equally frequent (for example: 65, 65, 50, 50, 40). We found 38 such orthogroups out of which 6 had more than two ORFs. For the 6 orthogroups with more than two ORFs, we identified the outgroup ORF using the population phylogenetic tree. We assigned the length of the outgroup ORF as the ancestral length of the orthogroup. It is possible that no outgroup could be identified but in our dataset we did not encounter such cases.

Predicting the Expected Number of Orthogroups with ORFs of Different Length Using Monte–Carlo Sampling

With the predicted lengths of the ancestors of every orthogroup, we calculated the probability that they remain the same for the number of generations between ancestral node and extant (leaf) nodes (254,462 generations for the hypothetical example used in the previous section). Next, we generated 10^6 pseudo-random numbers sampled from a uniform distribution with values ranging from 0 to 1, for each of the 758 orthogroups. For each of the 10^6 iterations, we calculated the total number of random values (out of 758) that exceed the corresponding probability that the ORF length remains the same as the ancestor (Monte–Carlo sampling). This number denotes the expected number of orthogroups that have ORFs with lengths different from that of the corresponding ancestors. In each of the 10^6 Monte–Carlo samples, we found that the expected number of orthogroups with length difference (median 37), does not exceed the corresponding observed number of orthogroups (172). Hence the observed numbers are greater than expected with a *P*-value less than 10^{-6}.

Analysis of ORF Length Changes

We analysed length changes in the orthogroups that contained ORFs with different lengths. To this end we first identified the longest ORF for each orthogroup. If there was not one unique longest ORF, we randomly picked one. Next, we compared the longest ORFs to all the other ORFs in the orthogroup using our BLASTP alignments, and classified these ORF pairs into four categories (i) both the ORFs shared both the start and the end positions (ii) both the ORFs shared the same start position but one ORF ended shorter (3′ change) (iii) both the ORFs shared the same end position but one ORF started further than the other (5′ change), and (iv) the shorter ORF was truncated in both sides, sharing neither the start nor the end with the longer ORF.

We analysed categories (ii) and (iii) further. First we analysed the length difference between the short and the longest ORFs (Δ). Next, we identified the codons in the longest ORF that overlap the terminal codons of the shorter ORFs. For example, the codon overlapping the stop codon. Conversely, we identified the nucleotide triplet (putative codon) that is located in the 3′UTR, at a distance Δ away from the stop codon of the shorter ORF. Likewise, we identified the putative codon in the 5′ UTR of the short ORFs that should positionally align to the start codon of the longest ORF.

We also used MAFFT (Katoh et al. 2002) to perform a global nucleotide alignment between the longest ORF and 5′ UTR (5′ length change) or 3′ UTR (3′ length change) of the short ORFs. We then retrieved the codon in the UTR of the short ORF that aligned to the start/stop codon of the long ORF, and checked the presence of gaps in the alignment.

Acknowledgements

M.K.L. acknowledges funding from the Deutsche Forschungsgemeinschaft priority program "Genomic Basis of Evolutionary Innovations" (SPP 2349), project BO 2544/20-1 awarded to E.B.-B. All the authors acknowledge support from the Open Access Publication Fund of the University of Münster.

Data Availability

All source data and analytical codes are freely available. Modeling scripts are available on GitHub:*BharatRavilyengar/DeNovoEvolution* (specifically Julia scripts ORFlen.jl ORFlen_defs.jl and analyseTree.jl). Scripts for *Drosophila melanogaster* data analysis are also available on Github: *MarieLebh/ORF_length_evolution*.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990:215(3):403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Baalsrud HT, Tørresen OK, Solbakken MH, Salzburger W, Hanel R, Jakobsen KS, Jentoft S. De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. Mol Biol Evol. 2017:35(3):593–606. https://doi.org/10.1093/molbev/msx311.

Blevins WR, Ruiz-Orera J, Messeguer X, Blasco-Moreno B, Villanueva-Cañas JL, Espinar L, Díez J, Carey LB, Albà MM. Uncovering *de novo* gene birth in yeast using deep transcriptomics. Nat Commun. 2021:12(1):604. https://doi.org/10.1038/s41467-021-20911-3.

Bloom JD, Labthavikul ST, Otey CR, Arnold FH. Protein stability promotes evolvability. Proc Natl Acad Sci USA. 2006:103(15):5869–5874. https://doi.org/10.1073/pnas.0510098103.

Bornberg-Bauer E, Chan HS. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. Proc Natl Acad Sci USA. 1999:96(19):10689–10694. https://doi.org/10.1073/pnas.96.19.10689.

Bucciantini M, Giannoni E, Chiti F, Baroni F, Formigli L, Zurdo J, Taddei N, Ramponi G, Dobson CM, Stefani M. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. Nature. 2002:416(6880):507–511. https://doi.org/10.1038/416507a.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC

Bioinformatics. 2009:10(1):421. https://doi.org/10.1186/1471-2105-10-421.
Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. Proto-genes and *de novo* gene birth. Nature. 2012: 487(7407):370–374. https://doi.org/10.1038/nature11184.
Dill KA. Theory for the folding and stability of globular proteins. Biochemistry. 1985:24(6):1501–1509. https://doi.org/10.1021/bi00327a032.
Dowling D, Schmitz JF, Bornberg-Bauer E. Stochastic gain and loss of novel transcribed open reading frames in the human lineage. Genome Biol Evol. 2020:12(11):2183–2195. https://doi.org/10.1093/gbe/evaa194.
Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol. 2005:6(3):197–208. https://doi.org/10.1038/nrm1589.
Fernández-Moreno MA, Farr CL, Kaguni LS, Garesse R. *Drosophila melanogaster* as a model system to study mitochondrial biology. Methods in molecular biology. Totowa (NJ): Humana Press; 2007. p. 33–49.
Grandchamp A, Czuppon P, Bornberg-Bauer E. Quantification and modeling of turnover dynamics of *de novo* transcripts in *Drosophila melanogaster*. Nucleic Acids Res. 2023b:52(1): 274–287. https://doi.org/10.1093/nar/gkad1079.
Grandchamp A, Kühl L, Lebherz M, Brüggemann K, Parsch J, Bornberg-Bauer E. Population genomics reveals mechanisms and dynamics of *de novo* expressed open reading frame emergence in *Drosophila melanogaster*. Genome Res. 2023a:33(6): 872–890. https://doi.org/10.1101/gr.277482.122.
Gubala AM, Schmitz JF, Kearns MJ, Vinh TT, Bornberg-Bauer E, Wolfner MF, Findlay GD. The goddard and saturn genes are essential for *Drosophila* male fertility and may have arisen *de novo*. Mol Biol Evol. 2017:34(5):1066–1082. https://doi.org/10.1093/molbev/msx057.
Hartl FU. Protein misfolding diseases. Annu Rev Biochem. 2017:86(1): 21–26. https://doi.org/10.1146/biochem.2017.86.issue-1.
Heames B, Schmitz J, Bornberg-Bauer E. A continuum of evolving *de novo* genes drives protein-coding novelty in *Drosophila*. J Mol Evol. 2020:88(4):382–398. https://doi.org/10.1007/s00239-020-09939-z.
Iyengar BR, Bornberg-Bauer E. Neutral models of *de novo* gene emergence suggest that gene evolution has a preferred trajectory. Mol Biol Evol. 2023:40(4):msad079. https://doi.org/10.1093/molbev/msad079.
Jacob F. Evolution and tinkering. Science. 1977:196(4295): 1161–1166. https://doi.org/10.1126/science.860134.
Jee J, Rasouly A, Shamovsky I, Akivis Y, Steinman SR, Mishra B, Nudler E. Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. Nature. 2016:534(7609):693–696. https://doi.org/10.1038/nature18313.
Kapun M, Barrón MG, Staubach F, Obbard DJ, Wiberg RAW, Vieira J, Goubert C, Rota-Stabelli O, Kankare M, Bogaerts-Márquez M, et al. Genomic analysis of European Drosophila melanogaster populations reveals longitudinal structure, continent-wide selection, and previously unknown DNA viruses. Mol Biol Evol. 2020:37(9): 2661–2678. https://doi.org/10.1093/molbev/msaa120.
Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002:30(14):3059–3066. https://doi.org/10.1093/nar/gkf436.
Kleppe AS, Bornberg-Bauer E. Robustness by intrinsically disordered c-termini and translational readthrough. Nucleic Acids Res. 2018:46(19): 10184–10194. https://doi.org/10.1093/nar/gky778.
Lange A, Patel PH, Heames B, Damry AM, Saenger T, Jackson CJ, Findlay GD, Bornberg-Bauer E. Structural and functional characterization of a putative *de novo* gene in *Drosophila*. Nat Commun. 2021:12(1):1667. https://doi.org/10.1038/s41467-021-21667-6.
Laurent SJ, Werzner A, Excoffier L, Stephan W. Approximate Bayesian analysis of *Drosophila melanogaster* polymorphism data reveals a recent colonization of Southeast Asia. Mol Biol Evol. 2011:28(7): 2041–2051. https://doi.org/10.1093/molbev/msr031.
Lebherz MK, Fouks B, Schmidt J, Bornberg-Bauer E, Grandchamp A. DNA transposons favour *de novo* transcript emergence through enrichment of transcription factor binding motifs. Genome Biol Evol. 2024:evae134. https://doi.org/10.1093/gbe/evae134.
Li H, Stephan W. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genet. 2006:2(10):e166. https://doi.org/10.1371/journal.pgen.0020166.
Middendorf L, Eicholt LA. Random, *de novo*, and conserved proteins: how structure and disorder predictors perform differently. Proteins: Struct Funct Bioinf. 2024:92(6):757–767. https://doi.org/10.1002/prot.26652.
Neme R, Tautz D. Phylogenetic patterns of emergence of new genes support a model of frequent *de novo* evolution. BMC Genom. 2013:14(1):117. https://doi.org/10.1186/1471-2164-14-117.
Neme R, Tautz D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to *de novo* gene emergence. eLife. 2016:5:e09977. https://doi.org/10.7554/eLife.09977.
Papadopoulos C, Callebaut I, Gelly J-C, Hatin I, Namy O, Renard M, Lespinet O, Lopes A. Intergenic ORFs as elementary structural modules of *de novo* gene birth and protein evolution. Genome Res. 2021:31(12):2303–2315. https://doi.org/10.1101/gr.275638.121.
Prabh N, Rödelsperger C. *De novo*, divergence, and mixed origin contribute to the emergence of orphan genes in *Pristionchus* nematodes. G3 Genes Genom Genet. 2019:9(7):2277–2286. https://doi.org/10.1534/g3.119.400326.
Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000:16(6):276–277. https://doi.org/10.1016/S0168-9525(00)02024-2.
Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. Long non-coding RNAs as a source of new peptides. eLife. 2014:3:e03523. https://doi.org/10.7554/eLife.03523.
Schmitz J, Bornberg-Bauer E. Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA. F1000Research. 2017:6:57. https://doi.org/10.12688/f1000research.10079.1.
Schrider DR, Houle D, Lynch M, Hahn MW. Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. Genetics. 2013:194(4):937–954. https://doi.org/10.1534/genetics.113.151670.
Vakirlis N, Acar O, Hsu B, Coelho NC, Van Oss SB, Wacholder A, Medetgul-Ernar K, Bowman RW, Hines CP, Iannotta J, et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. Nat Commun. 2020:11(1):781. https://doi.org/10.1038/s41467-020-14500-z.
Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, Coon JJ, Lafontaine I. A molecular portrait of de novo genes in yeasts. Mol Biol Evol. 2017:35(3):631–645. https://doi.org/10.1093/molbev/msx315.
Van Oss SB, Carvunis A-R. *De novo* gene birth. PLoS Genet. 2019:15(5):1–23. https://doi.org/10.1371/journal.pgen.1008160.
Wacholder A, Parikh SB, Coelho NC, Acar O, Houghton C, Chou L, Carvunis A-R. A vast evolutionarily transient translatome contributes to phenotype and fitness. Cell Syst. 2023:14(5):363–381.e8. https://doi.org/10.1016/j.cels.2023.04.002.
Witt E, Benjamin S, Svetec N, Zhao L. Testis single-cell RNA-seq reveals the dynamics of *de novo* gene transcription and germline

mutational bias in *Drosophila*. eLife. 2019:8:e47138. https://doi.org/10.7554/eLife.47138.
Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev Mol Cell Biol. 2014:16(1):18–29. https://doi.org/10.1038/nrm3920.
Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, Yu Y, Hou G, Zi J, Zhou R, et al. Rapid evolution of protein diversity by de novo origination in *Oryza*. Nat Ecol Evol. 2019:3(4):679–690. https://doi.org/10.1038/s41559-019-0822-5.
Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of *de novo* genes in *Drosophila melanogaster* populations. Science. 2014:343(6172):769–772. https://doi.org/10.1126/science.1248286.
Zheng J, Guo N, Wagner A. Selection enhances protein evolvability by increasing mutational robustness and foldability. Science. 2020:370(6521):eabb5962. https://doi.org/10.1126/science.abb5962.
Zhu YO, Siegal ML, Hall DW, Petrov DA. Precise estimates of mutation rate and spectrum in yeast. Proc Natl Acad Sci USA. 2014:111(22):E2310-8. https://doi.org/10.1073/pnas.1323011111.

Associate editor: Claudia Alvarez Carreño