# Population genomics reveals mechanisms and dynamics of de novo expressed open reading frame emergence in *Drosophila melanogaster*

Anna Grandchamp, Lucas Kühl, Marie Lebherz, et al.

*Genome Res.* 2023 33: 872-890 originally published online July 13, 2023
Access the most recent version at doi:10.1101/gr.277482.122

# Population genomics reveals mechanisms and dynamics of de novo expressed open reading frame emergence in *Drosophila melanogaster*

Anna Grandchamp,[1] Lucas Kühl,[1] Marie Lebherz,[1] Kathrin Brüggemann,[1] John Parsch,[2] and Erich Bornberg-Bauer[1,3]

[1]*Institute for Evolution and Biodiversity, University of Münster, 48149 Münster, Germany;* [2]*Division of Evolutionary Biology, Faculty of Biology, Ludwig-Maximilians-Universität München, 82152 Munich, Germany;* [3]*Max Planck Institute for Biology Tübingen, Department of Protein Evolution, 72076 Tübingen, Germany*

**Novel genes are essential for evolutionary innovations and differ substantially even between closely related species. Recently, multiple studies across many taxa showed that some novel genes arise de novo, that is, from previously noncoding DNA. To characterize the underlying mutations that allowed de novo gene emergence and their order of occurrence, homologous regions must be detected within noncoding sequences in closely related sister genomes. So far, most studies do not detect noncoding homologs of de novo genes because of incomplete assemblies and annotations, and long evolutionary distances separating genomes. Here, we overcome these issues by searching for de novo expressed open reading frames (neORFs), the not-yet fixed precursors of de novo genes that emerged within a single species. We sequenced and assembled genomes with long-read technology and the corresponding transcriptomes from inbred lines of *Drosophila melanogaster*, derived from seven geographically diverse populations. We found line-specific neORFs in abundance but few neORFs shared by lines, suggesting a rapid turnover. Gain and loss of transcription is more frequent than the creation of ORFs, for example, by forming new start and stop codons. Consequently, the gain of ORFs becomes rate limiting and is frequently the initial step in neORFs emergence. Furthermore, transposable elements (TEs) are major drivers for intragenomic duplications of neORFs, yet TE insertions are less important for the emergence of neORFs. However, highly mutable genomic regions around TEs provide new features that enable gene birth. In conclusion, neORFs have a high birth-death rate, are rapidly purged, but surviving neORFs spread neutrally through populations and within genomes.**

**[Supplemental material is available for this article.]**

De novo gene origination is a recently recognized process that describes the emergence of new genes from previously noncoding sequences via a series of mutational events (Kaessmann 2010; Schlötterer 2015; Levy 2019). It has been conjectured that over very short evolutionary time scales, most novel genes originate de novo but are then rapidly lost (Schmitz et al. 2018; Rödelsperger et al. 2019; Lange et al. 2021). Most probably, newly emerged de novo genes fail to establish a sufficient advantage to the host organism, and thus are easily purged by drift or negative selection. Over longer time scales, most observed novel genes stem from duplication events. Proteins derived from duplicated genes are more likely to have already structural elements such as domains and motifs, and functions which do not negatively interfere with the long established cellular network (O'Toole et al. 2018). Duplicated proteins gain new functional roles through mechanisms such as sub- or neo-functionalization (Rastogi and Liberles 2005; Innan and Kondrashov 2010; Konrad et al. 2011).

The emergence of functional proteins from de novo genes is difficult to rationalize with our current understanding of molecular genetics and evolution. Although believed to be highly unlikely until recently, de novo gene birth has by now been described in many species, including several fungi, plants, insects, mammals, and fishes (Zhao et al. 2014; Ruiz-Orera et al. 2015; Li et al. 2016; Wu and Knudson 2018). Some genes that emerge de novo not only provide new functions, but have even become essential to the species in which they emerged. For example, the human-specific de novo gene *ESRG* is required for the maintenance of pluripotency in human naive stem cells (Wang et al. 2014). *FYV5* (also known as *MDF1*), a de novo gene found only in *Saccharomyces cerevisiae*, has a function in suppressing sexual reproduction (Li et al. 2014). In humans, de novo genes have been shown to play a role in brain development (Kaessmann 2010; Li et al. 2010; Maze et al. 2015; Cutter et al. 2019). Several predicted de novo genes in the *Drosophila* genus were shown to have become essential for male fertility (Gubala et al. 2017; Rivard et al. 2021). Overall, many of the predicted de novo genes in metazoa seem to be associated with either developmental processes of the neuronal system or with reproduction. Both of these processes are well known for their fast genetic turnover and adaptation.

To ascertain that a candidate de novo gene did not emerge via a duplication or transposition event, the gene must not show homology with any other gene in the species in which it emerged and to any other gene in an outgroup species of the taxonomic group under study. However, as duplicated genes are more likely to evolve faster than single-copy genes (O'Toole et al. 2018), a

**Corresponding author: a.grandchamp@uni-muenster.de**
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.277482.122.




33:872–890 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/23; www.genome.org

duplicated gene could also lack a detectable orthology relationship. Most earlier studies of de novo genes relied on the comparative genomics of species that diverged several tens of millions of years (my) ago (Begun et al. 2006; Cai et al. 2008; Neme and Tautz 2013; McLysaght and Guerzoni 2015; Schmitz et al. 2018; Dowling et al. 2020). This led to considerable discussions regarding the reliability of their de novo status (Moyers and Zhang 2016; Domazet-Lošo et al. 2017; Weisman et al. 2020).

The only way to conclusively show that a gene emerged from scratch is to locate the corresponding homologous noncoding sequence in the closest outgroup species and study the putative mutations that transformed a noncoding sequence into a coding one. For a de novo gene to emerge, at least two events must occur: the emergence of a new open reading frame (ORF), and the gain of transcription. Nonetheless, genomes and synteny between two closely related species can be extensively reshuffled during evolution.

The extensive divergence of noncoding regions because of the lack of purifying selection implies that, the deeper the phylogeny becomes, the less likely the noncoding regions that are homologous to a predicted de novo gene can be identified. More sensitive phylogenetic procedures that rigorously refine the first-trawl-results from phylostratigraphic analyses have been suggested and implemented (McLysaght and Guerzoni 2015; Wu and Knudson 2018; Heames et al. 2020). As a key feature, they use gene syntenies to allow for highly sensitive matches (i.e., those with little sequence similarity) if de novo genes appear between the same gene pairs in two or more genomes. Using synteny works well in mammals (Jebb et al. 2020; Vakirlis et al. 2020) or in vertebrates in general but gene order is less conserved in taxa such as insects (Zdobnov and Bork 2007).

An improved strategy to study the emergence of de novo genes is to study genomes of high quality covering a very narrow time scale. Using several rice genomes, Zhang et al. (2019) managed to pinpoint precisely the mutations allowing emergence of de novo genes in closely related rice species. A recent state-of-the-art study used six nematode populations and confirmed the rapid turnover of de novo genes compared to duplicates (Prabh and Rödelsperger 2022). Several other studies included fewer species but focused on even shorter evolutionary time scales by using population data. By analyzing population data from wild fish populations, it has been corroborated that random emergence (and rapid loss) of transcripts are frequent and that (few) surviving transcripts gain stable and broader expressions, presumably via fast epigenetic control (Schmitz et al. 2020). Zhao et al. (2014) identified 106 de novo genes specifically expressed in the testis of laboratory strains of *Drosophila melanogaster*. Furthermore, Moyers and Zhang (2016) showed that 13.9% of *D. melanogaster* genes did not have a detectable ortholog in any outgroup species, and Heames et al. (2020) reported that 26.8% of the transcribed ORFs of *D. melanogaster* originated from orphan genes.

Despite this progress, all of these studies are hampered by either data disparity, the use of discontinuous evolutionary time scales, a lack of precisely annotated genomes for inter- and intraspecific comparison, or several of these issues. For these reasons, the full extent of de novo genes in a species, the mutations enabling their emergence, and their prevalence in natural populations, remain unknown.

In the present study, we aimed to overcome several of the limitations of earlier studies by using in-bred lines of *D. melanogaster* from various geographic locations to generate line specific genomes and transcriptomes. By using long-read sequence data (Oxford Nanopore Technologies [ONT]) and a common annotation strategy across all genomes, we also avoid recently recognized issues arising from the use of disparate methods such as gene detection failure, synteny loss, different expression signals because of different sequencing technologies, fragmented genomes that do not fully cover a transcript, or fragmented transcripts not covering the full genomic region, etc. (Weisman et al. 2022). Then, we aimed to localize de novo genes emerged in populations and to follow mutations and structural variations associated with the very early stages of de novo gene emergence, demonstrating their emergence from noncoding DNA and providing an early snapshot of the origin of de novo genes. Because most of these de novo genes have not yet become fixed within a species and are neither stably transcribed or translated across species boundaries, we refer to them as new expressed ORFs (neORFs), following the terminology proposed by Zhao et al. (2014).

## Results

### Genome sequencing, assembly, and annotation

We studied neORF emergence in seven inbred lines of *D. melanogaster*, including six from derived European populations and one from the ancestral species range in Zambia (Fig. 1A; Supplemental Table S1), using ONT long-read sequencing. We obtained 9–21 Gb of raw DNA sequence per line (see Data access), resulting in 66–158 × coverage (Table 1). The longest reads per genome ranged from 147–260 kb. An exceptionally long read of 8.3 Mb was obtained from the Zambian line. For each line, RNA was sequenced with short, paired reads (2 × 150 bp) with at least 92M reads/sample, resulting in a total of 8.8–12.3 Mb of RNA sequence per sample (Table 1). De novo genomes were assembled with Canu, and the assembly quality was improved following several steps (see Methods; Supplemental Data). BUSCO scores, which indicate the completeness of the genome assemblies, ranged from 97.7% (FI) to 99% (ZI) (Table 1). The genomes were aligned between lines, and the alignments did not show evidence of major reshuffling, indicating that chromosome sequence identities are mostly consistent across the lines (Fig. 1B). The total number of genes identified per line and referred to as singleton by BUSCO ranged from 95.6% (UA) to 98% (ZI) in the final assemblies. The total number of duplicated copies of genes was within the range of 1% (ZI) to 2.1% (TR). After scaffolding, all genomes contained the eight chromosome arms present in the *D. melanogaster* reference genome (X, 2L, 2R, 3L, 3R, 4, Y, mitochondria).

Established genes annotated in the reference genome of *D. melanogaster* were annotated in the seven new genomes. We detected a total number of genes ranging from 13,471 (ES) to 13,662 (TR), representing 97% to 99% of established genes from *D. melanogaster* (Supplemental Data). Based on a visual inspection and on the similarity of the pairwise genome alignments, we observed the orthologous genes distributions were consistent from one line to the other (Fig. 1C; Supplemental Tables S2–S5). The $d_S$ values (number of synonymous substitutions per synonymous site) were more variable, showing increased or decreased gene mutation diversity within a line, and between different populations (Prabh and Rödelsperger 2022). Most of the genes had a $d_S$ value lower than 0.1. Very few genes had a $d_S$ value superior to 0.5 (Supplemental Table S4). The segment 89 of Chromosome 3L contained only one gene, *CG43744*, which encodes an RNA-binding protein that binds to the EDEN element and mediates maternal mRNA translational

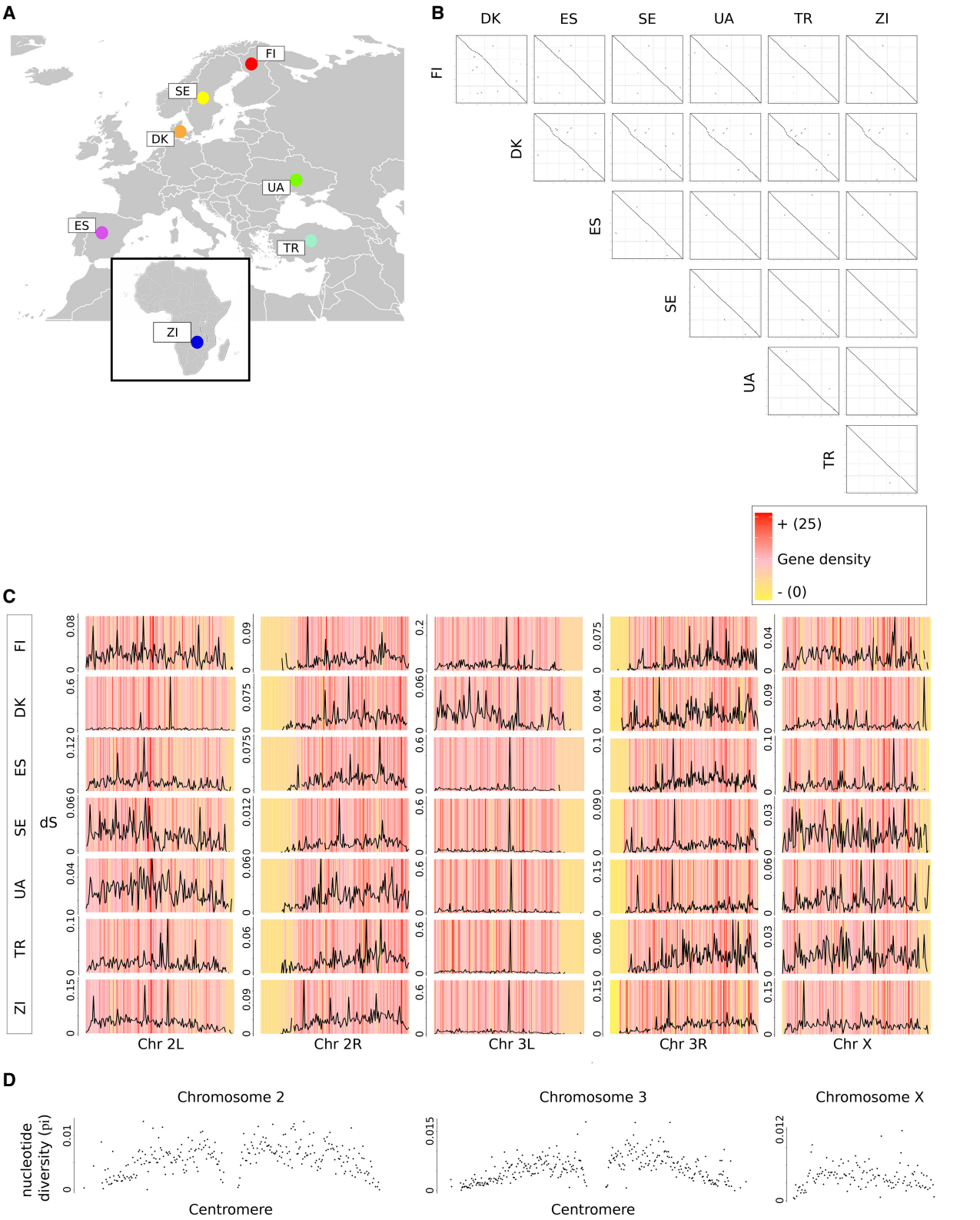


**Figure 1.** Genome alignments and gene contents. (*A*) Geographic origins of lines; (*B*) dot plots of all pairwise genome alignments; (*C*) gene density and $d_S$ values of genes in genomes of each line. The five major chromosome arms are shown for each line. The colors represent the density of established genes in successive genomic intervals of 160,000 bp. Red denotes highest density; yellow, lowest density. Black lines represent average $d_S$ values of established genes located in the respective interval. $d_S$ was calculated by aligning genes from all lines to orthologs from the reference genomes (see Methods section). (*D*) Average gene polymorphism in Chromosomes 2, 3, and X. The dots represent the nucleotide diversity $\pi$ of established genes in successive genomic intervals of 160,000 bp.

**Table 1.** Descriptive statistics of genome assemblies and annotations

| | FI (Finland) | DK (Denmark) | ES (Spain) | SE (Sweden) | UA (Ukraine) | TR (Turkey) | ZI (Zambia) |
|---|---|---|---|---|---|---|---|
| Estimated DNA base | 9.12 Gb | 14.49 Gb | 21.82 Gb | 15.96 Gb | 9.66 Gb | 20.38 Gb | 10.28 Gb |
| Estimated genome coverage | 66 | 105 | 158 | 115 | 70 | 148 | 75 |
| Size longest read | 147 kb | 220 kb | 180 kb | 260 kb | 160 kb | 220 kb | 8.26 Mb |
| Size assembly | 131 Mb | 129 Mb | 135 Mb | 131 Mb | 131 Mb | 133 Mb | 129 Mb |
| Estimated RNA base forward | 8.8 Mb | 10.7 Mb | 9.1 Mb | 9.1 Mb | 9.2 Mb | 11.9 Mb | 9.4 Mb |
| Estimated RNA base reverse | 9.2 Mb | 11.3 Mb | 9.4 Mb | 9.5 Mb | 9.4 Mb | 12.3 Mb | 9.6 Mb |
| Complete BUSCO (%) | 97.6 | 98.9 | 98.8 | 98.7 | 97.5 | 98.2 | 98.9 |
| Duplicated BUSCO (%) | 1.2 | 1 | 1.2 | 1 | 1.8 | 2.1 | 0.8 |
| DNA mapping rate (%) | 95.63 | 94.83 | 94.02 | 93.81 | 96.00 | 94.88 | 95.84 |
| RNA mapping rate (%) | 91.83 | 95.87 | 94.36 | 97.11 | 95.29 | 96.11 | 96.05 |
| Genome GC content (%) | 39.16 | 39.31 | 37.98 | 37.61 | 37.70 | 37.86 | 37.83 |
| # annotated genes | 13,471 | 13,534 | 12,594 | 13,566 | 13,535 | 13,662 | 13,480 |
| Chromosome arms | 8 | 8 | 8 | 8 | 8 | 8 | 8 |

repression (FlyBase) (Gramates et al. 2022). This gene shows a $d_S$ value superior to 0.6 in all lines except FI and DK. We also looked at the polymorphism of established genes between lines (Fig. 1D; Supplemental Table S6). In 2012, Mackay et al. (2012) published a genetic reference panel for *D. melanogaster*. They observed a reduced nucleotide diversity in centromeres and telomeres of autosomes compared to noncentromeric regions. Moreover, they observed a reduced nucleotide diversity on the X Chromosome compared to the autosomes. These results were also observed in *Drosophila simulans* (Begun et al. 2007). Our results show exactly the same pattern, with reduced nucleotide diversity $\pi$ between lines in telomeres and centromeres. Moreover, the average and median nucleotide diversity was lower for X-linked compared to autosomal genes (Autosomal; mean: $5.3 \times 10^{-3}$; X: mean: $4 \times 10^{-3}$).

Transcripts were mapped to their corresponding genomes, with percentages of mapping success ranging from 91.2% (FI) to 97.1% (SE) (Table 1; Supplemental Data).

## neORFs in *D. melanogaster* lines

We searched for neORFs independently in each line of *D. melanogaster*. A neORFs is here defined as an ORF detected within a spliced transcript, which is longer than 90 bp, is in frame with the direction of transcription, that possesses a putative 5′ and 3′ UTR and corresponds to a new transcript found in a noncoding region of one or more of the seven genomes. These neORFs show no detectable homology with any species other than *D. melanogaster* nor to the established genes of *D. melanogaster* (see Methods). On average, we found 1548 neORFs per line (from 1268 [SE] to 1840 [UA]) (Fig. 2A; Supplemental Tables S7–S14). The Zambian line did not differ from the European lines in terms of its total number of neORFs (Fig. 2A). Because the Zambian population is known to have a larger effective population size than European populations and to remove deleterious variation more effectively (Kapopoulou et al. 2018), this suggests that the vast majority of neORFs are not deleterious, but instead are effectively neutral with respect to

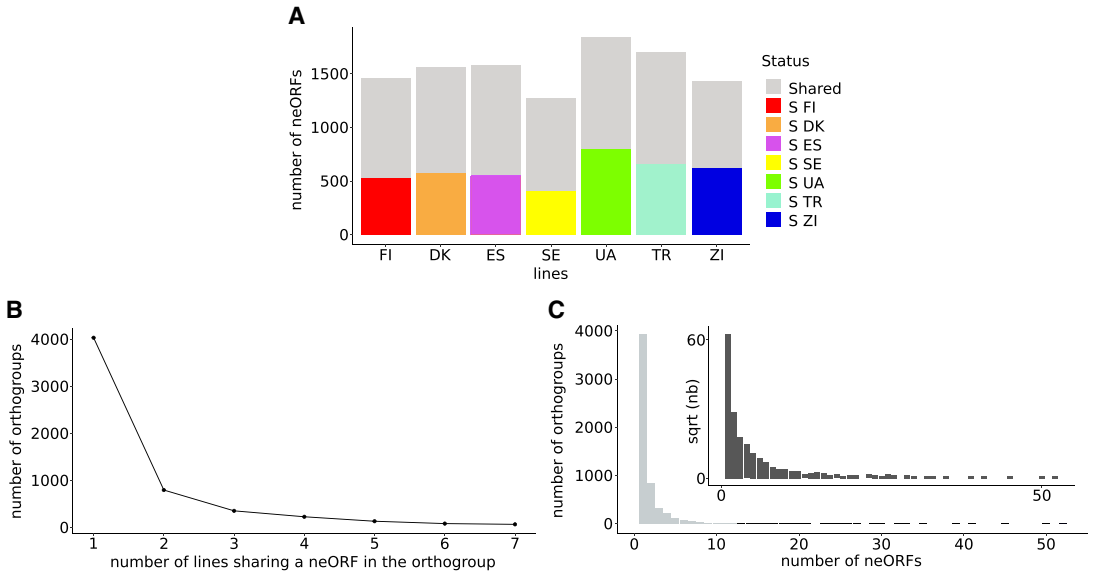


**Figure 2.** Characteristics of orthogroups from the seven lines of *D. melanogaster*. (*A*) Number of neORFs per each line. Colors represent neORFs that are specific to their line. Gray denotes neORFs with orthologs in other lines. (*B*) Number of orthogroups found in one line or shared by several lines. The *x*-axis represents the number of lines sharing a neORF in an orthogroup. The *y*-axis represents the number of orthogroups. (*C*) Number of neORFs per orthogroup. The *x*-axis represents the number of neORFs presents in an orthogroup, the *y*-axis represents the number of orthogroups corresponding to these numbers.

fitness. An average of 593 neORFs were specific to any one line, with the Zambian line containing more unique neORFs in proportion (45%) than any of the European lines. This is likely the consequence of the shared out-of-Africa demographic history of the European lines, the relatively high level of gene flow among European populations and the limited amount of European introgression into sub-Saharan African populations (Pool et al. 2012; Kapun et al. 2020). Homology search tools were used to gather neORFs shared by line in orthogroups. Overall, 5687 orthogroups were established (Fig. 2B; Supplemental Tables S15–S17). Most orthogroups (4010/5687; 71%) contained neORFs specific to a single line. Among these 4010 orthogroups, the vast majority (3989) comprise a single neORF, whereas the remaining ones comprise neORFs duplicated within the same genome. The number of orthogroups sharply declines for those containing neORFs found in two genomes and continues, more moderately, for orthogroups containing instances from 3 to 7 lines. The total number of orthogroups containing a neORF shared by all the seven lines was only 66. As some neORFs were found to be duplicated and present at different genomic positions within a single line, we investigated the number of copies of a neORF contained by an orthogroup (Fig. 2C). We observed that 16% to 29% of neORFs were present in at least two copies in the same genome of one line. In total, 35 orthogroups contained over 14 and up to 52 neORFs, showing the highest level of duplication, even though most orthogroups contained one neORF per line. As stated earlier, most of the orthogroups specific to one single line only contained one neORF, indicating that most duplications occured in neORFs that are shared by several lines, and thus are putatively older.

For each line, we next investigated the genomic position of the neORFs (Fig. 3A; Supplemental Table S14). The vast majority of neORFs (79% [ZI] to 81.7% [UA]) were located in *intergenic* regions. Between 8.5% (UA) and 12.5% (TR) of neORFs were found to be *antisense*, that is, overlapping with an annotated gene but transcribed in the opposite direction. Between 1.4% (DK, ZI) and 2.1% (FI) neORFs emerged inside an exon of an annotated gene (*ExonInside*). Between 1.2% (DK) and 2.2% (FI) neORFs partially overlapped to an existing gene and the upstream/downstream noncoding region. Finally, 4.4% (ES) to 6.7% (ZI) neORFs emerged inside an annotated intron. Most of these introns were antisense in all lines (64.2% [FI] to 80.6% [ES]). These percentages followed the same trends in all seven lines. The neORFs that emerged inside an exon or overlapping it did not correspond to the extension of existing ORFs inside a gene, as they showed no protein BLAST hit to the proteins expressed by the gene to which they overlapped. The neORFs inside an exon are new ORFs which have emerged fully inside a gene, but would code for a protein that does not correspond to any protein in the gene. Most likely, these ORFs were nested into an alternative frame of the main genic ORF. To avoid any bias, neORFs that emerged inside an exon were removed from the analysis in the study of synteny. Concerning neORFs overlapping an exon, the exonic overlap was retrieved (Supplemental Figs. S1–S4). In each line, about half of ExonLonger neORFs had new transcripts that started upstream from the start site of the preceding gene and ended downstream from the end site, with the exception of the Zambian line. The fraction of the neORF's transcript sequence covering the established gene was otherwise smaller than 50%.

For neORFs located inside of an existing gene in sense (ExonInside), we investigated which part of the gene they were overlapping with (Fig. 3B). In each line, most of these neORFs were located inside of an ORF of the gene. The fact that protein BLAST failed to identify them as existing proteins suggests that they are located in an alternate reading frame. Some of the neORFs also overlapped with the UTR regions or were fully located inside the UTR. The same analysis was performed for neORFs from the genomic positions "ExonLonger" and "Antisense" (Supplemental Table S18).

## Characteristics of neORFs and their encoded proteins

We next investigated several properties that have been reported to differ between de novo genes and established genes, that is, evolutionary older genes. First, we compared the length of de novo genic ORFs between lines. We find the length of ORFs in neORFs to be significantly higher for neORFs which are shared by all seven lines (average length 235 nt), than for those shared by fewer lines (average 216 nt) with high significance (*t*-test, *P*-value $1.1 \times 10^{-9}$) (Fig. 4A; Supplemental Table S19). Otherwise, no differences in length were found between the neORFs of the different conservation classes. The lengths of neORFs were also compared with the lengths of ORFs of de novo genes from *Drosophila* of different ages (Heames et al. 2020). The lengths of these ORFs were systematically longer than neORFs, and increased with the age of the de novo genes (Supplemental Figs. S5–S7).

In each line, around 45% of all neORFs had introns within their transcripts (Supplemental Tables S19, S20). We assessed the average number of exons in the unspliced neORFs, and searched for any correlation between the genomic position of neORFs and their number of exons (Fig. 4B). Independently of the lines, neORFs overlapping an existing exon ("ExonInside," "Antisense," "ExonLonger"), had significantly more exons (and therefore introns too) than the neORFs found in "Intronic" and "Intergenic" positions (Supplemental Table S19).

We next investigated if the neORFs had potential to be translated. Hexamer-scores were all low with average values below 0, as described by Dowling et al. (2020) (Supplemental Table S21). We observed no increase in hexamer score with the number of lines sharing a neORF, contrary to earlier reports (Schmitz et al. 2018), that compared genomes across species and, therefore, over much longer evolutionary time scales.

Aggregation propensity and intrinsic disorder of proteins encoded by neORFs were evaluated and correlated to the number of lines in which the neORF was found. The levels of aggregation propensity were systematically very low and below 20, indicating that none of the putative proteins encoded by neORFs would be prone to aggregate. Similarly, the levels of intrinsic disorder were low in each group. Still, the more lines a neORF was found in, the higher was its aggregation propensity (linear model, $P = 2 \times 10^{-16}$), and the lower was its level of intrinsic disorder (linear model, $P = 2 \times 10^{-16}$) (Fig. 4C,D; Supplemental Tables S19, S22, S23). When compared to de novo genes, we observed that this increase in aggregation propensity and decrease in intrinsic disorder with age was also true for the youngest de novo genes. However, these trends were reversed for the oldest de novo genes.

## Impact of transposable elements

Transposable elements (TEs) have been repeatedly shown to translocate not only themselves but to carry along other, possibly duplicated genes (Tan et al. 2021) and to integrate in all possible genomic positions, including regulatory regions, exons and introns (Zhang et al. 2011; Huff et al. 2016). TEs are, therefore, widely seen as drivers of evolution. To examine the impact of TEs on the emergence of neORFs, we tested if neORFs could be found: (i) inside a TE; (ii) overlapping a TE, or (iii) outside of a TE (Fig. 5A). We find that between

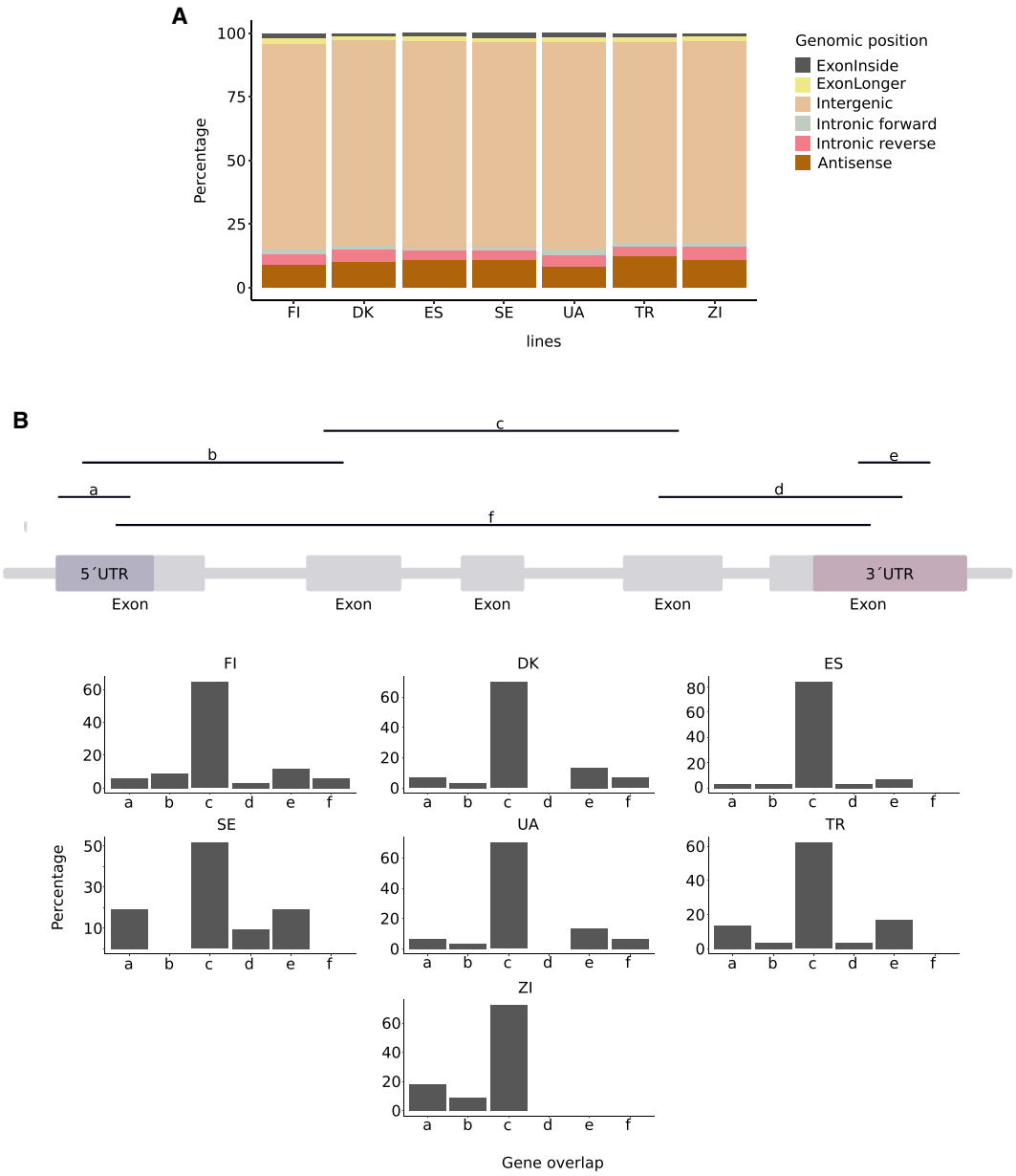


**Figure 3.** Genomic positions of neORFs in the seven lines of *D. melanogaster*. (*A*) Genomic positions in which neORFs are found. (*B*) Position of the neORF for neORFs that emerged inside of an existing gene (ExonInside). (*a*) The neORF is inside the 5′ UTR of the established gene. (*b*) The neORF overlaps with the 5′ UTR and at least one coding exon of the established gene. (*c*) The neORF is inside a CDS. (*d*) The neORF overlaps with a coding exon and the 3′ UTR of the established gene. (*e*) The neORF is inside the 3′ UTR of the established gene. (*f*) The neORF overlaps with all parts of the established gene.

40% (ES) and 51% (ZI) of neORFs were found to be outside of a TE (Supplemental Tables S24, S25; Supplemental Fig. S8). The fact that 16%–21% of neORFs overlapped a TE could suggest that the insertion of a TE in a noncoding sequence might have provided the sequence changes sufficient to create an ORF. Finally, between 31% and 44% of neORFs were found to be entirely located inside a TE. These results suggest that the emergence of neORFs is partly correlated with the presence of a TE (Supplemental Fig. S9).

For neORFs overlapping with a TE, we measured the percentage of the sequences involved in the overlap (Fig. 5B; Supplemental Figs. S10, S11). Most of these overlaps involved 0%–25% of the neORF sequence. As 16%–29% of neORFs were found to be present in at least two copies in the same genome of one line, we wondered if genes duplicated locally, or rather if the copies were located on different genomic positions. The homologous copies of one single neORF were often found on different chromosomes (Fig. 5C). Most of the neORFs found in two copies belong to chromosome arms 2R and 3R. Chromosome arms 2L and 3R also share many neORFs in all lines, except FI. neORFs found in 2L seem to have homologs in all other chromosomes except the mitochondria, which possesses no neORFs except in FI. As duplicated copies of neORFs were mainly found at different genomic locations, we investigated their overlap to TEs. The vast majority of duplicated neORFs was made up of neORFs found inside a TE (UA: 74% to ES: 84%, Supplemental Table S25), and a small amount overlapped to a TE (DK: 6% to UA: 11%). Finally, we investigated the family of TEs associated with

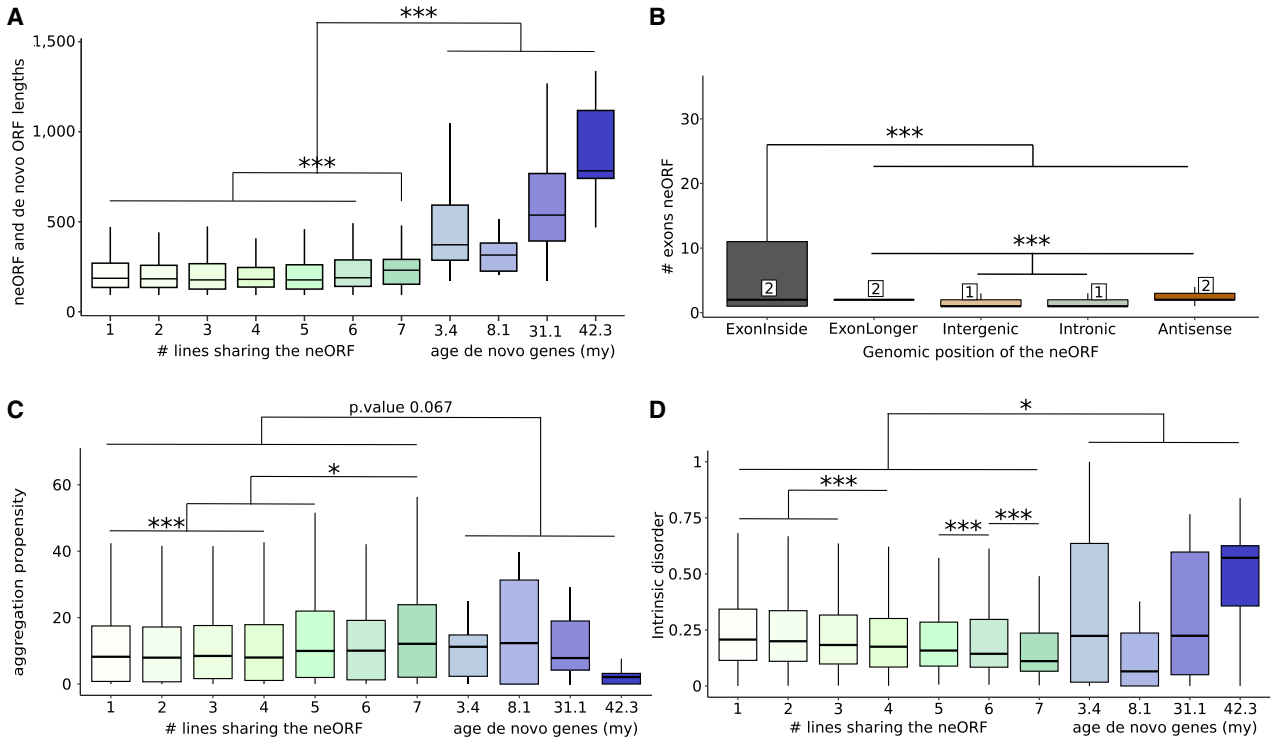


**Figure 4.** neORF properties. (*A*) Average neORF lengths. The *x*-axis represents the number of lines sharing a neORF. The number 1 means that the neORF is found only in one line. The *y*-axis represents the neORFs length. The four boxes on the *right* represent the average lengths of de novo genes found in *Drosophila* with different ages. (*B*) Number of exons in neORFs. The *x*-axis shows the genomic position of neORFs. The *y*-axis shows the average number of exons. The numbers in rectangles represent the median values. (*C*) Aggregation propensity of neORFs. The *x*-axis represents the number of lines sharing a neORF. The *y*-axis shows the aggregation propensity of putative proteins encoded by neORFs. The four boxes on the *right* represent the aggregation propensity of de novo genes found in *Drosophila* with different ages. (*D*) Intrinsic disorder of neORFs. The *x*-axis represents the number of lines sharing a neORF. The *y*-axis shows the intrinsic disorder of putative proteins encoded by neORFs. The four boxes on the *right* represent the intrinsic disorder of de novo genes found in *Drosophila* with different ages.

neORF duplications (Fig. 5D). We observed that long terminal repeat (LTR) retrotransposons were the TEs most associated with neORFs, followed by those with terminal inverted repeats. These results strongly suggest that transposable elements were responsible for most duplication events in duplicated neORFs, and again shows the importance of TEs in the origin of neORFs, and their spread within the genomes.

We next calculated the frequencies of TEs and of neORFs per 160,000 kb intervals on all chromosomes (Fig. 6). TEs were detected in each of the seven assemblies, with proportions ranging from 15.6% to 19.3% of a genome's content (Supplemental Figs. S12–S18). The seven genomes contained LINEs, low complexity families, LTR elements, satellites and simple repeats, SINEs. The TE distribution was highly biased along chromosomes with the highest number of TEs detected at telomeres. Their distribution was the opposite of that of established genes (Fig. 1C), which were mainly absent in telomeres. Somewhat unexpectedly, neORFs were distributed regularly along chromosomes, and were also present in telomeres. Therefore, the distribution of neORFs is less biased than the distribution of TEs: high density regions of TEs did not correspond to high density regions of neORFs.

## Homologous sequences of neORFs

We next attempted to identify, for every neORF in every lineage, homologous sequences in those lines in which no neORF (termed "query" hereafter) had been found. We denote as "homologous sequence" a DNA sequence found in a line where the query neORF is not found. A homologous sequence has a high sequence similarity to the neORF, but either lacks some or all elements of an ORF, either is not transcribed, or both. For this, we used a pipeline (see Methods; https://github.com/AnnaGrBio/Proto-gene_emergence) involving nucleotide BLAST and synteny detection. This would enable us to locate the homologous sequence derived from the query neORF and identify sequence differences ("converting mutations"; see also Zhang et al. 2019) which enabled the emergence of the neORF in first place, or the reverse.

First, neORFs shared by all lines were removed from the analysis. A total of 5613 query neORFs (out of 5687 orthogroups, see above) of which 3499 were without introns and 2114 were with introns, were investigated. Each neORF was used as a query to detect homologous sequence in all other lines in which the neORF was missing (searched in an average of 5.5 lines), giving rise to a total of 30,606 homology searches. Restrictive homology searches were performed initially against syntenic regions of the target lines, and then in full genomes when no hit was found. When several hits for a query were found in a single line, only the best nucleotide BLAST hit was kept for further analysis and identified as the "homologous sequence."

Of all neORF queries, 94.9% (DK-FI) to 99.8% (ES-ZI) were found to have homologous sequences in one or several target lines (Fig. 7A), resulting in 27,699 homologous sequences (90.5% of the search). Sequence identities between query neORFs and their target homologous sequence were between 80% and 100%,

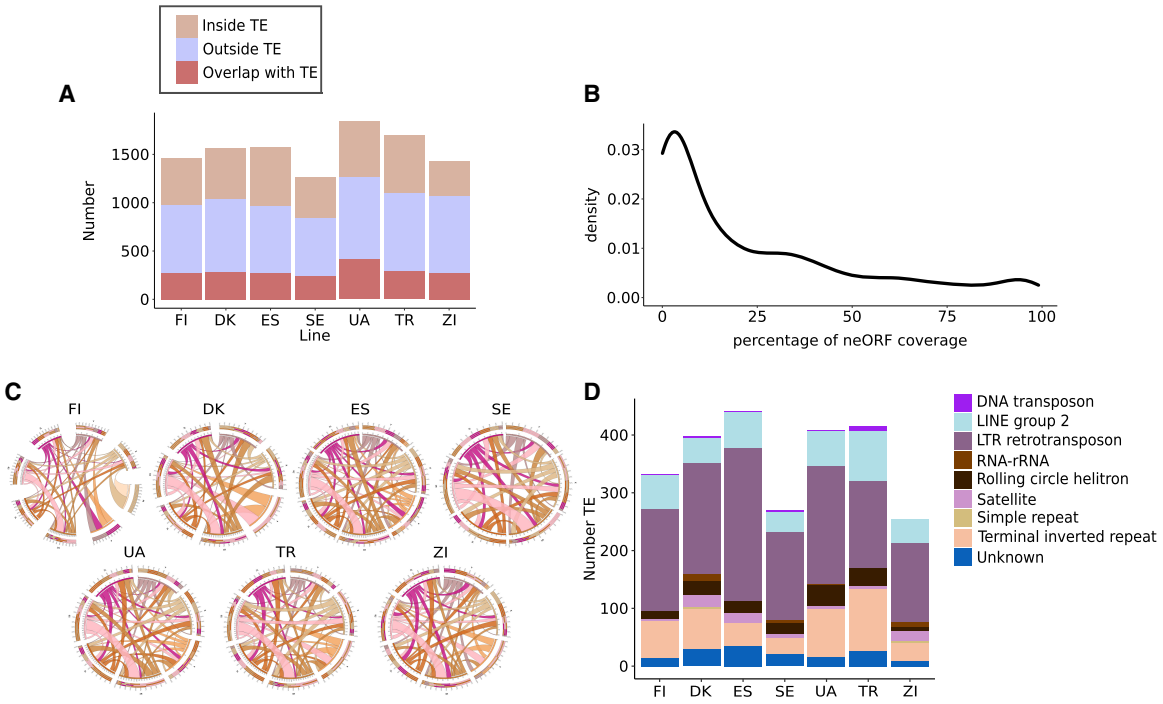


**Figure 5.** neORFs and transposable elements. (*A*) Overlap between neORFs and transposable elements inside lines. The *x*-axis represents the lines and the *y*-axis represents the number of neORFs. A brown color represents neORFs that are located inside of a transposable element; the violet color represents neORFs that do not overlap with a transposable elements; the pink color represents neORFs overlapping partially with a transposable element, but not entirely. (*B*) Percentage of the neORF covered by a TE for neORFs from the category "Overlap with a TE." (*C*) Circle plots representing the genomic location of neORFs duplicated inside a line's genome. A circle's borders represent chromosomes. The circle plots were created using Circos (Krzywinski et al. 2009). (*D*) TE families of TEs overlapping with neORF that are duplicated.

indicating a high reliability in the identification of homologous sequences. Of the alignments between queries and their homologous sequences, 43.1% had a sequence identity ranging from 95% to 100%, based on the Levenshtein ratio, which is based on the Levenshtein distance. This distance measures the difference between two sequences based on the minimum number of mutational events required to change one base to another (Fig. 7B).

## Converting mutations

The 27,699 homologous sequences were aligned to their homologous neORFs, after computationally splicing the neORFs that contained an intron. The alignments were analyzed to determine the converting mutations that distinguish neORFs from their homologous sequences. Among the 27,699 homologous sequences, 23,159 had at least one mutation that would not allow either their transcription, or their putative translation as no ORF was observed, or both. The remaining 4540 had no mutations. Their existence might be explained by the presence of a longer ORF overlapping with them, hiding a smaller ORF (Supplemental Table S26). These homologous sequences were removed from further analyses.

We determined six different features which might explain the different coding characters between queries and homologous sequences: (i) absence of a start codon; (ii) absence of a stop codon; (iii) insertion or deletion causing a frameshift mutation; (iv) presence of a premature stop codon in the first 75% fraction of the sequence or downstream start that would render the encoded protein to be shorter; (v) different sequence coverage in the alignment, causing the end or the beginning of the homologous sequence to be different from the neORF. This may result from, for example, the insertion of a TE in either of the two sequences; and (vi) absence of a detectable transcription event.

Among these six features, the lack of transcription was the most common difference between a neORF and its homologous sequence (Fig. 8). Of the homologous sequences, 85% were not transcribed. Furthermore, 39.5% of the homologous sequences have a frameshift mutation compared to their neORFs, which would result in a major change in the protein sequence and size, if any protein is encoded. A premature stop codon was present in 25.4% of the homologous sequences, resulting in a shorter ORF than their neORF. Finally, 9.2% of the homologous sequences have no stop codon, 9% have no ATG start codon, and 8.8% display a beginning or end of sequence which does not correspond to the neORF (Fig. 8). Most often, several of these features co-occurred in a homologous sequence, explaining why the sum of all percentages exceeds 100.

We next studied the combination of converting features observed in homologous sequences. 60 combinations of these six events were observed, some being represented by one single event. The most frequently observed combination in neORFs was one single event: a single lack of transcription event, which was the case for 11,778 homologous sequences. A frameshift mutation combined with a lack of transcription event was the second most common combination, and was found in 2223 homologous sequences. Furthermore, frameshift mutations combined with a premature stop codon, and frameshift mutations combined with the absence of transcription plus presence of a premature stop codon often occurred together (1059 and 2354 times, respectively). The absence of start and stop codons were the least frequently observed events. Overall, when comparing the presence of an ORF and a detectable transcription event, 50.9% of all homologous

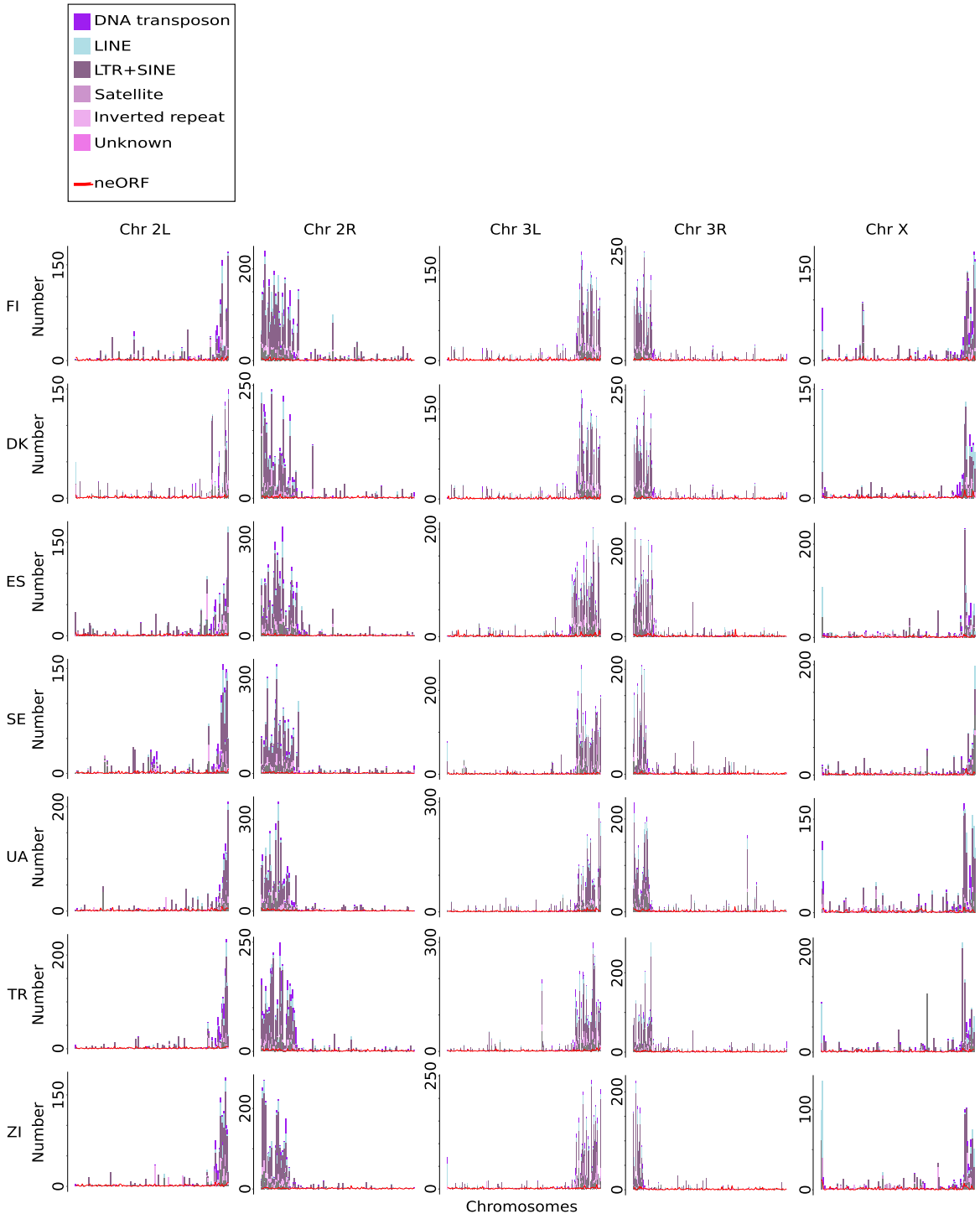


**Figure 6.** Genomic distribution of neORFs and transposable elements. Genomic distribution of TE (bars) and neORFs (red lines) along the five major chromosome arms. The *x*-axis represents chromosomes, in successive intervals of 160,000 bp. The *y*-axis represents the total number of elements (TE or neORFs) present in each interval.

sequences have all of the coding elements in the ORF but are not detected as being transcribed, whereas 14.5% of all homologous sequences are transcribed but do not possess all the coding elements of an ORF. Finally, 34.6% of homologous sequences do not have a coding ORF and are also not transcribed.

We investigated in further detail the nature of the mutations that constitute the difference between neORFs and their homologous sequences (Supplemental Table S26; Supplemental Figs. S19–S23). There were 36% of homologous sequences with no start codon that have not any of the nucleotides of an ATG. The presence

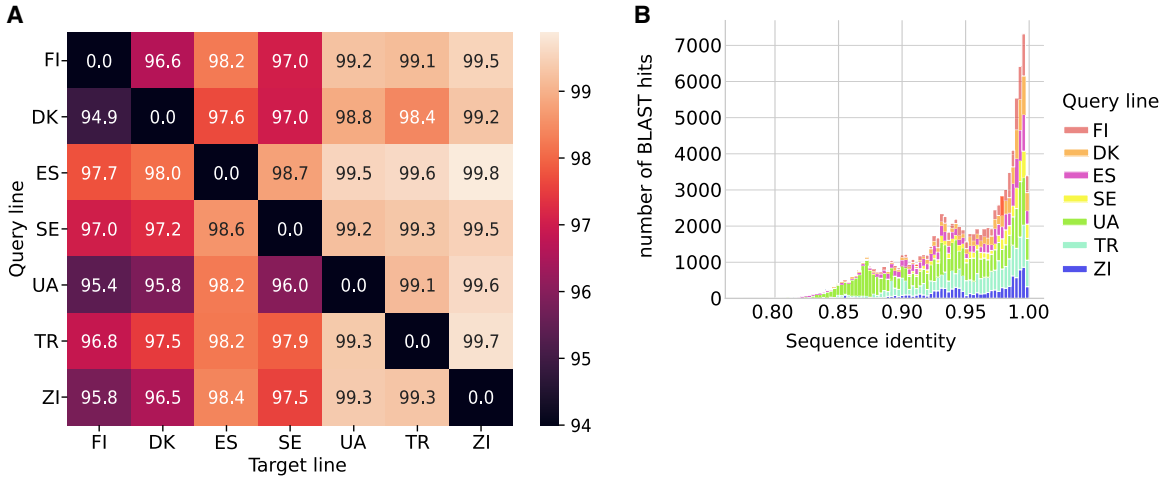


**Figure 7.** Detection of homologous sequences. (*A*) Percentage of nucleotide BLAST hits between query neORFs of each line and target lines. (*B*) Percentage of sequence similarity between neORFs and their homologous sequences, according to the Levenshtein ratio. Each color represents a line.

of two substitutions in the ATG was widely observed (52%). One substitution in the ATG was observed in 20% of all cases. We observed the same trend with the stop codons, in which all three nucleotides constituting the homologous stop codon were substituted in 25% of all cases.

When a frameshift mutation was observed, we studied how many nucleotides were missing or added in the homologous sequence. The most frequent frameshift was because of only one insertion or deletion. Finally, we studied in further detail the absence of a detectable transcription event in homologous sequences. The absence of a transcription event could be attributed to three different situations: (i) no transcript was found; (ii) a transcript was found but in the antisense direction to the ORF; or (iii) a transcript was present (in the forward or reverse direction) but did not cover the entire ORF. Moreover, 50% of the homologous sequences were located in a transcribed sequence, but the transcript was in opposite direction to the predicted ORF, even in cases when all coding elements were present. For 10% of the homologous sequences, a transcription event was observed, but did not cover the length of the entire transcript, in either the forward or reverse direction. Furthermore, 40% of the homologous sequences were not transcribed at all.

## Discussion

In the present study, inbred lines from seven populations of *D. melanogaster* were sequenced and their genomes and transcriptomes were assembled. We detected neORFs and studied their properties and mechanisms of emergence. Our methodology allowed us to detect very young neORFs that emerged within lines of a single species using a common sequencing and annotation framework. This study thus fills the gap between earlier phylogenetic studies that used long evolutionary distances, and population level studies that either used sparsely annotated reference genomes or did not investigate the genetic mechanisms underlying the creation of neORFs. A major result of our work is that neORF emergence is a frequent event, even when viewed from the perspective of populations.

### neORF birth and death

The average number of neORFs found per line (around 1500 neORFs) is in a similar range of numbers to that reported in other species. For example, in yeast, 1900 candidate protogenes were detected in *S. cerevisiae* for a much smaller genome and for much longer evolutionary divergence between species (Carvunis et al. 2012). In humans, Dowling et al. (2020) identified 4429 de novo transcribed ORFs specific to humans. Because most orthogroups (71%) detected in our study contained neORFs specific to a single line, we conclude that there is a rapid turnover within the species. Such a pervasive creation and rapid turnover of neORFs has already been suggested by crossspecies comparisons (Schmitz et al. 2018; Prabh and Rödelsperger 2022) but not yet shown on evolutionarily short time scales. Furthermore, the fairly equal distribution of neORF numbers across lines suggests a neutral birth-death process, at least over such short time scales. The comparability of our numbers with results from crossspecies studies, also including those in *Drosophila* (Heames et al. 2020), indicates that such a neutral regime extends to time scales spanning at least several speciation events. This turnover was further investigated in a follow-up study, where we use a dated tree and a mathematical model that allows us to estimate the birth and death rates of de novo transcripts (Grandchamp et al. 2023).

We detected 66 neORFs that were present in all seven lines and might be considered as established de novo genes. Again, the respective counts are comparable to earlier studies using different approaches: Heames et al. (2020) found 41 putative de novo genes that emerged specifically in *D. melanogaster*, whereas Zhou et al. (2008) reported 72 de novo genes and Zhao et al. (2014) identified 142 segregating and 106 fixed testis-expressed de novo genes in *D. melanogaster*.

Evidence of a translation event would be needed to classify our neORFs as protogenes. Zheng and Zhao (2022) detected evidence of translation of 993 unannotated proteins in *D. melanogaster*. When we searched for homology between our neORFs and the ORFs that they detected, we found that 64 ORFs were common to the two data sets (Supplemental Table S27), all of them being shared by several lines. As we show in this manuscript, most neORFs are line specific. This result suggests that most line-specific neORFs are not yet widely translated. Another explanation would be that our neORFs could not have been detected in Zheng and Zhao (2022), as they emerged by mutations in single lines of *D. melanogaster*. Determining how often the youngest neORFs are translated would be an important further step in understanding the emergence of new genes.

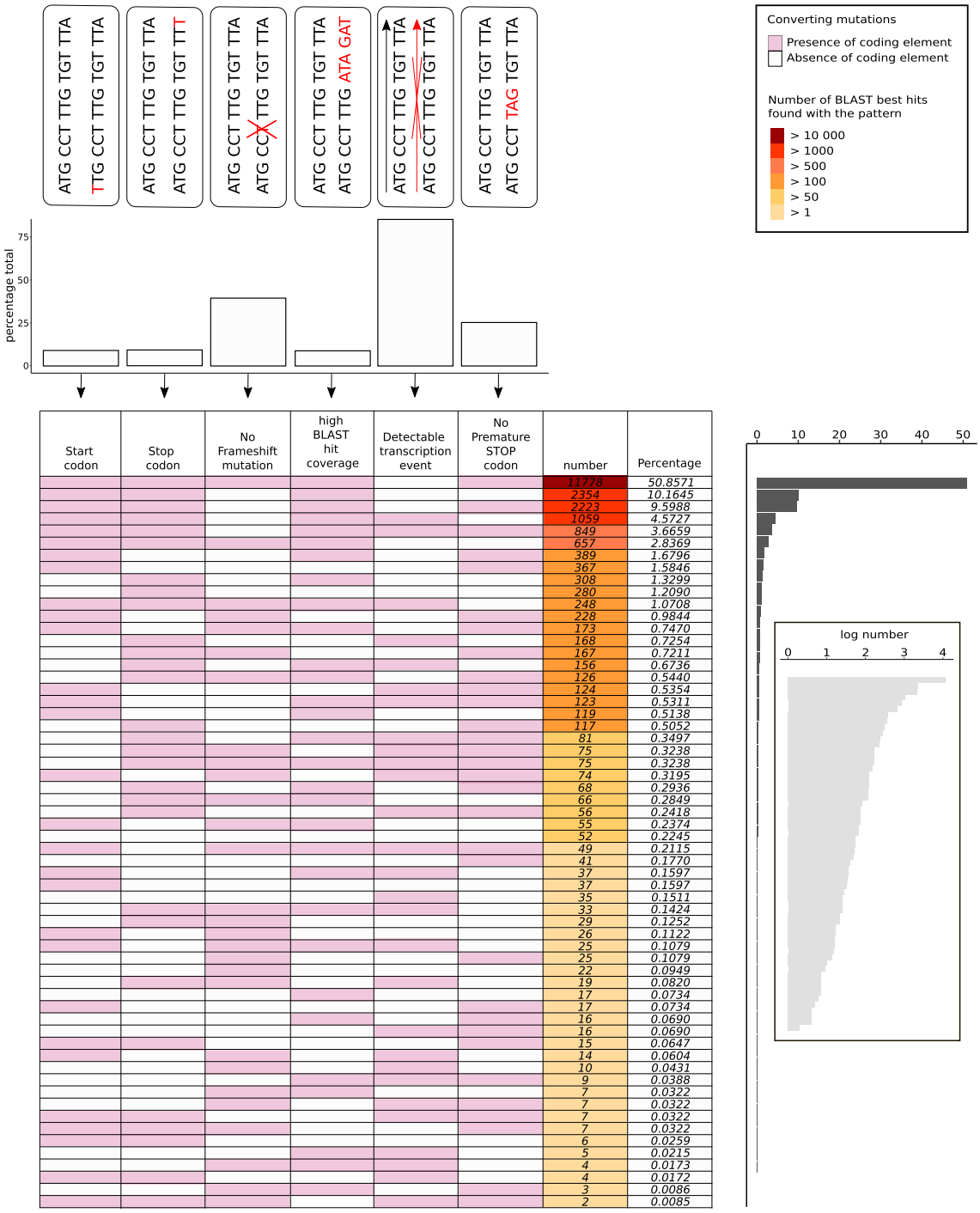


**Figure 8.** Converting mutations between neORFs and their homologous sequences. The figures on the *top* represent the six converting mutation that can occur between a neORF and its homologous sequence. They are generic examples, representing systematically the neORF on the *left* of the alignment and the homologous sequence on the *right*. The barplot on the *top* of the figure represents the percentage (*y*-axis) of six genic features that are not present in the homologous sequences of neORFs. The table represents the percentage of each combination of converting elements. The pink colors indicates that the considered element (ex start codon) is present in the homologous sequence. The gray color indicates that the element is not present in the homologous sequence.

## neORF properties

In each line, around 80% of the detected neORFs emerged in intergenic regions, whereas the next largest group (11%) emerged on antisense of other genes. These results are in line with earlier cross-species studies of humans (Dowling et al. 2020; Grandchamp et al. 2022), which showed that older de novo genes are more often exonic, whereas younger ones are more likely to be found in

intergenic regions. Again, this result strongly suggests a neutral process of neORF emergence and loss in the early stages of their formation. It seems intuitive that neORFs can emerge more easily in noncoding regions, as established genes are constrained by purifying selection (Finseth et al. 2014), whereas noncoding regions are less constrained by selection (Knibbe et al. 2007). Moreover, a neORF overlapping with an existing gene would probably have a lower likelihood of gaining an adaptive function compared to one emerging in a noncoding region.

When comparing properties of line-restricted neORFs to those that are widespread, we found several notable differences. Compared to established genes (older genes annotated in *D. melanogaster*), neORFs were shorter. The average length of a neORF was 250 nt, whereas the average length of coding sequences in established *D. melanogaster* genes is 1830 bp (Moriyama and Powell 1998). We find that, even over such a short evolutionary time scale, fixed neORFs (those present in all seven lines) were longer than less pervasive ones present in only one or a few lines (235 nt vs. 218 nt). In each line, almost half of all neORFs contained introns, with the average number of introns being highest in exonic neORFs, and lower in intergenic and intronic neORFs. This finding strengthens recent reports showing that de novo genes can contain introns (Wu et al. 2011; McLysaght and Guerzoni 2015; Zhang et al. 2019; Grandchamp et al. 2022), and shows that introns can appear at the earliest stages of neORF emergence.

We observed that the computationally predicted amount of intrinsic disorder in putative proteins encoded by neORFs decreased with their pervasiveness, and thus with their relative age, whereas their aggregation propensity increased, even though the level of intrinsic disorder and aggregation propensity were systematically very low. Because this prediction aligns with some studies (Dowling et al. 2020) but disagrees with others (Schmitz et al. 2018), this trend deserves further attention in experimental follow-up studies. However, our findings are consistent with the properties found in the de novo genes of *Drosophila* reported by Heames et al. (2020).

### Converting mutations

Our approach allowed us to track converting mutations and the features associated with neORF birth by mapping them precisely to homologous regions in other genomes. Our homology-based searches retrieved the majority (90.5%) of all homologous sequences with high levels of sequence coverage and identity. Contrary to most earlier studies, these assignments are nonambiguous with near perfect sequence identities. Even for most of the neORFs that have been translocated within chromosomes or to other chromosomes, homologous sequences could be identified, because of the short divergence times under consideration. Whereas earlier studies have reached discordant conclusions regarding the temporal order of gain of transcription versus gain of ORFs (Schlötterer 2015), we here were able to identify and quantify the frequency of features (or combinations thereof) that distinguish neORFs and their homologous sequences. In our study, we discovered that transcription is the feature most frequently missing in homologous sequences. As neORFs are not (yet) fixed, our study cannot predict the most recent ancestral state. Because, by definition, neORFs have no coding homologs, the most parsimonious scenario suggests that these neORFs emerged directly from noncoding DNA. However, a homologous sequence could also have emerged from the mutation in a fixed neORF. The number of neORFs shared between lines can be found in Supplemental Table S28. About half of all homologous sequences have all ORF specific features but lack detectable and/or complete transcription. Of the remaining half of homologous sequences, only 14.5% are fully transcribed, but lack all features of an ORF. About one third have neither a complete ORF, nor are they transcribed. Taken together, these observations suggest that transcription status changes very rapidly. Therefore, the presence of an ORF is not a sine-qua-non prerequisite for neORF emergence, but more often by chance may precede the gain of transcription than the other way around. However, accessing a transcription event by sequencing is more uncertain than determining the presence of an ORF. Transcription is dependent of several conditions, such as stress, environment or age. The lines were maintained and sequenced identically, limiting stress and environmental divergences, but parts of RNA expression may still have been missed in lines.

The rapid turnover of pervasive transcription has been widely documented in recent years (Clark et al. 2011; Wade and Grainger 2014; Neme and Tautz 2016). Underlying mechanisms (for review, see e.g., Kim and Jinks-Robertson 2012) include epigenetic events such as methylation turnover (Ginno et al. 2020), replication and spatial chromatin changes (Candelli et al. 2018), the birth and death of promoters or their interconversion into enhancers (Majic and Payne 2020) and their putative bidirectional functioning (Young et al. 2015). Antisense transcripts occur frequently (Pelechano and Steinmetz 2013) by exploiting transcription loops of genes transcripts. Indeed, among the 17% neORFs overlapping genic sequences in our data set, 15% were antisense to established genes. These data also lay the foundation for further in-depth investigations into the mechanisms of new transcripts formation and the causes of their transient behavior. Furthermore, the here reported abundance of mainly short ORFs in noncoding DNA makes it likely that many are randomly translated. This assumption is based on the study from Ingolia et al. (2014). In this study, the authors show that a very large part of the transcriptome is associated to ribosome activity, even noncoding RNA and UTRs regions. They refer to this translation as a pervasive translation event outside of a coding gene. Such a pervasive translation of non canonical ORFs have later been shown in other studies (e.g., Chen et al. 2020). In conclusion, the more de novo RNAs we find containing an ORF, the more likely some will be translated.

The coding potential of thousands of small ORFs in *Drosophila* has been shown using ribosomal profiling (Aspden et al. 2014; Patraquim et al. 2022), and pervasive transcription has also shown to evolve new functional proteins (Ruiz-Orera et al. 2018). Overall, we observed all possible combinations of features to be involved in the creation of new ORFs, implying that there are many different paths giving rise to new ORFs. 10% of homologous sequences lacked all six features, showing that evolution can quickly build ORFs during population divergence.

We considered the criterion "different size" as one of the five criteria to define the appearance of an ORF. It applies when the end or the beginning of the sequence of the neORF and its homologous sequence diverge completely. This can be because of genome reshuffling between lines, or an error in the genome assembly. It also can be explained by the insertion of a TE.

Finally, among the 27,699 homologous sequences of neORFs, 4540 of them had no mutation, and were masked in our pipeline to detect neORFs by the overlap of a longer ORF. This unexpected result suggests the possible existence of ORFs in alternate reading frames, or at least nested inside a longer ORF. Such an arrangement is quite common in annotated genes (Orr et al. 2020). Understanding the role of these hidden ORFs presents a new

challenge for scientific community (Couso and Patraquim 2017), and many of these ORFs seem to be also translated (Ruiz-Orera et al. 2015). However, it is unexpected to find such a complex structure in putative genes that emerged so recently.

### Transposable elements

Growing evidence has shown that TE insertions enabled the emergence of protein-coding genes or RNAs, and that some of them acquired essential functions over the course of evolution (Naville et al. 2016; Joly-Lopez and Bureau 2018). Among the best known examples of new genes that emerged thanks to transposable elements, are *RAG1* and *RAG2*, which are involved in the vertebrate immune system (Kapitonov and Koonin 2015; Huang et al. 2016), and the syncytin genes of mammals, which contributed to the emergence of the placenta (Malik 2012). When looking at the features enabling neORF emergence from homologous sequences, only 9% of homologous sequence showed strong reshuffling of the end or the beginning of their sequence. This suggests that only 9% of the homologous sequences gained an ORF via the insertion of a TE in their sequence. TEs are thus important, but not the dominant driving force in the creation of neORFs.

However, once emerged, a high number of neORFs were found to be associated with a TE, either by being located inside a TE (25%–30%), or by overlapping with a TE (15%–20%). Indeed, even though TE insertions are not the main events enabling neORF emergence, their presence in genomes may promote other mutations that contribute to neORFs emergence. This has also been suggested by (Lee et al. 2022), who studied presence-absence variants of genes in *C.elegans*, and showed that some of these de novo genes have signatures of active transposons. Concerning the 25%–30% of neORFs which emerged inside TEs, the question is whether to consider them as de novo genes. The proteins translated from these ORFs show no protein BLAST hit to any other known protein in *Drosophila* genomes, indicating that they did not emerge in protein coding TEs. Moreover, as they showed no homology with neORFs in other lines, their emergence in TEs is line-specific, suggesting a mutation of the TE rather than a specific and conserved coding function. Yet, thousands of noncoding RNAs have been shown to be derived from TEs (Lu et al. 2014; Bourque et al. 2018). Understanding the contribution of TEs to de novo gene emergence will be of key importance given their manifest involvement. In cases of neORFs having emerged inside a TE, further investigation would be required to understand if these TEs are still active in transposition, as well as which mechanisms allowed the emergence of an ORF inside them.

TEs are know to be enriched in genomic hotspots, which are often concentrated in telomeres (Venner et al. 2009; Robillard et al. 2016), and this observation was confirmed by our results. TEs that insert in genic regions are much more likely to be shaped by selection (Catlin and Josephs 2022), and may be quickly eliminated from genomic regions that are under strong purifying selection (Bourque et al. 2018). We found that neORFs are distributed regularly across chromosomes. Consequently, neORFs overlapping with TEs are more likely to overlap those more subject to selection than to telomeric TEs, which might explain their high potential of mutation. Activated transcription was found to be the most frequent feature associated to neORF emergence, and can also be because of the proximity of neORFs to TEs, as TEs have been shown to contribute and modify *cis*-regulatory DNA elements and modify transcriptional networks (Lippman et al. 2004; Bejerano et al. 2006; Bourque et al. 2008, 2018; Jacques et al. 2013; Wang et al. 2015; Trizzino et al. 2017).

Finally, among the average 20%–25% of neORFs that were duplicated inside a line, around 80% were located inside of a TE. Thus, TEs appear to play an important role in duplicating neORFs, and are able to translocate them within genomes. The ability of TEs to effect rapid translocation has already been reported at the population level (Bartolomé and Maside 2004; Kofler et al. 2015), but to the best of our knowledge, this is the first time that they have been reported as a duplication vector of neORFs. Most duplicated neORFs overlapped with LTR retrotransposons. LTR retrotransposons have been shown to be the most abundant transposable element in *Drosophila* species, representing 12% of the 20% genomic content in TEs (Mérel et al. 2020). LTR retrotransposons possess the necessary machinery to be active (McCullers and Steiniger 2017). Moreover, it has already been reported that TEs might be active between populations (Sessegolo et al. 2016), and play a role in the divergence of genomes sizes between populations (Vieira et al. 2002; Mérel et al. 2020). This strongly suggests that the duplication of neORFs inside lines was driven by TEs activity. To go further, if would be interesting to detect if TEs overlapping with neORFs are actively transposing inside the line were the duplication occurred.

In the present study, we studied the mutations observed between a neORF and its closest homolog. However, some neORF had several nucleotide BLAST hits in a target genome, suggesting also putative duplications of the homologous sequence. Further studies of emergence, duplication, and transposition of neORF and homologous sequence could give further insight on the role of TEs for the emergence of new genes.

## Methods

### Genetic material

To get a representation of the genetic diversity in Europe, we sequenced the genome of one isofemale line collected from each of six different European populations (Supplemental Table S1), which were sampled by members of the European *Drosophila* Population Genomics Consortium (Kapun et al. 2020, 2021). The sampled populations include Finland (FI), Sweden (SE), Denmark (DK), Spain (ES), Ukraine (UA), and Turkey (TR). We also included one isofemale line from Zambia (ZI), which was collected as part of the *Drosophila* Population Genomics Project (Pool et al. 2012). The Zambian population represents the species' presumed ancestral range in sub-Saharan Africa and is estimated to have diverged from the European populations ~14,000 yr ago (Li and Stephan 2006; Laurent et al. 2011). The lines were established as isofemale lines with five generations of single sib-sib inbreeding, then subsequently maintained in vials with *en masse* sib-sib mating of 50–100 individuals per generation.

### Nucleic acid extraction and sequencing

High molecular weight genomic DNA was extracted from a pool of 50 individuals (mixed female and male) from each line with the DNeasy Blood and Tissue Kit. RNA contamination was removed using chloroform extraction and isopropanol precipitation, and quality control was performed using agarose gels, a UV spectrometer, and NanoDrop measurements. For each line, DNA libraries were prepared with the Ligation Sequencing Kit LSK109 and the Flow Cell Priming Kit EXP-FLP001. Long-read DNA sequences were generated using ONT sequencing, with flow-cells 106D. The reads were generated in FAST5 format, and stored for each

line. The line from Zambia was sequenced twice, because the first round of sequencing did not generate enough reads.

Because our goal was to capture as many novel transcripts as possible, irrespective of their temporal, developmental or spatial expression profiles, we extracted total RNA from a pool of five individuals (two adult males, two adult females, and one larvae) from each iso-female line. To minimize the number of missed de novo transcripts, we aimed at high coverage. Note that missing some weakly expressed transcripts will only slightly influence our results: we are interested in the comparison between stably expressed de novo ORFs, their presumed ancestors and their relation to random sequences and just any ORF with transcription potential (i.e., whether transcribed or not). Total RNA following protocols widely used for *D. melanogaster* (TRIzol-based) (Emery 2007). DNA contamination was removed with DNase treatment, and control quality (agarose gels, UV spectrometer, NanoDrop measurements) was performed. The RNA samples were sent to the Institute of Clinical Molecular Biology of Kiel for sequencing. Libraries were constructed for TruSeq stranded RNA, and RNA was sequenced with RNA NovaSeq 6000 S1 2 × 150 bp, with 92M reads/sample.

## Sequence preparation and de novo genome assembly

For each line, the long DNA reads generated were converted into FASTQ format with guppy base calling from the Oxford Nanopore Technologies' base calling algorithms (https://nanoporetech.com/). Identified doublons were removed from the data sets. Reads shorter than 50 bp were removed from the assembly with Filtlong (https://github.com/rrwick/Filtlong). RNA reads were trimmed with Trimmomatic (0.32) to remove the adaptors AGATCGGAAGAGCAC ACGTCTGAACTCCAGTCA in forward and AGATCGGAAGAGC GTCGTGTAGGGAAAGAGTGT in reverse (Bolger et al. 2014). The quality of reads was assessed with FastQC (0.11.9) (Wingett and Andrews 2018).

For each line, DNA reads were converted into FASTA format with seqtk (https://github.com/lh3/seqtk) (Anaconda package 2015), and assembled with Canu (2.2) (Koren et al. 2017) with 137 Mb as a reference genome size from BDGP6.28 genome in Ensembl (106) (Yates et al. 2020). The quality of the assembly was assessed with BUSCO (5.4.3) with insecta_odb9 as a reference database and *fly* as reference species (Manni et al. 2021).

To improve the quality of the assembly, the contigs were first polished with the RNA reads specific to each line. The RNA reads of each line were mapped on the contigs of their respective assembly (generated previously with Canu), with the spliced aware software STAR (2.7) (Dobin et al. 2013). The SAM files were converted to BAM, sorted and indexed with SAMtools (1.13) programs (Bonfield et al. 2021; Danecek et al. 2021). The resulting genomes were polished with Pilon (1.24) (Walker et al. 2014). The quality of the new genome assemblies was assessed with BUSCO.

Subsequently, the quality of the assemblies was further improved by mapping the long DNA reads of each line to their respective reference genome with minimap2 (2.24) (Li 2018). The resulting SAM files were converted to BAM, sorted and indexed with SAMtools. As the seven lines were inbred for at least five generations, the assembled genomes had low levels of heterozygosity. Nevertheless, we resolved haplotypes with Purge Haplotigs programs using the `readhist`, `contigcov` and `purge` options (Roach et al. 2018). The resulting files were polished with Pilon (https://github.com/broadinstitute/pilon). The quality of the new contigs assembly was assessed with BUSCO. Finally, the contigs were scaffolded with the reference genome BDGP 6.28 of *D. melanogaster*. Scaffolding was performed with Ragstat (2.1) (Alonge et al. 2019). The quality of the assembly was assessed with BUSCO.

## Genome annotation and transcriptome assembly

In the seven genomes, repeats were modeled and classified with RepeatModeler (2.0) (Flynn et al. 2020), and masked with RepeatMasker (4.1.3) (https://www.repeatmasker.org). Homology-based methods were used to annotate the genomes with the software GeMoMa (1.9) (Keilwagen et al. 2019). At this stage, we only aimed to annotate the genes already known and annotated in the reference genome of *D. melanogaster*. Reference genes were used as a support for gene detection and exon/intron structure detection. The BDG release 6.28 of *D. melanogaster* was used as the reference genome with matching annotation data, and TBLASTN (Altschul et al. 1990) was used for homology searches. Pairwise alignments of genomes were performed between lines with Chromeister (Pérez-Wohlfeil et al. 2019). All established genes were retrieved in lines, aligned to the reference annotated genes with MAFFT (7.0) (Katoh et al. 2002) and pal2nal (13.0) (Suyama et al. 2006), and $d_N$ and $d_S$ values were calculated with Python program using Tajima's formula. $d_S$ values correspond to the number of synonymous substitutions per synonymous site, in our case by using pairwise alignments between each gene from lines and their homologous reference gene. Multiple sequence alignments were also performed with MAFFT (7.0) for established genes shared by all lines. The unbiased nucleotide diversity $\pi$ (Korunes and Samuk 2021) was calculated per gene with Python.

For each line, respective RNA reads were mapped on their reference genome with HISAT2 (2.2.1) (Zhang et al. 2021), using the splicing aware option. The resulting SAM files were converted to BAM format with SAMtools, and the BAM files were sorted and indexed. For each line, transcriptomes were assembled with StringTie (Pertea et al. 2015). FASTA, GTF, and GFF files for transcriptome assemblies were retrieved with TransDecoder (https://github.com/TransDecoder).

## Detection / assessment of neORF status

In each assembled transcriptome, the Transcripts Per Million (TPM) value was calculated for each transcript. Following earlier studies and statistical evaluations of transcript persistence (Schmitz et al. 2018; Dowling et al. 2020), transcripts with TPM value inferior to 0.5 were removed. ORFs in the remaining transcripts (trORFs) were detected with the software GetORF from EMBOSS (Rice et al. 2000) with a size ranging from 30 to 3000 amino acids, in forward direction of the transcript. Transcript names and the ORF positions in the spliced transcript were assessed. GetORF considers the end of a transcript as a stop codon, and indeed some false-positive ORFs were detected by the software. These ORFs were removed from the dataset, as only ORFs starting with a start codon and ending with a stop codon were taken into account.

To remove known annotated proteins from the ORF dataset, ORFs were translated into proteins, and used as a query for a protein BLAST search against *Drosophila* proteomes in the forward direction of transcription. Proteomes of the 11 *Drosophila* species available in the Ensembl database were downloaded as BLAST targets: *Drosophila ananassae*, *Drosophila erecta*, *Drosophila grimshawi*, *Drosophila mojavensis*, *Drosophila persimilis*, *Drosophila pseudoobscura*, *Drosophila sechellia*, *Drosophila simulans*, *Drosophila virilis*, *Drosophila willistoni*, and *Drosophila yakuba*. The reference genome of *D. melanogaster* was also included. All tORFs with no hit at an E-value of $\times 10^{-2}$ were stored as putative de novo tORFs. To refine the search, a second protein BLAST was performed against the

proteomes of 15 dipteran species present in Ensembl and covering all dipteran subfamilies: *Anopheles albimanus*, *Anopheles culicifacies*, *Anopheles darlingi*, *Anopheles farauti*, *Anopheles gambiae*, *Glossina austeni*, *Glossina morsitans*, *Lucilia cuprina*, *Lutzomyia longipalpis*, *Mayetiola destructor*, *Megaselia scalaris*, *Musca domestica*, *Phlebotomus papatasi*, *Stomoxys calcitrans*, and *Teleopsis dalmanni*.

All transcripts whose tORF had no BLAST hit were considered as neORFs. For transcripts containing several ORFs with no BLAST hit, only the longest tORF was retrieved and stored as "putative" neORF, following the methodology used by Dowling et al. (2020); Ruiz-Orera et al. (2015) (Supplemental Fig. S24). The genomic positions of the neORFs were determined with custom Python script. Genomic positions of unspliced and spliced neORFs were retrieved by combining the GTF file information and the position in the spliced transcript of the neORFs' ORFs. Each position was verified by reconstructing the ORF in the FASTA genomes. For each putative neORF, the transcript's name, number of exons present, genomics position (of the unspliced transcript), direction of the transcription (forward or reverse), TPM, FPKM, and coverage values were extracted and normalized. The transcripts that overlapped and might represent the alternative splicing of one single gene were also annotated.

BEDTools (Quinlan and Hall 2010) was used to attribute a "Position Category" to each neORF. Five "Position Categories" were established: Overlapping with an intergenic region, overlapping with an intron, inside a gene in frame, overlapping with a gene in antisense direction, and overlapping with a gene and the surrounding intergenic region. For the genomic positions: overlapping with a gene in antisense direction and overlapping with a gene, we used an overlap of 1 nucleotide as a cutoff. The starting point of the gene was defined as the first exon, which includes the 5′ UTR. The ending point was defined as the last exon, which ends with the 3′ UTR.

### Building orthogroups of neORFs between lines

NeORFs were identified in each of the seven lines of *D. melanogaster*. To identify neORFs that were common to several lines, and neORFs that were paralogs in one single line, neORFs from all lines were pooled together and scanned for orthogroups with the software OrthoFinder (Emms and Kelly 2019) (Supplemental Fig. S25).

### Characteristics of neORFs

Size, GC content, and the level of RNA expression (TPM and FPKM) were assessed for each neORF. To investigate if the putative translated proteins had specific properties, we translated neORF. For translated protein sequences, intrinsic disorder was assessed with Iupred (https://iupred2a.elte.hu). Aggregation propensity was assessed with TANGO (Fernandez-Escamilla et al. 2004).

De novo genes detected in the *Drosophila* clade were retrieved from Heames et al. (2020). Their properties were retrieved from their Supplemental Material, and were directly used for comparison with our neORFs as they had been extracted with the same tools. The overlap of neORFs with transposable elements was investigated with custom scripts. Some orthogroups contained several neORFs from the same line, originating from different regions of the genome. We investigated the duplication events within genomes, and mapped the TEs and neORFs distribution for each chromosome.

### Assessing mechanisms of neORF birth

NeORFs absent from at least one line were used as the query to search for homologous sequences lacking ORF and/or transcription in the lines in which they were not detected, in order to study converting mutations to a coding state. A Python pipeline was designed to find these homologs called homologous sequences, and investigate mutations further (Supplemental Fig. S26).

In each query orthogroup, the unspliced sequence of the neORF was retrieved. For orthogroups that contained several orthologs, when the orthologs were not 100% identical, the sequence which showed the highest similarity to the other ones was selected. When the neORF contained an intron, the intronic sequence was lowered in the FASTA file, and the exonic sequences were left in upper letters (Supplemental Tables S29, S30). NeORFs containing an intron and whose unspliced size was >10,000 bp were removed from the analyses, as such long sequences would have biased the homology search. "Query neORF" hereafter refers to neORF under investigation. "Target line" refers to the line in which a homologous sequence is searched.

#### Step 1. Identifying presence and absence of syntenic regions in target lines.

The two established genes neighboring the query neORF were identified in the target line. Corresponding genes were looked for in the GTF file of the target line, and the DNA region between these two genes was stored as "Target syntenic region." Whenever one of the two genes was missing, the second next surrounding gene was used. When a neORF was not surrounded by two genes but only by one gene and the end of the chromosome, the corresponding region was retrieved in the target line. When the two target surrounding genes were not found in syntenic region in the target line, the whole genome was used as a target "Target nonsyntenic region." The average size of the syntenic regions can be found in Supplemental Figures S27, S28.

#### Step 2. Detecting homologous sequences

The unspliced sequence of query neORF was used as a query for nucleotide BLAST homology search against the "Target syntenic region" when present, with a minimal coverage of 60% required. When several hits were found, the best hit was kept as a result. When no hit was found or when no "Target syntenic regions" was detected in Step 1, a second BLAST search was conducted against the whole target genome. NeORFs without hits were annotated.

When a neORF got several hits in a target line, only the best hit was considered for further analysis. Final best nucleotide BLAST hits were referred to as "homologous sequences." One query neORF can have a maximum six "homologous sequences," in a maximum of six outgroups lines.

#### Step 3. Assessing mechanisms of neORF birth

All homologous sequences were aligned to their neORF. For all query neORFs which contained an intron, the intron was removed in the alignment between the neORF and its homologous sequence, in both sequences. Several diverging features were investigated between the aligned neORF and homologous sequences. We investigated the presence of the following six features (Supplemental Fig. S26): (i) start codon, (ii) stop codon, (iii) frameshift mutation, (iv) anticipated stop codon, (v) different sequence hit size, and (vi) transcription event. When the homologous sequences did not align to the entire sequence of the neORF, meaning that the match started later in the sequence and/or ended earlier, the homologous sequence was annotated as "Different sequence hit." When the size of the alignment was the same, the start codon, stop codon, and frameshift mutations were searched

in the sequence. The presence of an anticipated stop codon was searched in the sequence.

Finally, the genomics positions of the homologous sequences were mapped to their reference genome, their orientation in the genome was assessed, and the presence of a transcription event was annotated.

### Programming and analyses

All statistical analyses were performed with R (R Core Team 2020). Data reshuffling, search and analyses were performed with Python (Van Rossum and Drake 1995), and the following modules were used: Biopython (Cock et al. 2009), Matplotlib (3.5.0) (Hunter 2007), numpy (1.20.3) (Harris et al. 2020), pandas (1.3.3) (McKinney 2010), seaborn (0.11.2) (Waskom 2021), gffutils (0.10.1) (http://daler.github.io/gffutils).

## Data access

All raw sequencing data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA929424. All processed sequencing data have been submitted to Zenodo archive (https://doi.org/10.5281/zenodo.7322757). The scripts to reproduce the work presented in this study are available as Supplemental Code and at GitHub (https://github.com/AnnaGrBio/Proto-gene_emergence).

## Competing interest statement

The authors declare no competing interests.

## Acknowledgments

We thank the European *Drosophila* Population Genomics Consortium (DrosEU), especially Maaria Kankare, Marta Pascual, Jessica Abbott, Mads Fristrup Schou, Iryna Kozeretska and Banu Önder for collecting flies, and Josefa Gonzalez and Elio Sucena for providing DrosEU lines to us. We also thank John Pool for providing the line from Zambia and Claudia Fricke for her help with fly breeding and maintenance. We thank Thomas Flatt and Martin Kapun for helpful advice and discussions. We thank Pablo Mirat for constructive feedback on the manuscript. We thank Margaux Aubel for her feedback on the figures. A.G. and E.B.-B. acknowledge funding by Alexander von Humboldt-Stiftung. This work was supported in part by the Deutsche Forschungsgemeinschaft priority program "The genomic basis of evolutionary innovations" (SPP2349; Project No. 503272152 awarded to J.P. [GZ:PA 903/12 -1] and E.B.-B. [GZ:BO 2544/20 -1]).

*Author contributions:* A.G. and E.B.-B. conceptualized the study; J.P. inbred and prepared the lines; A.G. and K.B. extracted the DNA and RNA; A.G. sequenced the DNA; A.G. and M.L. assembled the genomes; A.G. detected neORFs; A.G. and L.K. studied homologous sequences; A.G., E.B.-B., and J.P. validated and curated the data; A.G. and E.B.B. wrote the original draft; A.G., E.B.-B., J.P., M.L., L.K., and K.B. corrected the draft; A.G., J.P., and E.B.-B. acquired funding. All authors have read and agreed to the published version of the manuscript.

## References

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol* **20:** 224. doi:10.1186/s13059-019-1829-6

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

Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MAS, Brocard M, Couso J-P. 2014. Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. *eLife* **3:** e03528. doi:10.7554/eLife.03528

Bartolomé C, Maside X. 2004. The lack of recombination drives the fixation of transposable elements on the fourth chromosome of *Drosophila melanogaster*. *Genet Res* **83:** 91–100. doi:10.1017/S0016672304006755

Begun DJ, Lindfors HA, Thompson ME, Holloway AK. 2006. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics* **172:** 1675–1681. doi:10.1534/genetics.105.050336

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, et al. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila* simulans. *PLoS Biol* **5:** e310. doi:10.1371/journal.pbio.0050310

Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, James Kent W, Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature (London)* **441:** 87–90. doi:10.1038/nature04696

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30:** 2114–2120. doi:10.1093/bioinformatics/btu170

Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, Keane T, Davies RM. 2021. HTSlib: C library for reading/writing high-throughput sequencing data. *GigaScience* **10:** giab007. doi:10.1093/gigascience/giab007

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J-L, Ruan Y, Wei C-L, Ng HH, et al. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* **18:** 1752–1762. doi:10.1101/gr.080663.108

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, et al. 2018. Ten things you should know about transposable elements. *Genome Biol* **19:** 199. doi:10.1186/s13059-018-1577-z

Cai J, Zhao R, Jiang H, Wang W. 2008. *De novo* origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* **179:** 487–496. doi:10.1534/genetics.107.084491

Candelli T, Gros J, Libri D. 2018. Pervasive transcription fine-tunes replication origin activity. *eLife* **7:** e40802. doi:10.7554/eLife.40802

Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and *de novo* gene birth. *Nature (London)* **487:** 370–374. doi:10.1038/nature11184

Catlin NS, Josephs EB. 2022. The important contribution of transposable elements to phenotypic variation and evolution. *Curr Opin Plant Biol* **65:** 102140. doi:10.1016/j.pbi.2021.102140

Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD, et al. 2020. Pervasive functional translation of noncanonical human open reading frames. *Science* **367:** 1140–1146. doi:10.1126/science.aay0262

Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A, et al. 2011. The reality of pervasive transcription. *PLoS Biol* **9:** e1000625. doi:10.1371/journal.pbio.1000625

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25:** 1422–1423. doi:10.1093/bioinformatics/btp163

Couso J-P, Patraquim P. 2017. Classification and function of small open reading frames. *Nat Rev Mol Cell Biol* **18:** 575–589. doi:10.1038/nrm.2017.58

Cutter AD, Garrett RH, Mark S, Wang W, Sun L. 2019. Molecular evolution across developmental time reveals rapid divergence in early embryogenesis. *Evol Lett* **3:** 359–373. doi:10.1002/evl3.122

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10:** giab008. doi:10.1093/gigascience/giab008

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21. doi:10.1093/bioinformatics/bts635

Domazet-Lošo T, Carvunis A-R, Albà M, Šestak MS, Bakarić R, Neme R, Tautz D. 2017. No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Mol Biol Evol* **34:** 843–856. doi:10.1093/molbev/msw284

Dowling D, Schmitz JF, Bornberg-Bauer E. 2020. Stochastic gain and loss of novel transcribed open reading frames in the human lineage. *Genome Biol Evol* **12:** 2183–2195. doi:10.1093/gbe/evaa194
Emery P. 2007. RNA extraction from *Drosophila* heads. *Methods Mol Biol* **362:** 305–307. doi:10.1007/978-1-59745-257-1_20
Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20:** 238. doi:10.1186/s13059-019-1832-y
Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* **22:** 1302–1306. doi:10.1038/nbt1012
Finseth FR, Bondra E, Harrison RG. 2014. Selective constraint dominates the evolution of genes expressed in a novel reproductive gland. *Mol Biol Evol* **31:** 3266–3281. doi:10.1093/molbev/msu259
Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci* **117:** 9451–9457. doi:10.1073/pnas.1921046117
Ginno PA, Gaidatzis D, Feldmann A, Hoerner L, Imanci D, Burger L, Zilbermann F, Peters AH, Edenhofer F, Smallwood SA, et al. 2020. A genome-scale map of DNA methylation turnover identifies site-specific dependencies of DNMT and TET activity. *Nat Commun* **11:** 2680. doi:10.1038/s41467-020-16354-x
Gramates LS, Agapite J, Attrill H, Calvi BR, Crosby MA, dos Santos G, Goodman JL, Goutte-Gattat D, Jenkins VK, Kaufman T, et al. 2022. FlyBase: a guided tour of highlighted features. *Genetics* **220:** iyac035. doi:10.1093/genetics/iyac035
Grandchamp A, Berk K, Dohmen E, Bornberg-Bauer E. 2022. New genomic signals underlying the emergence of human proto-genes. *Genes* **13:** 284. doi:10.3390/genes13020284
Grandchamp A, Czuppon P, Bornberg-Bauer E. 2023. High turnover of *de novo* transcripts in *Drosophila melanogaster*. bioRxiv doi:10.1101/2023.02.13.528330
Gubala AM, Schmitz JF, Kearns MJ, Vinh TT, Bornberg-Bauer E, Wolfner MF, Findlay GD. 2017. The *goddard* and *saturn* genes are essential for *Drosophila* male fertility and may have arisen de novo. *Mol Biol Evol* **34:** 1066–1082. doi:10.1093/molbev/msx057
Harris CR, Millman KJ, Van Der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, et al. 2020. Array programming with NumPy. *Nature (London)* **585:** 357–362. doi:10.1038/s41586-020-2649-2
Heames B, Schmitz J, Bornberg-Bauer E. 2020. A continuum of evolving de novo genes drives protein-coding novelty in *Drosophila*. *J Mol Evol* **88:** 382–398. doi:10.1007/s00239-020-09939-z
Huang S, Tao X, Yuan S, Zhang Y, Li P, Beilinson HA, Zhang Y, Yu W, Pontarotti P, Escriva H, et al. 2016. Discovery of an active RAG transposon illuminates the origins of V(D)J recombination. *Cell* **166:** 102–114. doi:10.1016/j.cell.2016.05.032
Huff JT, Zilberman D, Roy SW. 2016. Mechanism for DNA transposons to generate introns on genomic scales. *Nature (London)* **538:** 533–536. doi:10.1038/nature20110
Hunter JD. 2007. Matplotlib: a 2D graphics environment. *Comput Sci Eng* **9:** 90–95. doi:10.1109/MCSE.2007.55
Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, Wills MR, Weissman JS. 2014. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep* **8:** 1365–1379. doi:10.1016/j.celrep.2014.07.045
Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet* **11:** 97–108. doi:10.1038/nrg2689
Jacques P-É, Jeyakani J, Bourque G. 2013. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet* **9:** e1003504. doi:10.1371/journal.pgen.1003504
Jebb D, Huang Z, Pippel M, Hughes GM, Lavrichenko K, Devanna P, Winkler S, Jermiin LS, Skirmuntt EC, Katzourakis A, et al. 2020. Six reference-quality genomes reveal evolution of bat adaptations. *Nature (London)* **583:** 578–584. doi:10.1038/s41586-020-2486-3
Joly-Lopez Z, Bureau TE. 2018. Exaptation of transposable element coding sequences. *Curr Opin Genet Dev* **49:** 34–42. doi:10.1016/j.gde.2018.02.011
Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20:** 1313–1326. doi:10.1101/gr.101386.109
Kapitonov VV, Koonin EV. 2015. Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. *Biol Direct* **10:** 20. doi:10.1186/s13062-015-0055-8
Kapopoulou A, Pfeifer SP, Jensen JD, Laurent S. 2018. The demographic history of African *Drosophila melanogaster*. *Genome Biol Evol* **10:** 2338–2342. doi:10.1093/gbe/evy185
Kapun M, Barrón MG, Staubach F, Obbard DJ, Wiberg RAW, Vieira J, Goubert C, Rota-Stabelli O, Kankare M, Bogaerts-Márquez M, et al. 2020. Genomic analysis of European *Drosophila melanogaster* populations reveals longitudinal structure, continent-wide selection, and previously unknown DNA viruses. *Mol Biol Evol* **37:** 2661–2678. doi:10.1093/molbev/msaa120
Kapun M, Nunez JC, Bogaerts-Márquez M, Murga-Moreno J, Paris M, Outten J, Coronado-Zamora M, Tern C, Rota-Stabelli O, Guerreiro MPG, et al. 2021. *Drosophila* Evolution over Space and Time (DEST): a new population genomics resource. *Mol Biol Evol* **38:** 5782–5805. doi:10.1093/molbev/msab259
Katoh K, Misawa K, Kuma K-I, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30:** 3059–3066. doi:10.1093/nar/gkf436
Keilwagen J, Hartung F, Grau J. 2019. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol Biol* **1962:** 161–177. doi:10.1007/978-1-4939-9173-0_9
Kim N, Jinks-Robertson S. 2012. Transcription as a source of genome instability. *Nat Rev Genet* **13:** 204–214. doi:10.1038/nrg3152
Knibbe C, Coulon A, Mazet O, Fayard J-M, Beslon G. 2007. A long-term evolutionary pressure on the amount of noncoding DNA. *Mol Biol Evol* **24:** 2344–2353. doi:10.1093/molbev/msm165
Kofler R, Nolte V, Schlötterer C. 2015. Tempo and mode of transposable element activity in *Drosophila*. *PLoS Genet* **11:** e1005406. doi:10.1371/journal.pgen.1005406
Konrad A, Teufel AI, Grahnen JA, Liberles DA. 2011. Toward a general model for the evolutionary dynamics of gene duplicates. *Genome Biol Evol* **3:** 1197–1209. doi:10.1093/gbe/evr093
Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res* **27:** 722–736. doi:10.1101/gr.215087.116
Korunes KL, Samuk K. 2021. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol Ecol Resour* **21:** 1359–1368. doi:10.1111/1755-0998.13326
Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19:** 1639–1645. doi:10.1101/gr.092759.109
Lange A, Patel PH, Heames B, Damry AM, Saenger T, Jackson CJ, Findlay GD, Bornberg-Bauer E. 2021. Structural and functional characterization of a putative de novo gene in *Drosophila*. *Nat Commun* **12:** 1667. doi:10.1038/s41467-021-21667-6
Laurent SJ, Werzner A, Excoffier L, Stephan W. 2011. Approximate Bayesian analysis of *Drosophila melanogaster* polymorphism data reveals a recent colonization of Southeast Asia. *Mol Biol Evol* **28:** 2041–2051. doi:10.1093/molbev/msr031
Lee BY, Kim J, Lee J. 2022. Intraspecific *de novo* gene birth revealed by presence–absence variant genes in *Caenorhabditis elegans*. *NAR Genom Bioinform* **4:** lqac031. doi:10.1093/nargab/lqac031
Levy A. 2019. How evolution builds genes from scratch. *Nature (London)* **574:** 314–316. doi:10.1038/d41586-019-03061-x
Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191
Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet* **2:** e166. doi:10.1371/journal.pgen.0020166
Li C-Y, Zhang Y, Wang Z, Zhang Y, Cao C, Zhang P-W, Lu S-J, Li X-M, Yu Q, Zheng X, et al. 2010. A human-specific *de novo* protein-coding gene associated with human brain functions. *PLoS Comput Biol* **6:** e1000734. doi:10.1371/journal.pcbi.1000734
Li D, Yan Z, Lu L, Jiang H, Wang W. 2014. Pleiotropy of the *de novo*-originated gene *MDF1*. *Sci Rep* **4:** 7280. doi:10.1038/srep07280
Li Z-W, Chen X, Wu Q, Hagmann J, Han T-S, Zou Y-P, Ge S, Guo Y-L. 2016. On the origin of de novo genes in *Arabidopsis thaliana* populations. *Genome Biol Evol* **8:** 2190–2202. doi:10.1093/gbe/evw164
Lippman Z, Gendrel A-V, Black M, Vaughn MW, Dedhia N, Richard McCombie W, Lavine K, Mittal V, May B, Kasschau KD, et al. 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature (London)* **430:** 471–476. doi:10.1038/nature02651
Lu X, Sachs F, Ramsay L, Jacques P-É, Göke J, Bourque G, Ng H-H. 2014. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat Struct Mol Biol* **21:** 423–425. doi:10.1038/nsmb.2799
Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The *Drosophila melanogaster* genetic reference panel. *Nature (London)* **482:** 173–178. doi:10.1038/nature10811
Majic P, Payne JL. 2020. Enhancers facilitate the birth of de novo genes and gene integration into regulatory networks. *Mol Biol Evol* **37:** 1165–1178. doi:10.1093/molbev/msz300
Malik HS. 2012. Retroviruses push the envelope for mammalian placentation. *Proc Natl Acad Sci* **109:** 2184–2185. doi:10.1073/pnas.1121365109

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol* **38:** 4647–4654. doi:10.1093/molbev/msab199

Maze I, Wenderski W, Noh K-M, Bagot RC, Tzavaras N, Purushothaman I, Elsässer SJ, Guo Y, Ionete C, Hurd YL, et al. 2015. Critical role of histone turnover in neuronal transcription and plasticity. *Neuron* **87:** 77–94. doi:10.1016/j.neuron.2015.06.014

McCullers TJ, Steiniger M. 2017. Transposable elements in *Drosophila*. *Mob Genet Elements* **7:** 1–18. doi:10.1080/2159256X.2017.1318201

McKinney W. 2010. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference (SCIPY 2010),* Austin, TX, Vol. 445, pp. 51–56. doi:10.25080/Majora-92bf1922-00a

McLysaght A, Guerzoni D. 2015. New genes from noncoding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philos Trans R Soc Lond B Biol Sci* **370:** 20140332. doi:10.1098/rstb.2014.0332

Mérel V, Boulesteix M, Fablet M, Vieira C. 2020. Transposable elements in *Drosophila*. *Mob DNA* **11:** 23. doi:10.1186/s13100-020-00213-z

Moriyama EN, Powell JR. 1998. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Res* **26:** 3188–3193. doi:10.1093/nar/26.13.3188

Moyers BA, Zhang J. 2016. Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. *Mol Biol Evol* **33:** 1245–1256. doi:10.1093/molbev/msw008

Naville M, Warren I, Haftek-Terreau Z, Chalopin D, Brunet F, Levin P, Galiana D, Volff J-N. 2016. Not so bad after all: retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. *Clin Microbiol Infect* **22:** 312–323. doi:10.1016/j.cmi.2016.02.001

Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent *de novo* evolution. *BMC Genomics* **14:** 117. doi:10.1186/1471-2164-14-117

Neme R, Tautz D. 2016. Fast turnover of genome transcription across evolutionary time exposes entire noncoding DNA to *de novo* gene emergence. *eLife* **5:** e09977. doi:10.7554/eLife.09977

Orr MW, Mao Y, Storz G, Qian S-B. 2020. Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Res* **48:** 1029–1042. doi:10.1093/nar/gkz734

O'Toole ÁN, Hurst LD, McLysaght A. 2018. Faster evolving primate genes are more likely to duplicate. *Mol Biol Evol* **35:** 107–118. doi:10.1093/molbev/msx270

Patraquim P, Magny EG, Pueyo JI, Platero AI, Couso JP. 2022. Translation and natural selection of micropeptides from long noncanonical RNAs. *Nat Commun* **13:** 6515. doi:10.1038/s41467-022-34094-y

Pelechano V, Steinmetz LM. 2013. Gene regulation by antisense transcription. *Nat Rev Genet* **14:** 880–893. doi:10.1038/nrg3594

Pérez-Wohlfeil E, Diaz-del Pino S, Trelles O. 2019. Ultra-fast genome comparison for large-scale genomic experiments. *Sci Rep* **9:** 10274. doi:10.1038/s41598-019-46773-w

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33:** 290–295. doi:10.1038/nbt.3122

Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, Duchen P, Emerson J, Saelao P, Begun DJ, et al. 2012. Population genomics of sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture. *PLoS Genet* **8:** e1003080. doi:10.1371/journal.pgen.1003080

Prabh N, Rödelsperger C. 2022. Multiple *Pristionchus pacificus* genomes reveal distinct evolutionary dynamics between de novo candidates and duplicated genes. *Genome Res* **32:** 1315–1327. doi:10.1101/gr.276431.121

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

Rastogi S, Liberles DA. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol* **5:** 28. doi:10.1186/1471-2148-5-28

R Core Team. 2020. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16:** 276–277. doi:10.1016/s0168-9525(00)02024-2

Rivard EL, Ludwig AG, Patel PH, Grandchamp A, Arnold SE, Berger A, Scott EM, Kelly BJ, Mascha GC, Bornberg-Bauer E, et al. 2021. A putative de novo evolved gene required for spermatid chromatin condensation in *Drosophila melanogaster*. *PLoS Genet* **17:** e1009787. doi:10.1371/journal.pgen.1009787

Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19:** 460. doi:10.1186/s12859-018-2485-7

Robillard É, Le Rouzic A, Zhang Z, Capy P, Hua-Van A. 2016. Experimental evolution reveals hyperparasitic interactions among transposable elements. *Proc Natl Acad Sci* **113:** 14763–14768. doi:10.1073/pnas.1524143113

Rödelsperger C, Prabh N, Sommer RJ. 2019. New gene origin and deep taxon phylogenomics: opportunities and challenges. *Trends Genet* **35:** 914–922. doi:10.1016/j.tig.2019.08.007

Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, Marqués-Bonet T, Albà MM. 2015. Origins of *de novo* genes in human and chimpanzee. *PLoS Genet* **11:** e1005721. doi:10.1371/journal.pgen.1005721

Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Cañas JL, Messeguer X, Albà M. 2018. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat Ecol Evol* **2:** 890–896. doi:10.1038/s41559-018-0506-6

Schlötterer C. 2015. Genes from scratch – the evolutionary fate of *de novo* genes. *Trends Genet* **31:** 215–219. doi:10.1016/j.tig.2015.02.007

Schmitz JF, Ullrich KK, Bornberg-Bauer E. 2018. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat Ecol Evol* **2:** 1626–1632. doi:10.1038/s41559-018-0639-7

Schmitz JF, Chain FJ, Bornberg-Bauer E. 2020. Evolution of novel genes in three-spined stickleback populations. *Heredity* **125:** 50–59. doi:10.1038/s41437-020-0319-7

Sessegolo C, Burlet N, Haudry A. 2016. Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. *Biology Lett* **12:** 20160407. doi:10.1098/rsbl.2016.0407

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34(suppl_2):** W609–W612. doi:10.1093/nar/gkl315

Tan S, Ma H, Wang J, Wang M, Wang M, Yin H, Zhang Y, Zhang X, Shen J, Wang D, et al. 2021. DNA transposons mediate duplications via transposition-independent and -dependent mechanisms in metazoans. *Nat Commun* **12:** 4280. doi:10.1038/s41467-021-24585-9

Trizzino M, Park Y, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, Perry GH, Lynch VJ, Brown CD. 2017. Transposable elements are the primary source of novelty in primate gene regulation. *Genome Res* **27:** 1623–1633. doi:10.1101/gr.218149.116

Vakirlis N, Carvunis A-R, McLysaght A. 2020. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife* **9:** e53500. doi:10.7554/eLife.53500

Van Rossum G, Drake FL Jr. 1995. *Python reference manual*. Centrum voor Wiskunde en Informatica, Amsterdam.

Venner S, Feschotte C, Biémont C. 2009. Dynamics of transposable elements: towards a community ecology of the genome. *Trends Genet* **25:** 317–323. doi:10.1016/j.tig.2009.05.003

Vieira C, Nardon C, Arpin C, Lepetit D, Biémont C. 2002. Evolution of genome size in *Drosophila*. Is the invader's genome being invaded by transposable elements? *Mol Biol Evol* **19:** 1154–1161. doi:10.1093/oxfordjournals.molbev.a004173

Wade JT, Grainger DC. 2014. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nat Rev Microbiol* **12:** 647–653. doi:10.1038/nrmicro3316

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9:** e112963. doi:10.1371/journal.pone.0112963

Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV, et al. 2014. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature (London)* **516:** 405–409. doi:10.1038/nature13804

Wang J, Vicente-García C, Seruggia D, Moltó E, Fernandez-Miñán A, Neto A, Lee E, Gómez-Skarmeta JL, Montoliu L, Lunyak VV, et al. 2015. MIR retrotransposon sequences provide insulators to the human genome. *Proc Natl Acad Sci* **112:** E4428–E4437.

Waskom M. 2021. seaborn: statistical data visualization. *J Open Source Softw* **6:** 3021. doi:10.21105/joss.03021

Weisman CM, Murray AW, Eddy SR. 2020. Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biol* **18:** e3000862. doi:10.1371/journal.pbio.3000862

Weisman CM, Murray AW, Eddy SR. 2022. Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Curr Biol* **32:** 2632–2639.e2. doi:10.1016/j.cub.2022.04.085

Wingett SW, Andrews S. 2018. FastQ Screen: a tool for multi-genome mapping and quality control. *F1000Res* **7:** 1338. doi:10.12688/f1000research.15931.1

Wu B, Knudson A. 2018. Tracing the *de novo* origin of protein-coding genes in yeast. *mBio* **9:** e01024–18. doi:10.1128/mBio.01024-18

Wu D-D, Irwin DM, Zhang Y-P. 2011. De novo origin of human protein-coding genes. *PLoS Genet* **7:** e1002379. doi:10.1371/journal.pgen.1002379

Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, et al. 2020. Ensembl 2020. *Nucleic Acids Res* **48:** D682–D688. doi:10.1093/nar/gkz966

Young RS, Hayashizaki Y, Andersson R, Sandelin A, Kawaji H, Itoh M, Lassmann T, Carninci P, Bickmore WA, Forrest AR, et al. 2015. The frequent evolutionary birth and death of functional promoters in mouse and human. *Genome Res* **25:** 1546–1557. doi:10.1101/gr.190546.115

Zdobnov EM, Bork P. 2007. Quantification of insect genome divergence. *Trends Genet* **23:** 16–20. doi:10.1016/j.tig.2006.10.004

Zhang Y, Romanish MT, Mager DL. 2011. Distributions of transposable elements reveal hazardous zones in mammalian introns. *PLoS Comput Biol* **7:** e1002046. doi:10.1371/journal.pcbi.1002046

Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, Yu Y, Hou G, Zi J, Zhou R, et al. 2019. Rapid evolution of protein diversity by de novo origination in *Oryza*. *Nat Ecol Evol* **3:** 679–690. doi:10.1038/s41559-019-0822-5

Zhang Y, Park C, Bennett C, Thornton M, Kim D. 2021. Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N. *Genome Res* **31:** 1290–1295. doi:10.1101/gr.275193.120

Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science* **343:** 769–772. doi:10.1126/science.1248286

Zheng EB, Zhao L. 2022. Protein evidence of unannotated ORFs in *Drosophila* reveals diversity in the evolution and properties of young proteins. *eLife* **11:** e78772. doi:10.7554/eLife.78772

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in *Drosophila*. *Genome Res* **18:** 1446–1455. doi:10.1101/gr.076588.108

*Received November 15, 2022; accepted in revised form June 6, 2023.*