High turnover of *de novo* transcripts in *Drosophila melanogaster*

Anna Grandchamp[1,†,*], Peter Czuppon[1,†], Erich Bornberg-Bauer[1,2]

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

[2]Department of Protein Evolution, Max Planck Institute for Biology, Tübingen, Germany

[†]these authors contributed equally to this work

[*]Corresponding author: a.grandchamp@uni-muenster.de

Abstract

Most of the transcribed genome in eukaryotes does not code for proteins but produces non-genic transcripts. Among these non-genic transcripts some are newly transcribed, when compared to outgroups, and are referred to as *de novo* transcripts. Despite their role as potential predecessors of *de novo* genes for genomic innovations, little is known about the rates at which *de novo* transcripts are gained and lost. This turnover rate estimation requires a precise comparison of the absence and presence of *de novo* transcripts between phylogenetically close samples and a model to describe transcript gain and loss along the corresponding phylogeny. To detect transcripts that are newly gained on a short evolutionary time scale, we use DNA long reads and RNA short reads from isolines derived from seven populations of *Drosophila melanogaster*. Transcripts from the seven lines were distributed into orthogroups according to three newly proposed definitions. Overall, each line contains between 2,320 and 2,809 unspliced *de novo* transcripts with most of them being specific to a single line. Depending on the definition of transcript orthogroups, we estimate that between 0.13 and 0.34 transcripts are gained per year and that each gained transcript is lost at a rate between 6.6×10^{-5} and 2×10^{-4} per year. This suggests frequent exploration of new genomic sequences mediated by a high turnover of transcripts. Our study therefore provides novel insight on non-genic transcript dynamics on a very short evolutionary time scale with implications for the process of *de novo* gene birth.

Introduction

In most multicellular organisms, only a small fraction of the genome codes for proteins (Sana et al., 2012; Piovesan et al., 2019). Intriguingly though, a large fraction of the non-genic genome is transcribed too, at least occasionally, e.g. under specific conditions such as stress or during development (Papantonis and Cook, 2010; Kim and Jinks-Robertson, 2012; Neme and Tautz, 2016; van Steensel and Furlong, 2019). This production of non-genic transcripts throughout the entire genome has been described as pervasive transcription (Clark et al., 2011; Hangauer et al., 2013; Kellis et al., 2014) and has been demonstrated through several techniques (reviewed in Wade and Grainger, 2014). Many of these non-genic transcripts perform important functions, e.g. rRNA and tRNA are indispensable constituents for the protein assembly by the ribosome (Palazzo and Lee, 2015). Another example are long non-coding RNAs (transcripts with more than 200 nucleotides), as some of them are involved in the regulation of splicing (Romero-Barrios et al., 2018) and transcription in general (Mercer et al., 2009; Ponting et al., 2009; Wang and Chang, 2011; Wang et al., 2011).

Among non-coding transcripts, some are found only in a single species or population without a known ortholog and are therefore called *de novo* transcripts (Neme and Tautz, 2016). Interestingly, many *de novo* transcripts have been found to be bound to ribosomes, indicating that part of their sequence might contain *Open Reading Frames* (ORFs) that are likely translated into small proteins (Aspden et al., 2014; Bazzini et al., 2014; Ruiz-Orera et al., 2014; Couso and Patraquim, 2017; Zhang et al., 2019; Patraquim et al., 2022). This protein-coding potential of *de novo* transcripts makes them important precursors of *de novo* genes, which are genes that arise from previously non-coding genomic regions (Carvunis et al., 2012). Up to now, there is still insufficient understanding of which factors predominantly contribute to *de novo* gene emergence. In particular, it is unclear if, or how often, ORFs are gained before transcription or vice versa (Reinhardt et al., 2013; Schlötterer, 2015; Van Oss and Carvunis, 2019). To understand the evolution and assess the rate of *de novo* gene emergence, it is therefore critical to quantify the evolutionary dynamics of *de novo* transcripts (Albà and Castresana, 2007; Schmitz and Bornberg-Bauer, 2017; Schmitz et al., 2018; Heames et al., 2020).

How quickly *de novo* transcripts are gained and lost is still largely unknown. Studies across multiple species indicate high turnover, i.e. gain or loss, of long non-coding RNAs (reviewed in Kapusta and Feschotte, 2014). For example, Necsulea et al. (2014) estimate that around 10,000 transcripts emerged and fixed *de novo* in primates after the split from rodents around 100 million years (MY) ago, indicating a gain of 100 transcripts per MY. A similar analysis, based on a comparison of transcriptomes from several rodent species, found a gain rate of 5-10 transcripts per MY (Kutter et al., 2012). These high

turnover rates of non-genic transcripts are consistent with another study, which studied multiple rodent species (Neme and Tautz, 2016). In all these studies, transcripts were compared between species separated by large phylogenetic distances. This provides important insight into the transcript turnover in conserved genomic regions. However, a large amount of transcripts could not be compared, because conserved genomic regions only cover a small fraction of the genome because of genome evolution and reshuffling. Moreover, the total turnover of transcripts is estimated without an underlying evolutionary model of transcript gain and loss. For example, the difference between the number of transcripts in rat and mouse genomes can be either due to a loss in the mouse species or a gain in the rat species, which is impossible to distinguish with the existing data (Kutter et al., 2012).

Here, we aim to overcome these limitations by using a tightly controlled setting and very short phylogenetic distances. We use deep sequencing data from seven populations of *Drosophila melanogaster* from different geographic locations. Because of the close evolutionary relationship of these populations, genomic rearrangements are highly unlikely to impact alignments and syntenies. We assembled a genome and a transcriptome for each line, and used comparative genomics approaches to detect newly emerged transcripts in each lines. We additionally determined the genomic location of the *de novo* transcripts and determined orthology with newly proposed transcript-specific orthogroup definitions. The occurrence of transcripts in different populations, not necessarily following a phylogenetic gradient, allows us to reconstruct their evolutionary history and to infer transcript gain and loss rates using the infinitely many genes model (Baumdicker et al., 2010). In contrast to previous studies, these estimates account for the transient evolutionary turnover of transcripts between populations of a single species, instead of solely the fixation (or loss) between different species. Our study provides a more detailed picture of transcript turnover rates, important information to understand the process of *de novo* gene birth.

Results

We aimed to quantify gain and loss rates of *de novo* transcripts based on sequencing data from seven populations of *D. melanogaster*. To this end, we compiled population-specific reference genomes and identified *de novo* transcripts. We propose three definitions of transcript orthology to cluster transcripts from the different populations into orthogroups. We then estimate transcript gain and loss rates from the data sets obtained from these definitions of transcript orthology.

Line-specific genome assemblies

Reference genomes were assembled separately for the seven isolines of *D. melanogaster* obtained from populations at different geographical locations (six from Europe, one from Zambia; Supplementary Information (SI), Section A.1). We mapped long DNA reads of each line to the reference genome of *D. melanogaster* and extracted the seven consensus genomes. The percentage of DNA reads that correctly mapped to the reference genome ranged from 94.3% to 97.42% (SI, Section A.2). RNA reads from each of the seven lines were then mapped to their respective reference genome to build line-specific transcriptomes. Percentages of correct RNA mappings ranged from 88.75% to 93.62% within lines, allowing high-quality reconstruction of transcriptome assemblies (SI, Section A.2). The percentages of SNPs between the aligned genomes of the seven lines ranged from 0.22% to 0.58% (SI, Section A.3), showing very low divergence as expected between lines from populations of the same species. The Zambian line diverged most from the other lines, which is consistent with its geographic separation from the European lines.

De novo transcripts in lines

De novo transcripts were identified in the seven lines of *D. melanogaster* (details in the Methods section). In total, an average of 28,021 transcripts were found per line (Table 1). Among these, an average of 2,901 transcripts per line were detected as *de novo* transcripts, as they showed no detectable homology to transcripts in other Diptera species, including the *D. melanogaster* reference species, and had a required minimal expression threshold of 0.5 transcripts per million reads (Table 1). Among these transcripts, some arose via alternative splicing from a single unspliced precursor. When merging splicing variants as a single transcript, the average number of *de novo* transcripts dropped by 14.5% to an average of 2,481 per line (Table 1).

lines	DK (Denmark)	ES (Spain)	FI (Finland)	SE (Sweden)	TR (Turkey)	UA (Ukraine)	ZI (Zambia)
# transcripts	29,675	27,901	28,212	27,022	27,357	27,786	28,198
# *de novo* transcripts	2,908	2,842	2,708	2,714	2,997	3,024	3,116
# *de novo* unspliced transcripts	2,344	2,417	2,327	2,320	2,529	2,620	2,809

Table 1: **Number of identified transcripts and *de novo* transcripts in the analysed lines**.

De novo transcripts were distributed uniformly across the chromosomes where they emerged (SI, Section A.3) and were found in all chromosomes except the mitochondrial one. According to their overlap with annotated genomic elements, six different genomic positions were defined to characterise the transcripts (details in the Methods section). Interestingly, *de novo* transcript distribution followed a similar pattern in all seven lines (Table 2): Around 60% of *de novo* transcripts overlapped with annotated genes in antisense, around 20% of the transcripts were found to be entirely intergenic, and just very few transcripts (around 3%) emerged inside introns. Except for these intronic transcripts, transcripts from all other genomic positions also overlapped, at least partially, with intergenic regions, making the intergenic region a very important pool of putative emergence of transcription initiation and termination. In addition, 8 to 12 *de novo* transcripts were consistently found to overlap with pseudogenes in each line. These pseudogenic transcripts were exceptionally long (11,000 to 150,000 nucleotides) (Supplemental deposit). Transcripts were also observed to overlap with non-coding RNA (ncRNA), but with a starting site upstream or a termination site downstream of it.

We found 0.23 – 0.48 introns per *de novo* transcript on average, which is less than the 1.8 introns per gene (average) in the reference genome of *D. melanogaster* (Hiller et al., 2009). *De novo* transcripts overlapping with exons in frame (abbreviated by exon longer), or overlapping with genes but in the opposite direction of transcription (antisense), had the highest number of introns (average of 0.48 and 0.38, Table 2). Intronic and intergenic transcripts had the lowest number of introns (average of 0.23 and 0.31).

position	DK	ES	FI	SE	TR	UA	ZI
number of *de novo* transcript							
exon longer	307	355	352	351	362	366	396
intergenic	577	571	537	415	664	731	707
intronic	63	79	69	59	61	106	67
ncRNA	108	111	140	115	131	112	210
pseudogene	11	12	8	10	13	10	12
antisense	1842	1714	1602	1764	1766	1699	1724
average number of introns							
exon longer	0.63	0.44	0.45	0.45	0.46	0.41	0.50
intergenic	0.35	0.36	0.26	0.33	0.30	0.32	0.26
intronic	0.27	0.22	0.17	0.22	0.20	0.19	0.36
ncRNA	0.44	0.27	0.27	0.34	0.32	0.29	0.35
pseudogene	-	-	-	-	-	-	-
antisense	0.42	0.37	0.38	0.36	0.41	0.38	0.34

Table 2: **Number of *de novo* transcripts per line ordered by genomic region and number of introns per *de novo* transcripts.**

Orthogroups of *de novo* transcripts

There is no standard procedure to identify orthogroups of *de novo* transcripts. We therefore propose three new definitions to assign the *de novo* transcripts from the seven lines of *D. melanogaster* into orthogroups. Definition 1 clusters transcripts as orthologs if their spliced sequences share at least 70% of coverage between all the members of the orthogroup and 75% of identity in reciprocal BLASTn. Definition 2 additionally requires that their transcription starting point is similar (within 500 base pairs). Definition 3 clusters transcripts as orthologs if their unspliced sequences share the same transcription starting and ending point, independently of the coverage and identity of the spliced sequences. To assess sequence similarity more explicitly, we also compared the overall number of nucleotides shared by the unspliced transcripts in the seven genomes, following a previous approach (Kutter et al., 2012).

The number of identified orthogroups differed between the definitions (Figure 1). Definition 1 gave the smallest number of orthogroups (9,945), Definition 2 clustered transcripts into 11,305 orthogroups, Definition 3 into 12,223 orthogroups. Definition 1 is solely based on transcript similarity and coverage and, contrarily to Definition 2, does not take into account the start position. Indeed, the fact that Definition 1 defines less orthogroups suggests that around 2,000 transcripts overlap but have a substantially different starting site. Definition 3 defined orthogroups based only on the start and stop position of the unspliced transcripts, without considering splicing, similarity and coverage. The fact that this definition gave the highest number of orthogroups, combined with results from Definitions 1 and 2, suggests that *de novo* transcripts diverge by their initiation and termination location even

though their sequences overlap. In other words, transcripts tend to emerge in close genomic regions between lines, which makes them overlap, but at different transcription starting sites, as they diverge more in their initiation machinery than in their coverage. Interestingly, when comparing the number of transcripts shared by lines, the three definitions showed the same frequency spectra (Fig. 1). Most orthogroups contain only one *de novo* transcript (present in a single line). The number of transcripts shared between lines decreases with the number of lines. This decrease is also visible for each genomic position. Because Definition 1 is the least restrictive, it clusters more transcripts from different lines together. Accordingly, the number of transcripts shared by 2-7 lines was higher in Definition 1 than in Definitions 2 and 3. The relative amount of transcripts specific to a single line represented 53% of the total amount of orthogroups with Definition 1, 66% with Definition 2 and 72% with Definition 3. The results from the alternative approach, which counts the number of nucleotides shared between the different lines, cannot be compared to the three definitions as transcripts are not considered *per se*. Still, the frequency spectrum shows the same pattern as found by the three definitions.

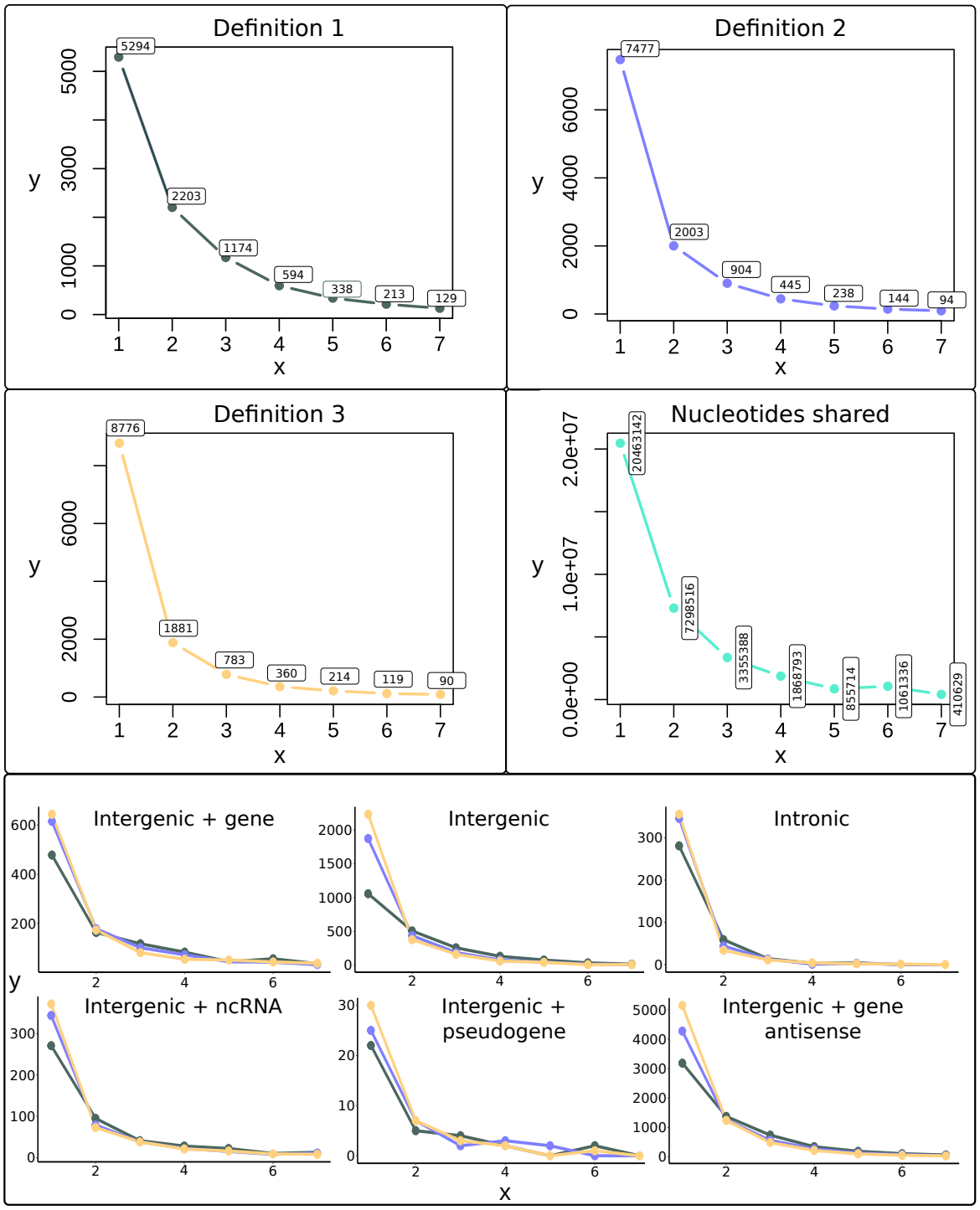

Figure 1: ***De novo* transcripts shared by lines.** The graphs from Definition 1, 2 and 3 represent the number of transcripts shared by the number of lines they are found in. The graph called "Nucleotides shared" shows the number of nucleotides transcribed in common in one to seven lines. The x-axis corresponds to the number of lines sharing a transcript, the y-axis shows the number of transcripts. The six graphs at the bottom show the number of transcripts per genomic position shared between the lines. Each color represents a different definition: Definition 1 in gray, Definition 2 in blue, Definition 3 in yellow.

Phylogenetic tree of *D. melanogaster* populations

To quantify the genomic relationship between the seven lines of *D. melanogaster*, a dated phylogenetic tree was estimated with the software BEAST (Bouckaert et al., 2014). The tree is based on the alignment of the 11,568 longest proteins per genes in common in the seven lines, retrieved from the seven genomes assembled *de novo* (Grandchamp et al., 2022a). The branch lengths are based on a previously estimated date of divergence from European lines to the ancestral African population, here approximated by the Zambian line, and set to 12,843 years (Laurent et al., 2011). The Zambian population is correctly identified as an outgroup population, as expected from its geographic isolation to the European populations (Fig. 2). The closest population to the Zambian in the estimated tree is the line from Spain. Lines from Sweden and Denmark are sister groups, as well as lines from Finland and Ukraine. To check robustness of the tree topology, we also estimated a phylogenetic tree with RAxML (Stamatakis, 2016) and found the same topology (SI, Sections A.4,A.5).

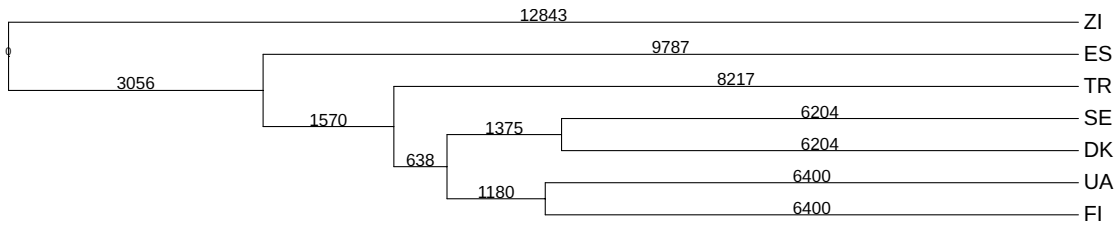

Figure 2: **Dated phylogenetic tree showing the genomic relatedness of the populations.** Branch lengths are proportional to genomic distances measured in years.

Gain and loss rates of *de novo* transcripts

We estimated the gain and loss rates for the transcript frequency spectra obtained by Definitions 2 and 3 (Fig. 1); results for Definition 1 are shown in Table B.4 in the SI. The parameter estimation relies on comparison of a theoretically calculated transcript frequency spectrum and the empirically observed one (details are provided in the Methods section). Briefly, the theoretical prediction is computed with the infinitely many genes model (Baumdicker et al., 2010), which assumes that every transcript can only be gained once, under two different assumptions on the phlyogeny: (i) the phylogeny is a standard coalescent, which arises from a model of neutral evolution in an unstructured population (Baumdicker et al., 2012) (coalescent method), or (ii) the phylogeny is given by the fixed tree estimated in the previous section (Collins and Higgs, 2012) (fixed-tree method). Additionally, we tested two different assumptions on the transcripts. The simpler model assumes that there is a single class of transcripts, i.e., all transcripts are gained and lost at the same rates. The more complex model distinguishes between two classes of transcripts, a class with a high turnover rate, i.e. fast gain and loss, and a class with a (relatively) low turnover rate, i.e. slow gain and loss. Lastly, we account for transcripts that have become fixed in all (sampled) populations and that cannot be lost anymore, i.e., their loss rate is equal to zero. The number of these fixed transcripts is denoted by C_{fixed}. Overall, we therefore have four different models with either three or five parameters to estimate: coalescent and fixed-tree phylogeny with one or two classes of transcripts. The estimated parameters are summarized in Table 3. To assess the robustness of the parameter estimation with respect to the Zambian outgroup, we conducted the analysis with data from all the populations or just from the European populations.

We find high gain and loss rates of transcripts. We estimate that between 0.13 and 0.34 new transcripts are gained per year, depending on the transcript orthology definition and estimation method, and every single transcript is lost at a rate of around 10^{-4} per year, i.e., the expected life span of a transcript is approximately 10,000 years. These rates are robust with respect to the analysed data sets (all populations or only European populations). We note that the parameter transformation to rates per year in the coalescent method strongly depends on the assumptions of the generation time, here assumed to be two weeks (Fernández-Moreno et al., 2007), and the effective population size, here set to 900,000 (Laurent et al., 2011). Uncertainty in these parameters can strongly impact the per year rate estimates of the coalescent method because the untransformed parameters are multiplied with the factor: *no. of generations per year*/(2× *effective population size*). However, we think that the overall magnitude should be consistent with the fixed-tree estimates, which seems to be the case with our chosen parameterization.

dataset	method	gain	loss	gain (slow)	loss (slow)	C_{fixed}
Definition 2						
all populations	1 class – coalescent	0.14	5.3×10^{-5}	–	–	71
Europe	1 class – coalescent	0.13	5×10^{-5}	–	–	91
all populations	2 classes – coalescent	0.18	2.1×10^{-4}	0.06	3.1×10^{-5}	41
Europe	2 classes – coalescent	0.15	2×10^{-4}	0.06	3.1×10^{-5}	50
all populations	1 class – fixed tree	0.23	8.9×10^{-5}	–	–	88
Europe	1 class – fixed tree	0.23	9.3×10^{-5}	–	–	100
all populations	2 classes – fixed tree	0.23	9.4×10^{-5}	2.1×10^{-3}	1.1×10^{-5}	0
Europe	2 classes – fixed tree	0.23	9.3×10^{-5}	1.1×10^{-4}	1×10^{-6}	0
Definition 3						
all populations	1 class – coalescent	0.18	6.6×10^{-5}	–	–	75
Europe	1 class – coalescent	0.17	6.3×10^{-5}	–	–	95
all populations	2 classes – coalescent	0.34	3.2×10^{-4}	0.05	3.2×10^{-5}	50
Europe	2 classes – coalescent	0.28	2.6×10^{-4}	0.05	3×10^{-5}	52
all populations	1 class – fixed tree	0.26	9.4×10^{-5}	–	–	90
Europe	1 class – fixed tree	0.27	1×10^{-4}	–	–	100
all populations	2 classes – fixed tree	0.26	1×10^{-4}	2.2×10^{-3}	1.1×10^{-5}	0
Europe	2 classes – fixed tree	0.27	1×10^{-4}	6.4×10^{-4}	5.1×10^{-6}	0

Table 3: **Estimated *de novo* transcript gain and loss rates.** The gain and loss rates are measured as rates per year, the parameter C_{fixed} is a number of transcripts. The details of the parameter transformation between the coalescent time scale and the 'per year' time scale are stated in the Methods. Briefly, we assume an effective population size of 900,000 for *D. melanogaster* in Europe (Laurent et al., 2011) and a generation time of two weeks (Fernández-Moreno et al., 2007), which gives 26 generations per year. The untransformed numerical estimates are stated in Table B.4 in the SI.

In the models with a fixed-tree phylogeny, the assumption of two transcript classes does not improve the fit to the data (compare star- and point-shaped symbols in Fig. 3). In fact, we find that the slow gain and loss rates replace the estimate of the fixed number of transcripts (C_{fixed}), which becomes zero in these models (Table 3). In contrast, the fit with two transcript classes improves the fit in the models based on a coalescent-like phylogeny (compare upward and downward triangles in Fig. 3).

In general, the coalescent method produces a better fit of the frequency spectrum than the presumably more accurate fixed-tree method. A similar observation has been made in the context of bacterial populations (Collins and Higgs, 2012). With our data, the worse fit of the fixed-tree compared to the coalescent method is exclusively due to the transcripts unique to a single population and transcripts shared by two populations. These numbers are most strongly influenced by external branches leading to the extant populations. In our estimated phylogeny (Fig. 2), which is the fixed tree, these external branches are long compared to the internal branches, especially for the most recently diverged populations (Sweden, Denmark, Ukraine and Finland). This is fundamentally different in a 'standard' coalescent tree, i.e. Kingman's coalescent, where the terminal branches are shorter (on

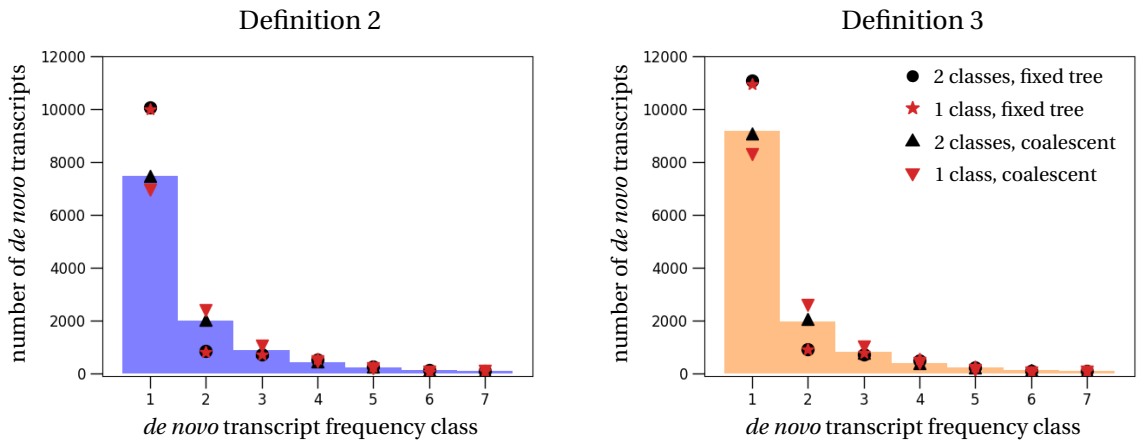

Figure 3: **Empirical and estimated transcript frequency spectra.** The histograms show the empirical *de novo* transcript frequencies obtained with Definition 2 (left) and Definition 3 (right). Symbols show the theoretical frequency spectrum computed with the parameters estimated by the different methods: black circle – 2 classes, fixed tree; red star – 1 class, fixed tree; black triangle – 2 classes, coalescent; red triangle (upside down) – 1 class, coalescent.

average) than the internal branches. The very long external branches of our fixed tree, which are a result of long times of isolation between populations, lead to a large number of transcript singletons, i.e. transcripts unique to a single population. Moreover, in our specific tree, the most recently diverged populations (Sweden, Denmark, Ukraine and Finland) split around the same time, which results in a star-like phylogeny when considering these populations only. This leaves little time for a transcript to arise in an ancestral population of exactly two extant populations, which is likely the reason why the number of transcripts shared by exactly two lines is underestimated.

Transcript gain and loss rates per genomic region

We estimate the gain and loss rates separately for different genomic regions, using the same methodology as outlined in the previous section. Table 4 shows the estimated parameters (rates per year) for the one class, fixed-tree method with the data set including all populations. We find that the loss rates of transcripts are estimated at around 10^{-4} across all genomic regions and also for both definitions underlying the data sets, which is consistent with the estimate from the aggregated data. The gain rates, in contrast, differ more strongly. One reason for this large variation might be that the genomic regions do not cover equal proportions of the genome. For example, intergenic regions cover around 68% of the *D. melanogaster* genome, while we approximate that the regions where transcripts overlap with non-coding RNAs only cover around 9% of the genome (details in the SI, Section B.2). Larger genomic coverage should therefore also result in a larger gain rate estimate because there are more positions where a *de novo* transcript could arise. To compare the gain rates in a meaningful way,

genomic region	Definition 2				Definition 3			
	gain	$\text{gain}_{\text{norm}}$	loss	C_{fixed}	gain	$\text{gain}_{\text{norm}}$	loss	C_{fixed}
exon longer	0.02	0.13	6.7×10^{-5}	30	0.03	0.2	7.2×10^{-5}	41
intergenic	0.06	0.09	9.8×10^{-5}	8	0.07	0.11	1×10^{-4}	15
intronic	8.5×10^{-3}	0.04	1.2×10^{-4}	2	0.02	0.09	1×10^{-4}	6
ncRNA	9.7×10^{-3}	0.11	8.3×10^{-5}	11	0.02	0.22	9×10^{-5}	14
antisense	0.14	0.93	9×10^{-5}	38	0.16	1.07	1×10^{-4}	38

Table 4: **Estimated *de novo* transcript gain and loss rates per genomic region.** We have used the one class, fixed-tree method to estimate the gain and loss rates of transcripts. Parameter estimates of gain and loss rates are per year. To compare the gain rates between different genomic regions in a meaningful way, we have normalized the gain rates, denoted by $\text{gain}_{\text{norm}}$, according to the coverage of the region in the genome. Normalization is done by rescaling the gain rates by their respective estimated coverage (details in SI, Section B.2).

we normalize them according to their respective coverage. Strikingly, we consistently find that the normalized gain rate of antisense transcripts is almost ten times larger than of all the other regions and that transcript gain in intronic regions is consistently the lowest.

The estimated transcript gain and loss rates from the coalescent-based method are stated in SI, Table B.2. The general pattern remains the same, i.e., we find consistent loss rates across regions and the lowest gain rate in intronic and the highest gain rate in antisense regions.

Discussion

De novo transcripts in the seven lines of *Drosophila melanogaster*

We investigated the emergence of *de novo* transcripts by using a unique setup based on seven inbred lines derived from seven geographically different populations of *Drosophila melanogaster*. In each line, between 2,708 and 3,116 transcripts showed no homology to any other annotated transcript in Diptera (including *D. melanogaster* annotated transcripts), suggesting their *de novo* emergence. Some of these detected *de novo* transcripts are the result of alternative splicing, reducing the amount of unspliced *de novo* transcripts to 2,327–2,809 transcripts per line. In total, their cumulative length covers 4–6% of the genome. We find that gain of a *de novo* transcript is a frequent event. Previous studies already detected high amounts of new transcripts when comparing species or expression levels in different organs of the same species. For example, Brown et al. (2014) identified 1,875 new candidate long non-coding RNAs (lncRNAs) producing 3,085 transcripts in *D. melanogaster*, with 2,990 of them having no overlap with protein-coding genes of *D. melanogaster* and known lncRNAs in outgroup species. Huang et al. (2015) determined that 4.5 to 6.7% of transcripts detected in the transcriptome of *D. melanogaster* were un-annotated in flybase, which amounted in 1,669 transcripts derived from intronic regions and 2,192 from intergenic regions. These previous studies focused on species differences rather than population differences and included data from several developmental stages. Our results came from *D. melanogaster* lines, with RNA from larvae and adults, which might explain why we detected fewer transcripts than previous studies. We particularly detected much fewer transcripts in intronic regions, which could suggest that intronic *de novo* transcripts are specific to some developmental stages, or that they are more likely to become fixed in the species.

Strikingly, we find that in all lines most *de novo* transcripts (around 60%) overlap with coding genes in opposite direction. These antisense transcripts are a common phenomenon in genomes (Katayama et al., 2005). Their functions and mechanisms of emergence were reviewed in Barman et al. (2019). For example, antisense transcription plays an important role in gene expression regulation, as antisense transcripts can hybridize with forward transcripts and prevent translation of the forwardly transcribed transcript (Pelechano and Steinmetz, 2013). Despite their importance in gene regulation, *de novo* emergence of antisense transcripts has not yet been intensively studied. Our results suggest that *de novo* emergence of antisense transcripts is common on the population level. As most of the antisense transcripts are specific to one line, their emergence could play a role in regulating genes that are under constraints in the specific environment of the population, which they emerged in.

Unexpectedly, splicing was observed in some *de novo* transcripts. Still, the number of introns and of alternative splicing was on average low and smaller than the average number of introns from protein-coding genes. lncRNAs are expected to become spliced by the same mechanisms as protein-coding genes (Krchňáková et al., 2019; Quinn and Chang, 2016). It is interesting to see that *de novo* transcripts are already capable of splicing, despite being newly emerged. Particularly, *de novo* transcripts that are antisense and overlapping with an existing gene in frame had the highest number of introns on average. Introns from the genes to which they overlap could thus putatively be reused by antisense transcripts. This would also explain why intronic and intergenic *de novo* transcripts had, in contrast, the lowest number of introns.

High rates of gain and loss suggest high transcript turnover

Approaches to determine orthology of protein-coding genes are either based on phylogeny, or on graph-based methods that rely on sequence similarity (Gabaldón, 2008; Kuzniar et al., 2008; Kristensen et al., 2011; Ambrosino and Chiusano, 2017). However, as *de novo* transcripts are not associated to known functions, a partial overlap of two transcripts does not have the same meaning as the overlap of two coding sequences. Moreover, two transcripts from distinct lines with sequences partially overlapping can have different transcription initiation and termination sites, suggesting different sources of emergence. We distributed transcripts into orthogroups according to three different definitions that build on different interpretations of orthology. All definitions resulted in similar patterns of transcripts shared across populations (Fig 1).

Independent of the definition, we find a high turnover rate of transcripts. Using the data from Definitions 2 and 3, the gain rate is estimated to be about 0.23 transcripts per year and a transcript is lost at a rate of 10^{-4} per year. Interestingly, the definition did not affect substantially the predicted rates of gain and loss. However, Definition 2 predicted less orthogroups than Definition 3, suggesting that new transcripts diverge more in their initiation and termination machinery than in their coverage. The high turnover of transcript gain and loss seems to be strongly influenced by gain and loss of transcription starting and termination sites. This result is consistent with other studies (Ward and Kellis, 2012; Wade and Grainger, 2014; Young et al., 2015). For example, high turnover of starting sites for transcription has been observed between human and mice (Frith et al., 2006), together with high reshuffling in upstream regulator regions (Brown and Feder, 2005; Odom et al., 2007; Wittkopp and Kalay, 2012; Cotney et al., 2013; Ballester et al., 2014; Vierstra et al., 2014; Villar et al., 2014, 2015). In *Bacillus subtilis*, 174 transcription units were found to have a new termination signal in

the genome, further down than the original termination site, which had been inactivated (Nicolas et al., 2012). In fact, several studies in yeast and other eukaryotes demonstrated that modifications in transcription termination were involved in the abundant production of non-genic transcripts (Grosso et al., 2015; Rutkowski et al., 2015; Vilborg et al., 2015). To maintain constant transcript numbers over time, the large gain rate of transcription is countered by different transcript removal mechanisms. These mechanisms can be complementary and occur at different levels (Wade and Grainger, 2014; Lasa et al., 2011; Singh et al., 2014). For example, *de novo* transcripts can be directly degraded in the nucleus or the cytoplasm, as it has been shown in *Saccharomyces cerevisiae* (Jacquier, 2009; Porrua and Libri, 2015; Candelli et al., 2018b). In addition, transposable elements alter the landscape of pervasive transcription by pausing or terminating neighbouring transcription (Candelli et al., 2018a).

Rates of *de novo* transcript gain vary among genomic regions

We classified *de novo* transcripts according to their genomic position. The proportion of transcripts found per genomic region was similar for all lines (Table 2). While the loss rate of *de novo* transcripts was consistent across all genomic regions, we found differences between the gain rates (Table 4). The normalized gain rate was the highest for *de novo* transcripts overlapping with genes in the opposite direction of their transcription (antisense). This finding is in line with studies showing that antisense transcription is frequent and is a major driver of evolution as it regulates genes expression (Katayama et al., 2005; He et al., 2008). Moreover, antisense transcripts can also be involved in diseases (Barman et al., 2019), and their frequent gain and loss could be adaptive in lines facing different environments. We also consistently find that the gain rate in intronic regions is lowest. Gain of a new intronic transcript might require both the acquisition of a starting and an ending site inside the intron, contrarily to the other transcripts which could exploit an already existing site. The gain of these two elements has to occur in a region limited in size (the intron), which might explain why gain of transcript inside an intron is less expected.

More unexpectedly, the gain of transcripts in intergenic regions or of transcripts overlapping with a gene occurs at a similar rate. We would have expected that transcripts emerging in overlap with a coding region have a higher chance to be removed by purifying selection (Bourque et al., 2018). Transcripts overlapping to existing genes could, however, have a higher chance to acquire a coding function and by that become part of the gene splicing process, which might explain our results. Moreover, intergenic regions are more subject to mutations (Liu and Zhang, 2022), which could quickly delete any newly emerged transcript, as the acquisition of a function in non-coding regions might be a slower process

than in regions containing (positively) selected exons.

Mutation rate varies across the genomic regions according to different selective constrains (Liu and Zhang, 2022), which could affect the varying transcript gain and loss rates. Yet, we find that the transcript loss rate seems to be relatively constant across all the genomic regions. Therefore, under the assumption that mutations are the predominant mechanism of transcript loss, the constant loss rate across the genome questions the potential influence of mutation rate variation on transcript gain rates. Transposable elements have also been shown to be involved in reshuffling transcription initiation and termination (Feschotte, 2008). In the seven lines, transposable elements are located mainly in the telomeres, whereas *de novo* transcripts have a continuous distribution along the chromosomes (SI, section A.3). However, further studies about whether or not transposable elements overlap with *de novo* transcripts should provide more information about their putative involvement.

Limitations of the estimation procedure

The gain and loss rate estimation is based on the comparison of the empirically observed transcript frequency spectrum and a theoretical prediction, either obtained from coalescent theory or using a fixed phylogenetic tree. In line with a study about gene gain and loss in bacteria (Collins and Higgs, 2012), we find that the coalescent-based estimates produce a better fit than rate estimates from the fixed tree approach. This is surprising because the fixed tree should be a more appropriate representation of the evolutionary history of the studied *D. melanogaster* populations than a coalescent tree. One improvement to our model would be to explicitly account for gene flow between populations. This would have two effects. First, it would change the branch length, possibly even the structure, of the phylogenetic tree of the seven lines. For example, accounting for gene flow Kapopoulou et al. (2020) found a much more recent divergence time between the ancestral African and European *D. melanogaster* populations. Second, the inclusion of gene flow in the infinitely many genes model would change the theoretical prediction of the frequency spectrum. We suspect that this would likely improve the fit of the fixed-tree method because gene flow reduces the number of transcripts predicted to be found in a single population, which is the frequency class that is strongly overestimated by the theory.

In fact, gene flow has been included in both coalescent (Baumdicker and Pfaffelhuber, 2014) and fixed-tree (Zamani-Dahaj et al., 2016) models already. We have applied the coalescent method with gene flow and did not find substantial differences to the parameter estimates without gene flow (details in SI, Section B.4). As for the fixed-tree method, Zamani-Dahaj et al. (2016) have assumed an

unobserved source population from which transcripts (the authors study genes) are transmitted to the observed population. Instead of assuming an unobserved source population, which is a reasonable approach for bacteria that can incorporate genes from the environment, we believe that explicitly accounting for population structure based on the geographical proximity of our populations would be more sensible in our setting. We therefore refrained from applying their approach to our data, but stress that including population structure and gene flow into the fixed-tree method would be a useful future extension.

Conclusion and future prospects

To summarize, we have estimated gain and loss rates of *de novo* transcripts from seven *D. melanogaster* lines. These estimates show that *de novo* transcripts are gained and lost at high rates. Gain rates seem to vary across the genome, being higher in antisense and lower in intronic regions. In contrast, loss rates of *de novo* transcripts seem to be similar across the genome.

While the new data provided in this study are obtained from a tightly controlled setting using inbred lines and consistent protocols for assembly and annotation of genomes and transcriptomes, future broader data sets could still help refine or confirm generality of our findings. For example, multiple genomes per population rather than using a single genome only would improve the estimates. Second, including more populations would help alleviate the problem of population structure and gene flow in the construction of the tree. Third, the date of divergence from the Zambian population was set to $\sim$13,000 years according to Laurent et al. (2011). However, this time was found to be around 4,000 years according to Kapopoulou et al. (2020). Indeed, the support of future studies to determine the most accurate date of divergence between Drosophila lines will help adjusting the model more precisely. Finally, including genomes from several populations from species that follow a phylogenetic gradient would enable modelling gain-loss dynamics beyond species boundaries. These results would, in turn, provide insight into the fixation probabilities of transcripts under changing environmental conditions and after going through population bottlenecks.

In the broader context of genome evolution, our results are a first step to a more mechanistic and less phenomenological treatment and understanding of *de novo* transcript, and consequently *de novo* gene, evolutionary dynamics.

Methods

Line-specific reference genomes

RNA short reads from whole genome illumina sequencing and DNA long reads from whole genome nanopore sequencing of seven isofemale lines of *D. melanogaster* were downloaded from NCBI (accession PRJNA929424). Among the seven lines, six come from European populations: Finland (FI), Sweden (SE), Denmark (DK), Spain (ES), Ukraine (UA) and Turkey (TR), and one is from Zambia (ZI) (SI, section A.1) The RNA seq samples were built based on two males, two females and one larvae (Grandchamp et al., 2022b) pooled together. The DNA was extracted from 50 individuals per line. The line from the Zambian population is considered as the ancestral outgroup as the *D. melanogaster* from sub-Saharan Africa diverged from the European populations around 12,843 years ago (Li and Stephan, 2006; Laurent et al., 2011).

Reads shorter than 100 base pairs (bp) were removed from the long DNA reads by using Fitlong (github rrwick/fitlong). We used the genome of *D. melanogaster* BDGP6.28, downloaded from Ensembl (Yates et al., 2020), as a reference genome. Usually, when working with populations, RNA reads from each lines are directly mapped to a reference genome of the studied species, or the closest one available. Instead, here we first built a specific reference genome for each of the seven lines by mapping the DNA to the reference genome and extracting the consensus genome. This approach allowed us to detect genomic polymorphisms between lines, and to increase the precision for mapping line-specific transcripts to the genome. Accordingly, the DNA reads of each line were mapped to the reference genome of *D. melanogaster* using BWA-MEM (Houtgast et al., 2018). The resulting SAM files were converted into BAM with Samtools view (Danecek et al., 2021). The BAM files were sorted and indexed with Samtools-sort and Samtools-index. Mapping statistics were obtained with Samtools-flagstats (SI, Section A.2), and alignments were visualised with Integrative Genome Viewer (IGV) (Thorvaldsdóttir et al., 2013). This procedure mapped 95-98% of the DNA to the reference genome. For each line, a consensus genome was extracted, by using Samtools-mpileup (Ramirez-Gonzalez et al., 2012), bcftools (Narasimhan et al., 2016) and its function vcfutils.pl (Afgan et al., 2016) (Supplemental deposit). The genomic regions that were not covered by mapping were completed by the corresponding region of the reference genome of *D. melanogaster*. Percentages of polymorphisms were assessed between lines, as well as genomic GC contents (SI, Section A.3). The polymorphism between lines was systematically lower than the threshold of 2%, confirming that the DNA belongs to a single species.

Transcriptome assembly

The RNA reads of each line were cleaned from adaptors with Trimmomatic (Bolger et al., 2014) and trimmed with FastQC (de Sena Brandine and Smith, 2019). The seven reference genomes were indexed with HISAT2 (Kim et al., 2019), and RNA reads from each line were mapped to their respective reference genome with HISAT2 using the spliced aware option. The resulting SAM files were converted to BAM format with Samtools. The BAM files were sorted and indexed. In each line, the transcriptomes were assembled with StringTie (Pertea et al., 2015) (Supplemental deposit). The corresponding GTF and GFF files were built with TransDecoder from Trinity (Grabherr et al., 2011). A GTF file based on the transcriptome assembly was created with GTF to FASTA from GffRead (Pertea and Pertea, 2020) (Supplemental deposit). For each line, a final file was generated including the genomic position of unspliced transcripts, transcription orientation, and the size of the unspliced transcripts. Transcript coverage was recorded as *Transcripts Per Million* (TPM). The GTF file of the reference *D. melanogaster* genome was retrieved to access the position of established genes and transposable elements genes, which were annotated in the seven new reference genomes.

Detection of de novo transcripts in the seven lines

The transcriptome of *D. melanogaster* BDGP6.28 was downloaded from NCBI, as well as the transcriptome of six other species (*Drosophila simulans, Drosophila sechellia, Drosophila virilis, Drosophila ananassae, Drosophila yakuba, Drosophila erecta*). To weed out as many transcripts of older origin as possible, these transcriptomes were merged and used as a target for nucleotide BLAST search with all line transcripts as query. Nucleotide BLASTs were performed with a cutoff of $E = 10^{-2}$, in the forward direction of the target transcripts. Transcripts with no hit were considered as preliminary *de novo* transcripts. Preliminary *de novo* transcripts were then used as query to BLASTn search against the transcriptome of seven other Dipteran species: *Aedes aegypti, Anopheles arabiensis, Culex quiquefasciatus, Lucilia cuprina, Musca domestica, Stomoxys calcitrans, Teleopsis dalmanni*, downloaded from Ensembl, with a cutoff of $E = 10^{-2}$, in the forward direction of the target transcripts. The remaining *de novo* transcripts without a BLAST hit were then filtered by TPM value. Transcripts with a TPM < 0.5 were removed from the data set to exclude very low frequency transcripts, which are potentially false positives. The remaining transcripts were considered as *de novo* transcripts (Supplemental deposit).

The genomic positions of unspliced transcripts were retrieved from the seven GTF files generated from the transcriptome assemblies, and the overlap with genomic components was assessed with bedtools

(Quinlan and Hall, 2010) and a code developed for this purpose. The transcripts were distributed in six genomic positions: a) "Overlapping with an intergenic region and a fragment of a gene in the identical direction", referred to as "exon longer", b) "Overlapping with an intergenic region only", referred to as "intergenic", c) "Overlapping with an intergenic region and a fragment of a gene but in the opposite direction", referred to as "antisense", d) "Overlapping with an intergenic region and a pseudogene", referred to as "pseudogenic", e) "Overlapping with an intergenic region and an annotated non-coding RNA", referred to as "ncRNA", and f) "Inside of an intron" referred as "intronic".

Orthogroups of de novo transcripts between lines

De novo transcripts from the seven lines were searched for orthology relationships between them. Constructing orthogroups of transcripts is more complex than constructing orthogroups of protein-coding genes. Two protein-coding genes are commonly grouped together into an orthogroup if the sequence similarity and coverage of their encoded protein exceeds a certain threshold, which suggests a common origin and an homologous function of the encoded protein. On the other hand, two transcripts can overlap but have arisen from a completely different transcription process if their transcription starting and ending sites do not correspond. Moreover, two transcripts can have similar starting and ending sites, but have different splicing in one line and between lines, giving rise to spliced transcripts with low sequence homology. Indeed, depending on the scientific question, transcript homology should be defined differently.

To cover a large number of possible scenario, we established three different definitions of transcript orthology that are depicted in Figure 4:

Definition 1: A set of transcripts are considered orthologous if their spliced sequences share at least 70% of coverage between all of the members of the orthogroup, and 75% of identity in reciprocal BLASTn.

Definition 2: A set of transcripts are considered orthologous if their spliced sequences share at least 70% of coverage between all of the members of the orthogroup and 75% of identity in reciprocal BLASTn, and the transcription starting sites of all of them can be found in a window of 500 bp of the genome.

Definition 3: A set of transcripts are considered orthologous if their unspliced sequences start at a genomic position that is within a window of 500 bp, and end in a window of 500 bp.

Definitions 1 and 2 compare spliced transcripts. Definition 2, in addition to sequence similarity, takes into account the genomic position of the transcription starting site, thus making orthogroup affiliation more restrictive than Definition 1. In contrast, Definition 3 builds on the starting and ending sites but do not consider sequence similarity. In this definition, two splice variants of one single transcript would be clustered together, independently of the sequence similarity of the spliced variants. Indeed, if two transcripts emerge at the same position in two lines but with different splicing variants, they would still be detected by this definition.

To build the orthogroups following Definitions 1 and 2, *de novo* transcripts from each line were used as target for BLASTn (Altschul et al., 1990) searches against the *de novo* transcripts of the six other lines. With a 70% coverage threshold, we did not encounter any ambiguous orthogroup definitions, e.g. when multiple transcripts overlapped with 70%, but not for all pairwise comparisons. A python script was used to sort the orthogroups according to our definitions. For Definition 3, a python script was built to directly assess the overlapping positions and nucleotides of *de novo* transcripts between lines. To test the accuracy of our definitions, we used a set of 11,000 proto-genes from Grandchamp et al. (2022a), and distributed them into orthogroups by using the software OrthoFinder (Emms and Kelly, 2019) and our script for Definition 1. We found 93% of similarity between the results from OrthoFinder and our script (Orthofinder: 5687 orthogroups, Our script: 6124 orthogroups) (Supplemental deposit). The difference of 7% is due to a higher threshold of similarity between all members of an orthogroup in our pipeline compared to OrthoFinder.

Additionally, we also used an alternative approach to estimate the similarity of transcripts between the lines. Instead of comparing transcripts with each other and treating them as a single object, we compare transcribed nucleotides. For example, if two transcripts of 200 bp found in two lines have an overlap of 50 bp, we will consider these 50 nucleotides as "*de novo* transcribed nucleotides" common to the two lines, and the remaining 150 nucleotides as "*de novo* transcribed nucleotides" specific to each line. We used a python script to classify the nucleotides of transcripts into these categories. We refer to this approach as 'Nucleotide similarity'.

All scripts are deposited at `https://github.com/AnnaGrBio/Transcripts-gain-and-loss`.

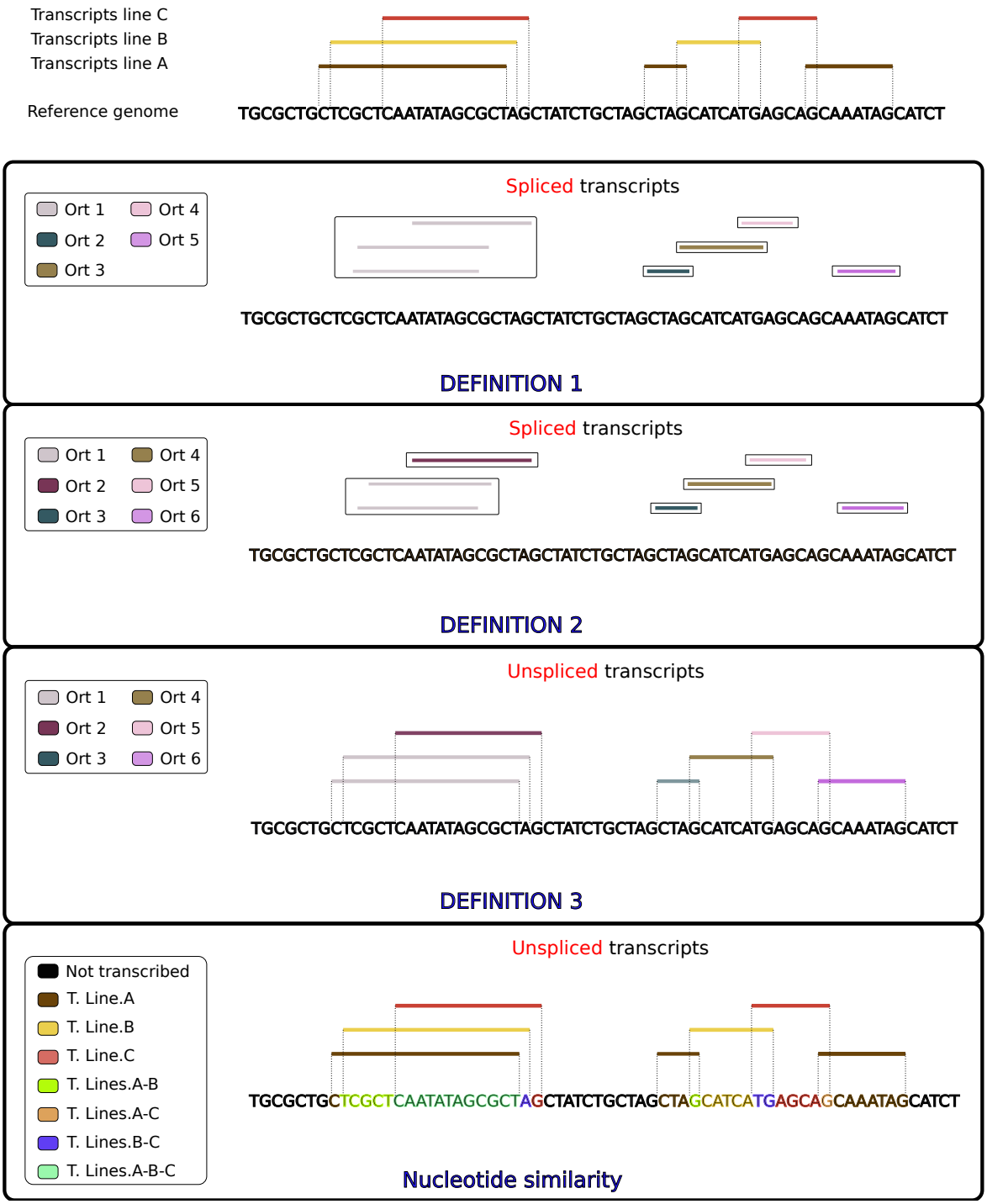

Figure 4: **Transcript classifications**. The figure illustrates the different *de novo* transcript definitions and the additional approach to categorize transcript nucleotides. The black nucleotides represent the reference genomes (all of the exact same size), the colored rows at the top show the transcripts. As an example, we show three lines.

Phylogeny of the seven *D. melanogaster* populations

One of the estimation methods of *de novo* transcript gain and loss rates (explained below) relies on a dated phylogenetic tree. To obtain an estimate of the phylogenetic history of the seven lines of *D. melanogaster*, we downloaded the genomes of each line assembled *de novo* in Grandchamp et al. (2022b). We did not use the reference genomes constructed by mapping in that study, as these genomes do not contain events of reshuffling, indels or transposable element insertions that are more precise to reconstruct the phylogenetic tree. Furthermore, we note that because of constant gene flow between populations, a strict phylogenetic classification in form of a binary tree is neither possible nor desirable but serves as a first approximation to group populations such that gain and loss rates can be estimated.

For each line, all coding sequences (CDS) were retrieved with GffRead (Pertea and Pertea, 2020). For all genes, longest CDS common to each line (11,568) were stored in orthogroups. Sequences from the 11,568 orthogroups were aligned with MAFFT (Katoh and Standley, 2013) and the 11,568 alignments were concatenated in one single alignment with python. This concatenated alignment was converted to nexus format with Seqmagick (fhcrc.github.io/seqmagick) and used as input for the software BEAST (Bouckaert et al., 2014; Drummond et al., 2012) for constructing a dated tree (details in SI, Section A.5).

To set the priors for constructing this tree, we used the software BEAUti from the BEAST package. To set the parameters, we followed the methodology as suggested by Drummond et al. (2012). The Gamma category count was set to 4. Blosum62 was used as the substitution model for the amino acid sequences and HKY with $k = 2$ for the nucleotide sequences as empirical frequencies. The Yule model was selected as the model of speciation (Harding, 1971; Nee et al., 1994). The divergence time between Zambian and European populations was set to 12,843 years, as inferred by Laurent et al. (2011). The prior for the molecular clock and speciation rates were set as gamma distributions with $\alpha = 0,001$ and $\beta = 1,000$. Further, an informative prior was set with the ZI populations as an outgroup. The mean was set to $\hat{u} = 14,000$ with a standard deviation $\sigma = 400$. The Markov Chain Monte Carlo (MCMC) options were chosen as follows: chain length 6 million, trace $\log(\text{frequency}) = 1,000$.

Once the prior set, the phylogenetic tree was constructed with BEAST. The best tree was selected with Tree annotator from BEAST with the following options: burning percentage $= 10$, target tree type "Maximum clade credibility tree", node heights: "mean heights". The Bayesian parameters were estimated with Tracer 1.7 (Rambaut et al., 2018). The tree visualisation and adjustment was done with Figtree (http://tree.bio.ed.ac.uk/software/figtree/). The accuracy of the tree was tested by

generating another phylogenetic tree based on the same alignment with RAxML (Stamatakis, 2016). The two estimated trees had the same topology, meaning that the inferred phylogeny of the seven populations given the data is robust (SI, Section A.5).

Evolutionary model of gain and loss of *de novo* transcripts

We estimate gain and loss rates of *de novo* transcripts using a model that describes transcript gain and loss dynamics along the constructed phylogenetic tree of the *D. melanogaster* populations from the previous section. This evolutionary model is based on the *infinitely many genes model*, which has been developed to describe the gain and loss dynamics of genes in prokaryotes (Baumdicker et al., 2010, 2012). In analogy to the infinitely many genes model, we assume that a specific *de novo* transcript arises only once during the short evolutionary time frame that we study. This means that any *de novo* transcript that is shared between different populations must have been gained in one of the ancestral populations common to both extant populations. Furthermore, we assume that *de novo* transcripts are selectively neutral, i.e., they do not confer a fitness advantage or disadvantage (Palazzo and Koonin, 2020).

De novo transcripts are gained with probability u per unit of time and each transcript is lost with probability v per unit of time. In the following we will either use a generation or years as a unit of time. To translate between generations and years, we assume that the generation time of *D. melanogaster* is approximately two weeks (Fernández-Moreno et al., 2007). The total number of *de novo* transcripts gained after t units of time, denoted by $g(t)$, is then described by the following dynamics (e.g. Eqs. (6), (7) in Collins and Higgs, 2012):

$$\frac{dg}{dt} = u - vg(t) \qquad \Rightarrow \qquad g(t) = \frac{u}{v}\left(1 - e^{-vt}\right) . \tag{1}$$

The equilibrium number of *de novo* transcripts therefore is u/v.

Estimating *de novo* transcript gain and loss rates

Next, our goal is to estimate the gain and loss rates of *de novo* transcripts. We will use two different strategies, based on (i) coalescent theory and (ii) on the estimated phylogeny of the sampled populations. In both cases we compare the empirical transcript frequency spectrum to theoretical predictions. The transcript frequency spectrum contains information about the number of transcripts shared by a certain number of populations. We denote by $D^n = (d_1^n, ..., d_n^n)$ the empirical transcript

frequency spectrum and by $T^n = (t_1^n, ..., t_n^n)$ the theoretical prediction for the transcript frequency spectrum, where n is the number of populations that are studied (here $n = 7$). To estimate the parameters, we follow the approach from Collins and Higgs (2012) and use a weighted least squares method that gives the following summary statistic:

$$\chi^2 = \sum_{k=1}^{n} \frac{(d_k^n - t_k^n)^2}{t_k^n} \,. \tag{2}$$

We now outline how to compute the theoretical transcript frequency spectrum, either using a coalescent-based approach or using the estimated phylogeny of the sampled populations, which we refer to as the fixed-tree method.

(i) Coalescent method (Baumdicker et al., 2010, 2012): The coalescent describes the phylogeny of a neutrally evolving population of individuals under the assumption of no population structure (Kingman, 1982); for a comprehensive introduction to coalescent theory we refer to Wakeley (2008). Under these assumption, the genealogy of the sample is given by a standard coalescent and the theoretical frequency spectrum can be computed analytically (Baumdicker et al., 2010). The assumption on population structure is likely violated in our data. While the European populations seem to have only a relatively weak population structure (Kapun et al., 2020), the Zambian population is clearly isolated from the European populations. Even if there is gene flow between the European and Zambian populations, as recently inferred (Kapopoulou et al., 2020), the rate is still lower as the rate of gene flow between the European populations, thus violating the assumption of no population structure. We therefore conducted our analysis also on the set of European populations only, which can be assumed to be approximately unstructured (Kapun et al., 2020). The obtained estimates are largely robust with respect to the assumption of population structure (Table 3).

The frequency spectrum is given in transformed gain and loss rates to transform them to the time scale of the coalescent, i.e., we write $\theta = 2N_e u$ and $\rho = 2N_e v$. N_e denotes the effective population size, which here becomes the effective population size of *D. melanogaster* of the seven lines. Then θ is the average number of gained transcripts in $2N_e$ generations and ρ is the rate at which a transcript is lost in $2N_e$ generations. We denote by G_k^n the number of transcripts shared by k populations out of the n sampled populations. Similar to Collins and Higgs (2012), we also study two transcript classes: transcripts with high and low turnover. To distinguish their respective rates, we write θ_s, ρ_s for the "slow" class and θ_f, ρ_f for the "fast" class of transcripts. We drop the indices if we use only one class of transcripts in the following. The expected frequency spectrum, denoted by $\mathbb{E}[\cdot]$, in its most general

form is then given by (Baumdicker et al., 2010, Theorem 5 extended to multiple classes)

$$\begin{aligned}
\mathbb{E}\left[G_k^n\right] &= \sum_{i\in\{s,f\}} \frac{\theta_i}{k} \frac{n\cdots(n-k+1)}{(n-1+\rho_i)\cdots(n-k+\rho_i)}\,, \text{ for } 1\leq k<n\,,\\
\mathbb{E}\left[G_n^n\right] &= \sum_{i\in\{s,f\}} \frac{\theta_i}{n} \frac{n\cdots 1}{(n-1+\rho_i)\cdots\rho_i} + C_{\text{fixed}}\,,
\end{aligned} \tag{3}$$

where C_{fixed} denotes the number of *de novo* transcripts that are fixed in all *D. melanogaster* populations, which means that they have a loss rate equal to zero.

(ii) Fixed-tree method (Baumdicker et al., 2012; Collins and Higgs, 2012): In contrast to a random phylogeny described by a standard coalescent (previous paragraphs), here we use the estimated dated phylogenetic tree of the seven lines (previous section). We therefore calculate the expected transcript frequency spectrum by computing the number of transcripts gained and lost along each branch of this fixed phylogenetic tree. These gains and losses are then combined across the entire phylogeny to obtain the frequency spectrum. We follow the approach outlined in Collins and Higgs (2012). In brief, the theoretical frequency spectrum for a fixed tree with n genomes (the tip nodes), and thus $n-1$ internal nodes is given by (Collins and Higgs, 2012, Eq. (14))

$$\mathbb{E}\left[G_k^n\right] = \sum_{i=1}^{2n-1} p_i(k)g(t_i), \quad \text{for } 1\leq k\leq n-1\,, \tag{4}$$

where t_i is the branch length leading to node i, $g(t_i)$ is the expected number of transcripts at node i that were gained on the branch leading to it, and $p_i(k)$ is the probability that a transcript at node i will be retained in k tip nodes that descend from i. The function $g(t)$ is given by Eq. (1). Note that the branch leading to the root has length infinity, so that the number of transcripts in the root is given by u/v. For the values of p_i, one first notes that if i is a tip node, then $p_i(1)=1$ and $p_i(k)=0$ for $k\neq 1$. If i is an internal node, one can compute the probabilities for retaining a transcript in k descendants by a recursion that involves survival and loss probabilities of a transcript along the branches. The survival probability is given by $s(t)=e^{-vt}$ and the loss probability is $l(t)=1-e^{-vt}$. Denoting by t_{i_1} and t_{i_2} the two descending branch lengths from node i that lead to nodes i_1 and i_2, the recursion is defined by (Collins and Higgs, 2012, Eqs. (16), (17))

$$\begin{aligned}
p_i(k) &= s(t_{i_1})l(t_{i_2})p_{i_1}(k) + l(t_{i_1})s(t_{i_2})p_{i_2}(k) + s(t_{i_1})s(t_{i_2})\sum_{l=0}^{k} p_{i_1}(l)p_{i_2}(k-l), \text{ for } 1\leq k\leq n\,;\\
p_i(0) &= \left(l(t_{i_1}) + s(t_{i_1})p_{i_1}(0)\right)\left(l(t_{i_2}) + s(t_{i_2})p_{i_2}(0)\right)\,.
\end{aligned} \tag{5}$$

As for the coalescent method, this procedure can be extended to multiple classes of transcripts with

low and high turnover. For two classes we apply this procedure separately with the set of rates for each class and then sum up the expected frequency spectra. Moreover, if we include a class of fixed transcripts, we add the constant value C_{fixed} to the value $\mathbb{E}[G_n^n]$.

Numerical implementation of the parameter estimation

We implemented both methods to compute the theoretical frequency spectrum in python and used the integrated 'minimize' function from SciPy using the method 'SLSQP' to obtain parameter estimates through Eq. (2). We estimated parameters using one or two classes of transcripts plus the number of fixed transcripts, i.e. we estimated either three or five parameters. We constrained the parameters so that the mean number of transcripts per population, the value u/v or θ/ρ, fits the empirical observation $\sum_{i=1}^{7} id_i^n/n$. We assumed this to improve convergence of the minimization procedure by reducing the number of parameters to be estimated, which was necessary in the multi-class case. In addition, we conditioned the loss parameter(s) to be between 0 and 1,000 and the gain parameter(s) to be between 0 and 20,000.

Initial values for the parameters in the optimization routine were chosen by fitting the mean number of transcripts per population and the pairwise-population differences if we fitted one class of transcripts, as reasoned in Baumdicker et al. (2012) (more details in the SI, Section B.1). When fitting two classes of transcripts we divided the initial guess of the gain rate by two and set the initial gain and loss rate estimates of the slow transcript class to the initial values of the fast class divided by 100. For some data sets it was necessary to adjust the initial parameter choices to ensure convergence of the minimization algorithm. All choices of initial parameter values are stated in Section B.1 in the SI.

Programming and analyses

All statistical analyses were performed with R (R Core Team, 2022). Data reshuffling, analyses, orthology searches and modelling were performed with python (Van Rossum and Drake, 2009), and can be accessed at: `https://github.com/AnnaGrBio/Transcripts-gain-and-loss`.

Data access

The genomic DNA and RNA sequences are available under NCBI Bioproject PRJNA929424. Additionally, a package containing processed data is available in the zenodo archive: (https://doi.org/10.5281/zenodo.7681079), and is referred in the main text as "Supplemental deposit". The archive contains, for each line, the polymorphic genomes, transcriptome assemblies, *de novo* transcripts, orthology comparison results, orthogroups according to the 4 classifications, proteins alignments for the phylogeny. Supplemental figures, informations, analyses and models are found in the Supplementary Information (SI). All programs are stored in the github link (https://github.com/AnnaGrBio/Transcripts-gain-and-loss).

Competing interest statement

The authors declare no competing interests.

Authors contribution

AG, PC and EBB conceptualised the study; AG assembled the genomes and transcriptomes, detected *de novo* transcripts, built the orthogroups and the phylogeny; PC conducted the mathematical modelling and parameter estimation; AG, PC and EBB validated and discussed the results; AG and PC wrote the original draft; AG, PC, EBB edited and commented on the draft; AG, PC and EBB acquired funding.

Funding

AG and EBB acknowledge funding by Alexander von Humboldt-Stiftung. This work was supported in part by the Deutsche Forschungsgemeinschaft priority program "Genomic Basis of Evolutionary Innovations" (SPP 2349), project BO 2544/20-1 awarded to EBB and project CZ 294/1-1 awarded to PC.

References

Afgan, E., Baker, D., Van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*, 44(W1):W3–W10, 2016.

Albà, M. and Castresana, J. On homology searches by protein Blast and the characterization of the age of genes. *BMC Evolutionary Biology*, 7(1):1–8, 2007.

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

Ambrosino, L. and Chiusano, M. L. Transcriptologs: a transcriptome-based approach to predict Orthology relationships. *Bioinformatics and Biology Insights*, 11:1177932217690136, 2017.

Aspden, J. L., Eyre-Walker, Y. C., Phillips, R. J., Amin, U., Mumtaz, M. A. S., Brocard, M., and Couso, J.-P. Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. *Elife*, 3:e03528, 2014.

Ballester, B., Medina-Rivera, A., Schmidt, D., Gonzàlez-Porta, M., Carlucci, M., Chen, X., Chessman, K., Faure, A. J., Funnell, A. P., Goncalves, A., et al. Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *Elife*, 3:e02626, 2014.

Barman, P., Reddy, D., and Bhaumik, S. R. Mechanisms of antisense transcription initiation with implications in gene expression, genomic integrity and disease pathogenesis. *Non-coding RNA*, 5(1):11, 2019.

Baumdicker, F., Hess, W. R., and Pfaffelhuber, P. The diversity of a distributed genome in bacterial populations. *The Annals of Applied Probability*, 20(5):1567 – 1606, 2010. doi: 10.1214/09-AAP657.

Baumdicker, F. and Pfaffelhuber, P. The infinitely many genes model with horizontal gene transfer. *Electronic Journal of Probability*, 19:1 – 27, 2014. doi: 10.1214/EJP.v19-2642.

Baumdicker, F., Hess, W. R., and Pfaffelhuber, P. The Infinitely Many Genes Model for the Distributed Genome of Bacteria. *Genome Biology and Evolution*, 4(4):443–456, 2012. doi: 10.1093/gbe/evs016.

Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., Vejnar, C. E., Lee, M. T., Rajewsky, N., Walther, T. C., et al. Identification of small ORF s in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO journal*, 33(9):981–993, 2014.

Bolger, A. M., Lohse, M., and Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 10(4):e1003537, 2014.

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., et al. Ten things you should know about transposable elements. *Genome Biology*, 19(1):1–12, 2018.

Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O., Booth, B. W., Wen, J., Park, S., Suzuki, A. M., et al. Diversity and dynamics of the Drosophila transcriptome. *Nature*, 512(7515):393–399, 2014.

Brown, R. P. and Feder, M. E. Reverse transcriptional profiling: non-correspondence of transcript level variation and proximal promoter polymorphism. *BMC Genomics*, 6(1):1–10, 2005.

Candelli, T., Challal, D., Briand, J.-B., Boulay, J., Porrua, O., Colin, J., and Libri, D. High-resolution transcription maps reveal the widespread impact of roadblock termination in yeast. *EMBO journal*, 37(4): e97490, 2018a.

Candelli, T., Gros, J., and Libri, D. Pervasive transcription fine-tunes replication origin activity. *Elife*, 7:e40802, 2018b.

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., et al. Proto-genes and *de novo* gene birth. *Nature*, 487(7407): 370–374, 2012.

Clark, M. B., Amaral, P. P., Schlesinger, F. J., Dinger, M. E., Taft, R. J., Rinn, J. L., Ponting, C. P., Stadler, P. F., Morris, K. V., Morillon, A., et al. The reality of pervasive transcription. *PLoS Biology*, 9(7):e1000625, 2011.

Collins, R. E. and Higgs, P. G. Testing the Infinitely Many Genes Model for the Evolution of the Bacterial Core Genome and Pangenome. *Molecular Biology and Evolution*, 29(11):3413–3425, 2012. doi: 10.1093/molbev/mss163.

Cotney, J., Leng, J., Yin, J., Reilly, S. K., DeMare, L. E., Emera, D., Ayoub, A. E., Rakic, P., and Noonan, J. P. The evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell*, 154(1):185–196, 2013.

Couso, J.-P. and Patraquim, P. Classification and function of small open reading frames. *Nature Reviews Molecular Cell Biology*, 18(9):575–589, 2017.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., et al. Twelve years of SAMtools and BCFtools. *Gigascience*, 10(2):giab008, 2021.

de Sena Brandine, G. and Smith, A. D. Falco: high-speed FastQC emulation for quality control of sequencing data. *F1000Research*, 8, 2019.

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8):1969–1973, 2012.

Emms, D. M. and Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1):1–14, 2019.

Fernández-Moreno, M. A., Farr, C. L., Kaguni, L. S., and Garesse, R. *Drosophila melanogaster* as a Model System to Study Mitochondrial Biology. In *Methods in Molecular Biology*, pages 33–49. Humana Press, 2007. doi: 10.1007/978-1-59745-365-3_3.

Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, 9(5): 397–405, 2008.

Frith, M. C., Ponjavic, J., Fredman, D., Kai, C., Kawai, J., Carninci, P., Hayshizaki, Y., and Sandelin, A. Evolutionary turnover of mammalian transcription start sites. *Genome Research*, 16(6):713–722, 2006.

Gabaldón, T. Large-scale assignment of orthology: back to phylogenetics? *Genome Biology*, 9(10):1–6, 2008.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, 29(7):644, 2011.

Grandchamp, A., Berk, K., Dohmen, E., and Bornberg-Bauer, E. New genomic signals underlying the emergence of human proto-genes. *Genes*, 13(2):284, 2022a.

Grandchamp, A., Kühl, L., Lebherz, M. K., Brüggemann, K., Parsch, J., and Bornberg-Bauer, E. Population genomics reveals mechanisms and dynamics of *de novo* proto-gene emergence in *Drosophila melanogaster*. *bioRxiv*, pages 2022–11, 2022b.

Grosso, A. R., Leite, A. P., Carvalho, S., Matos, M. R., Martins, F. B., Vítor, A. C., Desterro, J. M., Carmo-Fonseca, M., and de Almeida, S. F. Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. *Elife*, 4:e09214, 2015.

Hangauer, M. J., Vaughn, I. W., and McManus, M. T. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genetics*, 9(6):e1003569, 2013.

Harding, E. The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability*, 3(1):44–77, 1971.

He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N., and Kinzler, K. W. The antisense transcriptomes of human cells. *Science*, 322(5909):1855–1857, 2008.

Heames, B., Schmitz, J., and Bornberg-Bauer, E. A continuum of evolving *de novo* genes drives protein-coding novelty in Drosophila. *Journal of Molecular Evolution*, 88(4):382–398, 2020.

Hiller, M., Findeiß, S., Lein, S., Marz, M., Nickel, C., Rose, D., Schulz, C., Backofen, R., Prohaska, S. J., Reuter, G., et al. Conserved introns reveal novel transcripts in *Drosophila melanogaster*. *Genome Research*, 19(7):1289–1300, 2009.

Houtgast, E. J., Sima, V.-M., Bertels, K., and Al-Ars, Z. Hardware acceleration of BWA-MEM genomic short read mapping for longer read lengths. *Computational Biology and Chemistry*, 75:54–64, 2018.

Huang, W., Carbone, M. A., Magwire, M. M., Peiffer, J. A., Lyman, R. F., Stone, E. A., Anholt, R. R., and Mackay, T. F. Genetic basis of transcriptome diversity in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences*, 112(44):E6010–E6019, 2015.

Jacquier, A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nature Reviews Genetics*, 10(12):833–844, 2009.

Kapopoulou, A., Kapun, M., Pieper, B., Pavlidis, P., Wilches, R., Duchen, P., Stephan, W., and Laurent, S. Demographic analyses of a new sample of haploid genomes from a Swedish population of *Drosophila melanogaster*. *Scientific Reports*, 10(1), 2020. doi: 10.1038/s41598-020-79720-1.

Kapun, M., Barrón, M. G., Staubach, F., Obbard, D. J., Wiberg, R. A. W., Vieira, J., Goubert, C., Rota-Stabelli, O., Kankare, M., Bogaerts-Márquez, M., Haudry, A., Waidele, L., Kozeretska, I., Pasyukova, E. G., Loeschcke, V., Pascual, M., Vieira, C. P., Serga, S., Montchamp-Moreau, C., Abbott, J., Gibert, P., Porcelli, D., Posnien, N., Sánchez-Gracia, A., Grath, S., Sucena, E., Bergland, A. O., Guerreiro, M. P. G., Onder, B. S., Argyridou, E., Guio, L., Schou, M. F., Deplancke, B., Vieira, C., Ritchie, M. G., Zwaan, B. J., Tauber, E., Orengo, D. J., Puerma, E., Aguadé, M., Schmidt, P., Parsch, J., Betancourt, A. J., Flatt, T., and González, J. Genomic Analysis of European *Drosophila melanogaster* Populations Reveals Longitudinal Structure, Continent-Wide Selection, and Previously Unknown DNA Viruses. *Molecular Biology and Evolution*, 37(9):2661–2678, 2020. doi: 10.1093/molbev/msaa120.

Kapusta, A. and Feschotte, C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends in Genetics*, 30(10):439–452, 2014.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C., Suzuki, M., Kawai, J., et al. Antisense transcription in the mammalian transcriptome. *Science*, 309(5740): 1564–1566, 2005.

Katoh, K. and Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., Ward, L. D., Birney, E., Crawford, G. E., Dekker, J., et al. Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences*, 111(17):6131–6138, 2014.

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8):907–915, 2019.

Kim, N. and Jinks-Robertson, S. Transcription as a source of genome instability. *Nature Reviews Genetics*, 13 (3):204–214, 2012.

Kingman, J. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982. doi: https://doi.org/10.1016/0304-4149(82)90011-4.

Krchňáková, Z., Thakur, P. K., Krausová, M., Bieberstein, N., Haberman, N., Müller-McNicoll, M., and Staněk, D. Splicing of long non-coding RNAs primarily depends on polypyrimidine tract and 5' splice-site sequences due to weak interactions with SR proteins. *Nucleic Acids Research*, 47(2):911–928, 2019.

Kristensen, D. M., Wolf, Y. I., Mushegian, A. R., and Koonin, E. V. Computational methods for Gene Orthology inference. *Briefings in Bioinformatics*, 12(5):379–391, 2011.

Kutter, C., Watt, S., Stefflova, K., Wilson, M. D., Goncalves, A., Ponting, C. P., Odom, D. T., and Marques, A. C. Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression. *PLoS Genetics*, 8(7): e1002841, 2012. doi: 10.1371/journal.pgen.1002841.

Kuzniar, A., van Ham, R. C., Pongor, S., and Leunissen, J. A. The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics*, 24(11):539–551, 2008.

Lasa, I., Toledo-Arana, A., Dobin, A., Villanueva, M., de los Mozos, I. R., Vergara-Irigaray, M., Segura, V., Fagegaltier, D., Penadés, J. R., Valle, J., et al. Genome-wide antisense transcription drives mRNA processing in bacteria. *Proceedings of the National Academy of Sciences*, 108(50):20172–20177, 2011.

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

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

Liu, H. and Zhang, J. Is the mutation rate lower in genomic regions of stronger selective constraints? *Molecular Biology and Evolution*, 39(8):msac169, 2022.

Mercer, T. R., Dinger, M. E., and Mattick, J. S. Long non-coding RNAs: insights into functions. *Nature Reviews Genetics*, 10(3):155–159, 2009.

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., and Durbin, R. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, 32(11):1749–1751, 2016.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J. C., Grützner, F., and Kaessmann, H. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, 505 (7485):635–640, 2014.

Nee, S., May, R. M., and Harvey, P. H. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1309):305–311, 1994.

Neme, R. and Tautz, D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to *de novo* gene emergence. *Elife*, 5:e09977, 2016.

Nicolas, P., Mäder, U., Dervyn, E., Rochat, T., Leduc, A., Pigeonneau, N., Bidnenko, E., Marchadier, E., Hoebeke, M., Aymerich, S., et al. Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. *Science*, 335(6072):1103–1106, 2012.

Odom, D. T., Dowell, R. D., Jacobsen, E. S., Gordon, W., Danford, T. W., MacIsaac, K. D., Rolfe, P. A., Conboy, C. M., Gifford, D. K., and Fraenkel, E. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genetics*, 39(6):730–732, 2007.

Palazzo, A. F. and Koonin, E. V. Functional Long Non-coding RNAs Evolve from Junk Transcripts. *Cell*, 183 (5):1151–1161, 2020. doi: 10.1016/j.cell.2020.09.047.

Palazzo, A. F. and Lee, E. S. Non-coding RNA: what is functional and what is junk? *Frontiers in Genetics*, 6: 2, 2015.

Papantonis, A. and Cook, P. R. Genome architecture and the role of transcription. *Current Opinion in Cell Biology*, 22(3):271–276, 2010.

Patraquim, P., Magny, E. G., Pueyo, J. I., Platero, A. I., and Couso, J. P. Translation and natural selection of micropeptides from long non-canonical RNAs. *Nature Communications*, 13(1):1–15, 2022.

Pelechano, V. and Steinmetz, L. M. Gene regulation by antisense transcription. *Nature Reviews Genetics*, 14 (12):880–893, 2013.

Pertea, G. and Pertea, M. GFF utilities: GffRead and GffCompare. *F1000Research*, 9, 2020.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3): 290–295, 2015.

Piovesan, A., Antonaros, F., Vitale, L., Strippoli, P., Pelleri, M. C., and Caracausi, M. Human protein-coding genes and gene feature statistics in 2019. *BMC Research Notes*, 12(1):1–5, 2019.

Ponting, C. P., Oliver, P. L., and Reik, W. Evolution and functions of long noncoding RNAs. *Cell*, 136(4): 629–641, 2009.

Porrua, O. and Libri, D. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nature Reviews Molecular Cell Biology*, 16(3):190–202, 2015.

Quinlan, A. R. and Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

Quinn, J. J. and Chang, H. Y. Unique features of long non-coding RNA biogenesis and function. *Nature Reviews Genetics*, 17(1):47–62, 2016.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, 67(5):901–904, 2018.

Ramirez-Gonzalez, R. H., Bonnal, R., Caccamo, M., and MacLean, D. Bio-samtools: Ruby bindings for SAMtools, a library for accessing BAM files containing high-throughput sequence alignments. *Source code for biology and medicine*, 7(1):1–6, 2012.

Reinhardt, J. A., Wanjiru, B. M., Brant, A. T., Saelao, P., Begun, D. J., and Jones, C. D. *De Novo* ORFs in Drosophila Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-coding Sequences. *PLoS Genetics*, 9(10):e1003860, 2013. doi: 10.1371/journal.pgen.1003860.

Romero-Barrios, N., Legascue, M. F., Benhamed, M., Ariel, F., and Crespi, M. Splicing regulation by long noncoding RNAs. *Nucleic Acids Research*, 46(5):2169–2184, 2018.

Ruiz-Orera, J., Messeguer, X., Subirana, J. A., and Alba, M. M. Long non-coding RNAs as a source of new peptides. *Elife*, 3:e03523, 2014.

Rutkowski, A. J., Erhard, F., L'Hernault, A., Bonfert, T., Schilhabel, M., Crump, C., Rosenstiel, P., Efstathiou, S., Zimmer, R., Friedel, C. C., et al. Widespread disruption of host transcription termination in HSV-1 infection. *Nature Communications*, 6(1):1–15, 2015.

Sana, J., Faltejskova, P., Svoboda, M., and Slaby, O. Novel classes of non-coding RNAs and cancer. *Journal of Translational Medicine*, 10(1):1–21, 2012.

Schlötterer, C. Genes from scratch – the evolutionary fate of *de novo* genes. *Trends in Genetics*, 31(4):215–219, 2015. doi: 10.1016/j.tig.2015.02.007.

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

Schmitz, J. F., Ullrich, K. K., and Bornberg-Bauer, E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nature Ecology & Evolution*, 2(10):1626–1632, 2018.

Singh, S. S., Singh, N., Bonocora, R. P., Fitzgerald, D. M., Wade, J. T., and Grainger, D. C. Widespread suppression of intragenic transcription initiation by H-NS. *Genes & Development*, 28(3):214–219, 2014.

Stamatakis, A. The RAxML v8. 2. X Manual. *Heidleberg Institute for Theoretical Studies*, 2016.

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, 2013.

Van Oss, S. B. and Carvunis, A.-R. *De novo* gene birth. *PLOS Genetics*, 15(5):e1008160, 2019. doi: 10.1371/journal.pgen.1008160.

Van Rossum, G. and Drake, F. L. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

van Steensel, B. and Furlong, E. E. The role of transcription in shaping the spatial organization of the genome. *Nature Reviews Molecular Cell Biology*, 20(6):327–337, 2019.

Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R. S., Stehling-Sun, S., Sabo, P. J., Byron, R., Humbert, R., et al. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science*, 346(6212):1007–1012, 2014.

Vilborg, A., Passarelli, M. C., Yario, T. A., Tycowski, K. T., and Steitz, J. A. Widespread inducible transcription downstream of human genes. *Molecular Cell*, 59(3):449–461, 2015.

Villar, D., Flicek, P., and Odom, D. T. Evolution of transcription factor binding in metazoans—mechanisms and functional implications. *Nature Reviews Genetics*, 15(4):221–233, 2014.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J., et al. Enhancer evolution across 20 mammalian species. *Cell*, 160(3):554–566, 2015.

Wade, J. T. and Grainger, D. C. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nature Reviews Microbiology*, 12(9):647–653, 2014.

Wakeley, J. *Coalescent Theory: An Introduction*. Roberts & Company, 2008.

Wang, K. C. and Chang, H. Y. Molecular mechanisms of long noncoding RNAs. *Molecular Cell*, 43(6):904–914, 2011.

Wang, K. C., Yang, Y. W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B. R., Protacio, A., Flynn, R. A., Gupta, R. A., et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature*, 472(7341):120–124, 2011.

Ward, L. D. and Kellis, M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science*, 337(6102):1675–1678, 2012.

Wittkopp, P. J. and Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, 13(1):59–69, 2012.

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., et al. Ensembl 2020. *Nucleic Acids Research*, 48(D1):D682–D688, 2020.

Young, R. S., Hayashizaki, Y., Andersson, R., Sandelin, A., Kawaji, H., Itoh, M., Lassmann, T., Carninci, P., Bickmore, W. A., Forrest, A. R., et al. The frequent evolutionary birth and death of functional promoters in mouse and human. *Genome Research*, 25(10):1546–1557, 2015.

Zamani-Dahaj, S. A., Okasha, M., Kosakowski, J., and Higgs, P. G. Estimating the Frequency of Horizontal Gene Transfer Using Phylogenetic Models of Gene Gain and Loss. *Molecular Biology and Evolution*, 33(7): 1843–1857, 2016. doi: 10.1093/molbev/msw062.

Zhang, L., Ren, Y., Yang, T., Li, G., Chen, J., Gschwend, A. R., Yu, Y., Hou, G., Zi, J., Zhou, R., et al. Rapid evolution of protein diversity by *de novo* origination in *Oryza*. *Nature Ecology & Evolution*, 3(4):679–690, 2019.