DNA Transposons Favor De Novo Transcript Emergence Through Enrichment of Transcription Factor Binding Motifs

Marie Kristin Lebherz [1], Bertrand Fouks[2,3,4], Julian Schmidt[1], Erich Bornberg-Bauer [1,5,*], Anna Grandchamp [1,6,*]

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

[2]CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France

[3]UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398, Montpellier, France

[4]CIRAD, UMR AGAP Institut, F-34398, Montpellier, France

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

[6]Present address: Aix-Marseille Univ, INSERM, TAGC, Marseille, France

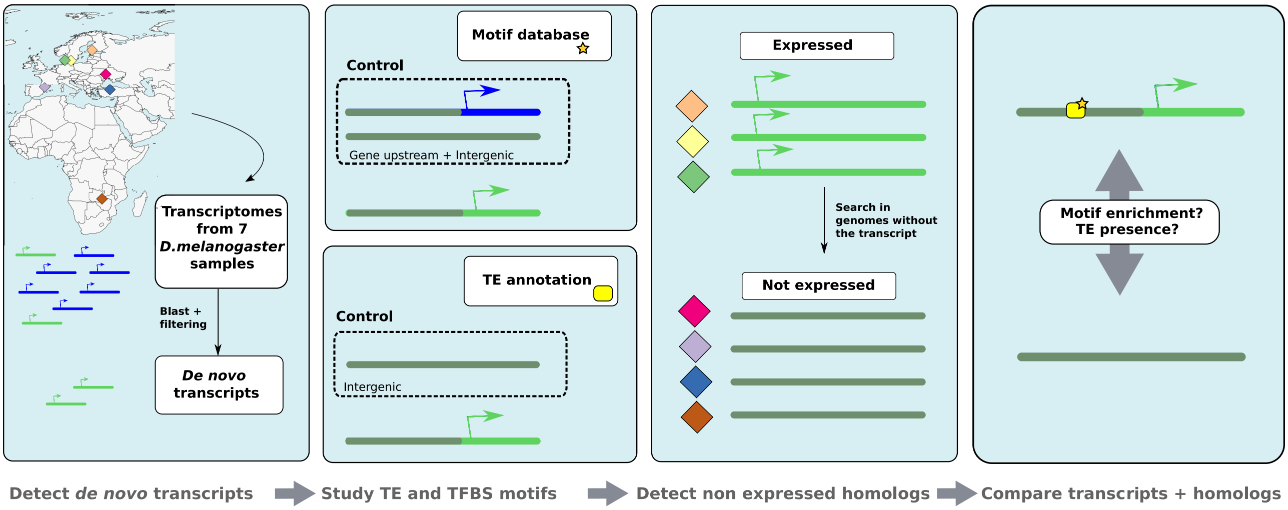
*Corresponding authors: E-mails: ebb-admin@uni-muenster.de; a.grandchamp@uni-muenster.de.

Accepted: June 15, 2024

Abstract

De novo genes emerge from noncoding regions of genomes via succession of mutations. Among others, such mutations activate transcription and create a new open reading frame (ORF). Although the mechanisms underlying ORF emergence are well documented, relatively little is known about the mechanisms enabling new transcription events. Yet, in many species a continuum between absent and very prominent transcription has been reported for essentially all regions of the genome. In this study, we searched for de novo transcripts by using newly assembled genomes and transcriptomes of seven inbred lines of *Drosophila melanogaster*, originating from six European and one African population. This setup allowed us to detect sample specific de novo transcripts, and compare them to their homologous nontranscribed regions in other samples, as well as genic and intergenic control sequences. We studied the association with transposable elements (TEs) and the enrichment of transcription factor motifs upstream of de novo emerged transcripts and compared them with regulatory elements. We found that de novo transcripts overlap with TEs more often than expected by chance. The emergence of new transcripts correlates with regions of high guanine-cytosine content and TE expression. Moreover, upstream regions of de novo transcripts are highly enriched with regulatory motifs. Such motifs are more enriched in new transcripts overlapping with TEs, particularly DNA TEs, and are more conserved upstream de novo transcripts than upstream their 'nontranscribed homologs'. Overall, our study demonstrates that TE insertion is important for transcript emergence, partly by introducing new regulatory motifs from DNA TE families.

Graphical Abstract


Transcriptomes from 7 D.melanogaster samples
Blast + filtering
De novo transcripts
Motif database
Control
Gene upstream + Intergenic
TE annotation
Control
Intergenic
Expressed
Search in genomes without the transcript
Not expressed
Motif enrichment? TE presence?
Detect de novo transcripts
Study TE and TFBS motifs
Detect non expressed homologs
Compare transcripts + homologs

Key words: *Drosophila melanogaster*, transposable elements, transcription factor motifs, de novo transcripts.

Significance

In the present study, we used inbred lines of *Drosophila melanogaster* to detect earlier stages of de novo emerged transcripts in samples. We determined and studied the impact of transposable elements (TEs) and TFBS motifs on the emergence of de novo transcripts. We show that the insertion of DNA transposons plays a role in de novo transcripts emergence. We demonstrate enrichment of transcription factor binding motif (motifs whose identity to a reference motif is low) upstream de novo transcripts compared to regions upstream annotated genes and control non transcribed intergenic sequences.This enrichment is even more frequent upstream de novo transcripts overlapping with DNA TEs. Our findings help elucidate main molecular drivers of transcription gain, namely insertions of DNA TEs and enrichment in transcription factor motifs with lower similarity to the reference.

Introduction

Recent studies showed that a nonnegligible proportion of new genes can emerge de novo from noncoding regions of the genome (Tautz and Domazet-Lošo 2011; Bornberg-Bauer et al. 2015, 2021; Schlötterer 2015; McLysaght and Hurst 2016; Rödelsperger et al. 2019; Van Oss and Carvunis 2019). Several de novo genes have been shown to become essential, bearing important organismal functions, e.g. male fertility (Gubala et al. 2017) and cold resistance (Baalsrud et al. 2018). The process of de novo gene emergence entails two main steps (Carvunis et al. 2012) in which a nongenic region needs to acquire coding properties in one individual/subpopulation. For that, the region of emergence requires both the gain of an open reading frame (ORF) and the acquisition of transcription (Schlötterer 2015; Durand et al. 2019). Moreover, it requires minimal properties of the UTR regions and of the ORF such that the new transcript is translated. Once a nongenic region becomes coding in one individual or a set of individuals, it reaches the stage of a "proto-gene". A proto-gene can be considered as a de novo gene as soon as it has been fixed in the species. It was shown that, while proto-gene emergence is frequent in individuals, most proto-genes revert to a noncoding status via a fast turnover, and only a tiny fraction of them become fixed in a species (Neme and Tautz 2016; Durand et al. 2019; Rödelsperger et al. 2019; Schmitz et al. 2020; Grandchamp et al. 2024). While the gain of ORFs in the emergence of proto-genes has been well studied (Carvunis et al. 2012; Rödelsperger et al. 2019; Wang et al. 2020b; Zhuang and Cheng 2021; Delihas 2022; Grandchamp et al. 2023), how transcription is acquired remains poorly understood.

The transcription of a gene is initiated at the core promoter which is located upstream the gene's 5′ untranslated region (UTR) (Butler and Kadonaga 2002; Haberle and Stark 2018). Core promoters contain specific transcription factor binding-sites (TFBs), such as the *TATA box* or the *Initiator sequence*, that are recognized by transcription factors (TFs) (Boeva 2016). Binding motifs with low identity to the consensus sequence are referred to as "minimal motif" (Wang et al. 2020a). Transcription factors recruit the protein complexes required for transcription (Butler and Kadonaga 2002). However, transcription of low amounts of transcripts can also be initiated by a core promoter alone (reviewed in Haberle and Stark (2018) and Small and Arnosti (2020)). Promoters can also initiate antisense transcripts by initiating transcription in both direction (Scruggs et al. 2015). Furthermore, proximal and distal enhancers regulate the levels of transcription. Proximal enhancers (also called proximal promoters) are located directly upstream of core promoters, while distal enhancers influence transcription over long distances (Kim and Shiekhattar 2015; Haberle and Stark 2018). Both contain TFBs motifs and can increase the amount of transcription initiated by the promoter (Haberle and Stark 2018), independently of their locations and directions (Haberle and Stark 2018). Enhancers often support bi-directional transcription, producing short but unstable transcripts in both directions (Meers et al. 2018; Small and Arnosti 2020). Enhancers and promoters can also occasionally be converted into each other (Majic and Payne 2020), and promoters can be interconnected by successive mutations without completely losing their activity (Kurafeiski et al. 2019).

In a noncoding region, the gain of transcription can result following random point mutations creating a minimal motif and subsequently lead to stable transcription (Kapusta and Feschotte 2014; Palazzo and Lee 2015), as genomes generally contain many such cryptic functional sites with minimal promoter motifs (Kapusta and Feschotte 2014). Other genomic mutations include the insertion of TEs. TEs are mobile DNA sequences that can move and in some cases, amplify in genomes. They can be divided into two classes, based on their transposition mode: RNA and DNA transposons, which are further divided into sub classes and families based on their sequence characteristics (McCullers and Steiniger 2017). Several studies reported major genomic rearrangements caused by TEs, as well as their role in adaption (Delprat et al. 2009; Bourque et al. 2018; Thybert et al. 2018). For example, syncytin genes, enabling cell-cell fusion in mammalian placenta, are derived from retroviruses (Malik 2012). TEs have also aided the evolution of the placenta in mammals by affecting enhancer activity (Chuong et al. 2013). Other mechanisms can influence transcription levels, such as DNA methylation, which represses a gene's transcription in vertebrates via the modulation of TFBs activity (Law and Jacobsen 2010). In invertebrates, methylation patterns are also associated with the regulation of transcription (Dixon and Matz 2022), but the correlation between transcription and methylation is less clear than in vertebrates (Lyko et al. 2000; Dunwell and Pfeifer 2014). Transcription is a highly dynamic and plastic process with high rates of transcripts gain and loss in closely related species, as well as among populations and individuals (Zhao et al. 2014; Neme and Tautz 2016; Schmitz et al. 2018, 2020; Grandchamp et al. 2023, 2024; Iyengar and Bornberg-Bauer 2023), suggesting fast transcript turnover. However, the mechanisms promoting de novo transcripts, i.e. transcription initiation from noncoding regions, remains elusive.

In this study, we investigate the mechanisms underlying new transcript emergence at short evolutionary time scales by studying de novo transcripts in seven samples of inbred lines of *Drosophila melanogaster*, originating from different geographical locations and devoid of allelic diversity (Grandchamp et al. 2023). By using long-read sequencing and short-read sequencing to generate genome and transcriptome assemblies and a common annotation methodology across all genomes, the genomes and transcriptomes present a unique opportunity to precisely categorize de novo transcripts in each *Drosophila* sample by mapping them exactly on the genomes of their individuals and thus investigate the molecular basis underlying the gain of transcription. Indeed, our dataset allows us to compare directly the related DNA sequences that are transcribed in one or several samples but not in the others. In particular, we studied the role of TE insertions and motif enrichment upstream of de novo transcripts that emerged in each *Drosophila* sample. Overall, our analyses reveal that the emergence of transcription is aided by an enrichment of motifs upstream of a DNA sequence, and that this motif enrichment is itself favored by nearby insertion of DNA transposons.

Materials and Methods

Detection of de novo Transcripts and Their 'Non-transcribed Homologs'

To investigate the molecular mechanisms enabling new transcript emergence, we searched for de novo transcripts and their 'nontranscribed homologs' in the transcriptomes and genomes, respectively, of seven lines of *D. melanogaster*, six inbred European lines and one from Zambia (NCBI Bioproject PRJNA929424) (Grandchamp et al. 2023). For each sample, the genomes have been assembled using nanopore long-read sequencing DNA extracted from 50 randomly selected individuals from each line pooled together. The transcriptome assemblies were generated using RNA-seq samples. In each line, RNA was extracted from two males, two females, and one larva pooled together. Each pool was sequenced with illumina paired reads (2 × 150 bp) with at least 92 million reads per sample, and low quality reads below 50 bp were removed with FastQC (Wingett and Andrews 2018). Transcripts were defined as

Table 1 List of reference species used to build the reference database for the blast search

	Species	Accession number	Assembly
1	*Aedes aegypti*	GCA_002204515.1	AaegL5
2	*Anopheles sinensis*	GCA_000472065.2	AsinS2
3	*Culex quinquefasciatus*	GCA_000209185.1	CpipJ2
4	*Drosophila ananassae*	GCA_000005115.1	dana_caf1
5	*Drosophila erecta*	GCA_000005135.1	dana_caf1
6	*Drosophila grimshawi*	GCA_000005155.1	dgri_caf1
7	*Drosophila melanogaster*	GCA_000001215.4	BDGP6.32
8	*Drosophila mojavensis*	GCA_000005175.1	dmoj_caf1
9	*Drosophila persimilis*	GCA_000005195.1	dper_caf1
10	*Drosophila pseudoobscura*	GCA_000001765.2	Dpse_3.0
11	*Drosophila sechellia*	GCA_000005215.1	dsec_caf1
12	*Drosophila simulans*	GCA_000754195.3	ASM75419v3
13	*Drosophila virilis*	GCA_000005245.1	dvir_caf1
14	*Drosophila willistoni*	GCA_000005925.1	dwil_caf1
15	*Drosophila yakuba*	GCA_000005975.1	dyak_caf1
16	*Megaselia scalaris*	GCA_000341915.1	Msca1
17	*Teleopsis dalmanni*	GCA_002237135.2	ASM223713v2

being de novo (i.e. newly emerged) if they met our four criteria: i) detected in one or several samples of the seven inbred line transcriptomes with a TPM value (transcripts per million) above 0.5 (Grandchamp et al. 2024) (transcripts with TPM above 2TPM only were also studied as a backup (supplementary material SI-S10, Supplementary Material online)); ii) no homology to any other annotated transcripts (cRNA and ncNRA) in the *D. melanogaster* reference transcriptome (Table 1); iii) no homology with annotated transcripts (cRNA and ncRNA) of eleven outgroup *Drosophila* and five Diptera species (Table 1); iv) no overlap of transcript genome location with TEs greater than 80%. All de novo transcripts with a TE overlap greater than 80% were considered to be newly expressed TEs and thus treated as a separate category in the followup analysis.

Nucleotide BLAST (version 2.12) (Altschul et al. 1990) with the plus strand option was used to assess homology between inbred *Drosophila melanogaster* samples and reference transcripts. The lack of homology was defined if a transcript did not return a BLAST hit (with a threshold E-value of 0.05), as well as none of its splicing variant.

Bedtools (version 2.3, intersect with default parameters) (Quinlan and Hall 2010) was used to map de novo transcripts onto their respective genome. De novo transcripts overlapping with a gene in sense or antisense direction were filtered out, keeping only intergenic de novo transcripts.

To better understand the frequency of transcription gain and loss, we quantified the amounts of de novo transcripts shared across inbred *D. melanogaster* samples. To that end, a BLAST search (plus strand option, E-value of 0.05) of our de novo transcripts were performed against the transcripts of the other samples. Transcripts were deemed to be homologous if they met those three criteria: i) the transcription start sites of transcripts match up in a 200 nucleotide window; ii) the transcription termination sites of transcripts match up in a 200 nucleotide window; iii) transcripts share at least 80% identity.

To precisely categorize the mechanisms underlying the gain of transcription, direct comparisons of the same nucleotide sequences exhibiting different transcription status is mandatory. We, therefore, used de novo transcripts, which were not found across all samples, and their location onto their respective genome to find their 'nontranscribed homologs'. The unspliced sequences of those de novo transcripts were retrieved using bedtools (get fasta with the -s option)(Quinlan and Hall 2010). Those unspliced sequences were then used to identify similar/identical nucleotide sequences in the genome of other samples, which do not posses this de novo transcript, using a nucleotide BLAST search (default settings, E-value cut-off 0.05) (Altschul et al. 1990). A nucleotide sequence was defined as a 'nontranscribed homolog', if BLAST hits had 80% query coverage with the de novo transcript. If a transcript had multiple 'nontranscribed homologs' in the same sample, only the nucleotide sequence with the lowest E-value, highest percent identity and highest query coverage, was retained.

'Non-transcribed homologs' were searched per transcript instead of per orthogroup. The original dataset was reduced to ensure data consistency i) Alternative spliceforms were reduced to one spliceform per orthogroup; ii) Orthogroup containing samples duplication were removed (iii) All orthogroup member and 'nontranscribed homologs' have their initiation and termination positions in the same window (± 200 nt).

The Contribution of Transposable Elements to The Gain of Transcription

To unravel the importance of TEs in the emergence of de novo transcripts, de novo annotations of TEs were performed in each sample, using the reasonaTE pipeline from the TransposonUltimate software (Riehl et al. 2022). This pipeline was chosen as it combines, compiles, and filters TE annotations from 13 tools with different annotation approaches (Riehl et al. 2022). De novo TE annotations of each *D. melanogaster* sample genome was used to infer their relative overlap with de novo transcripts, as well as with their upstream and downstream regions, with 'nontranscribed homologs' and their upstream regions, and as a control with random intergenic regions of 1,100 bp length obtained using bedtools (Quinlan and Hall 2010). Relative overlap was calculated by dividing the overlap length between a sequence and a TE obtained with bedtools (Quinlan and Hall 2010) with the full of length of the sequence. Up- and down-stream regions were defined as 1,000 bp length before and after a given sequence, respectively, with a 100 bp overlap with the given sequence (for a total of 1,100 bp length). A given sequence could overlap with more than one TE. In this case relative overlap was calculated using all overlapping TEs, and the number of TEs as well as their class and family were calculated.

Moreover, to evaluate features associated with gain of transcription at the genome scale, the distribution of de novo transcripts and TE density within a 10 kb sliding window, as well as CpGoe, were plotted along chromosomes for each *D. melanogaster* sample, using an R script adapted from (Ylla et al. 2021) https://github.com/guillemylla/Crickets_Genome_Annotation.

Motif Enrichment and Gain of Transcription

Motif Datasets

The presence of specific DNA motifs upstream a gene is a major factor enabling transcription. We therefore searched for such motif enrichment upstream of de novo transcripts and control sequences, using custom python scripts along the Bio-python motifs (Cock et al. 2009) package. To that end, two motif databases were downloaded as position frequency matrices (PFM) from JASPAR: the JASPAR Core insects (non redundant) database (Castro-Mondragon et al. 2022) and the JASPAR Pol II database (Fornes et al. 2020), containing 146 TFBS motifs of *D. melanogaster* and 13 core promoter motifs, respectively. While the JASPAR Core insects database was used to find general promoter and proximal enhancer motifs, the JASPAR Pol II database was restricted to the main core promoter motifs. PFM were used to calculate for each motif a position weight matrix (PWM). The PWM was then used to determine a position specific scoring matrix (PSSM). TFBMs show optimal binding to transcription factors when their sequence perfectly matches the highest matrix score. However, they can still bind transcription factors with lower similarity to the reference matrix. We have defined two relative scores of matrix similarity to detect motif enrichment: a relative score of 0.8, referred to as the 'low identity score', which considers all motifs present if their sequence is at least 80% similar to the PFM; and a relative score of 0.95, referred to as the 'high identity score', which considers all motifs present if their sequence is at least 95% similar to the PFM. For each motif, an absolute score threshold was defined based on Formula 1, and motifs with PSSM scores superior to their absolute score were considered for analysis. Proximal promoters are usually located between −200 to +200 bp. However, distal promoters/enhancers can be distributed further upstream. Moreover, there are no published datasets of our inbred lines of *D.melanogaster* species sequenced using a 5′ end-capture approach available, making the transcription initiation site less precise. The computational requirements for motif searching are quite high, and searching for enhancers at high nucleotide distances (e.g. more than 10,000 bp) is risky—because further away enhancers more likely regulate other genes (too). Therefore, motif enrichment was estimated for upstream sequences set to 1,000 bp before transcript start and 100 bp after it of de novo transcript. Such a length is longer than the standard 500 bp region chosen by other studies investigating motifs upstream genes (Corà et al. 2004; Lis and Walther 2016) but is in within the range of lengths commonly used when studying upstream sequences (Reineke et al. 2011; Peng et al. 2016; Wolf et al. 2016). We also studied motif enrichment for upstream sequences of 'non-transcribed' homologs and for random intergenic sequences of 1,100 bp (obtained with bedtools Quinlan and Hall 2010, N = 53,300) as negative controls, and for upstream sequences of annotated genes as a positive control. We also restricted our upstream sequences to 200 bp before a given sequence start and 100 bp after it, to estimate the core promoter binding motifs enrichment as those motifs are expected to be closer to the start of a transcript than general promoter and proximal enhancer (Butler and Kadonaga 2002).

Formula 1:

Absolute score threshold:

$$(pssm.max - psssm.min) * relativescore + pssm.min \qquad (1)$$

In order to detect significantly enriched TFBS motifs, the total amount of motifs from the TFBS database were counted in all upstream regions and in the intergenic control sequences. Anova tests were used to detect all motifs that are significantly more abundant in gene/de novo transcript/TE upstream regions compared to the intergenic control. Hence significantly enriched motifs refer to motifs that are enriched in at least one category of de novo transcript upstream regions (or gene upstream region) compared to the intergenic control.

All comparisons of transcripts with other sequence types were performed using Generalized Linear Mixed Models (GLMMs) using the package glmmTMB (Magnusson et al. 2017), retaining the best model after simplifying the model with a step-wise factor deletion.

Transcripts vs. 'Non-transcribed Homologs'

To unravel the differences among sequences leading to transcription, four GLMMs were built using a binomial distribution. The first one assessed the importance of TE overlap, number and presence / absence in gaining transcription. This model includes as a dependent variable the type of sequence (transcript or 'nontranscribed' homolog), as fixed factors the relative overlap with TEs, the number of overlapping TEs, the presence or absence of overlapping TEs, the regions of the sequence (upstream, sequence, downstream), and their interactions. Moreover to account for pseudo-replication, the orthogroup ID of the sequence (single ID shared among transcript and 'nontranscribed homologs') and *D. melanogaster* sample were added as random variables into the GLMM. A second model to account only for motif enrichment was built with as fixed factors the number of minimal and optimal TFBS motifs and the minimal and optimal number of core promoter. A third model to account simultaneously for TEs and the different motifs was built by adding as fixed factor the number of the different motifs (motifs, cores, low and high). Finally, a fourth model was built to disentangle the impact of different TE classes (DNA vs. RNA transposon) on transcript and 'nontranscribed homologs', by adding the TE class as a fixed factor.

Transcripts vs. Genes and Intergenic Regions

To understand how transcripts differ from genic and intergenic regions, three GLMMs were built. The first GLMM compares the relative overlap of transcripts with TEs with the different sequence types, using a zero-inflated Gamma distribution and as dependent variable: the sequence type, as fixed factor: the relative of overlap with TEs, and a random variable: the *D. melanogaster* sample. The second GLMM compares the sequence types in term of motif numbers, using a poisson distribution and as dependent variable: the number of motifs / cores, as fixed factor: the sequence type, and a random variable: the *D. melanogaster* sample. The third GLMM accounts for differences of sequence features among the different sequence types, using a zero-inflated Gamma distribution and as dependent variable: the sequence type, as fixed factor: the log TPM, the guanine-cytosine (GC) content, spliced length, and exon number, and a random variable: the *D. melanogaster* sample. As the data-sets were of unequal sample size among the different sequence types and to ensure the robustness of our results, *P*-values of the best GLMM was bootstrapped using data-sets with equal sample size, using the package boot (Canty and Ripley 2017).

Furthermore, the density of de novo transcript per 100 kb was correlated to its distance to the center of the chromosome and the density of TEs, using GLMMs with (i) as a dependent variable the number of de novo transcript within a 100 kb window; (ii) as a random variables the chromosome and sample, and (iii) as an explanatory variable the distance from the center of the chromosome (scaled) and the density of TE per 100 kb (scaled). Furthermore, the levels of CpGoe of de novo transcripts was correlated with their relative overlap with TEs, using a GLMM with CpGoe value as a dependent variable, length of overlap with a TE as explanatory variable and chromosome and sample as a random variable.

For all models we did not account if explanatory variables co-vary within our models as covariance matrices were not implemented within them. It is highly likely that some explanatory variables co-vary: i.e. the log of TPM, the GC content, spliced length and exon number. Nevertheless, accounting for co-variance of those variables for the model may just reveal a weaker of one variable compared to another one, but it is unlikely to change completely the result of the model. For the other models, either only one explanatory variable was used or they are unlikely to co-vary (TFBS motif number with TE number) or were accounted for separately.

Visualization

All graphs and statistics were created with R version >4.1 (Team 2022). The packages dplyr (Wickham et al. 2022), tidyverse (Wickham et al. 2019) and data.table (Dowle and Srinivasan 2021) were used for data preparation. The plots were mainly done with ggplot2 (Wickham et al. 2016) and its extensions ggpubr (Kassambara and Kassambara 2020).

Results

General Characteristics of de novo Transcripts

To characterize the molecular basis underlying gains of transcription, we used a conservative approach to define de novo transcripts, ensuring detection of strictly de novo transcript (see Materials and Methods). Such definition and filtering led to the discovery of between 403 (Sweden [SE]) and 628 (Ukraine [UA]) de novo transcripts across *D. melanogaster* samples (mean = 504 ± 28.04 (standard error), Fig. 1a, Supplementary text, Supplementary Material online). De novo transcripts were unevenly distributed among and along chromosomes, with the highest numbers of de novo transcripts in 3L and 3R chromosome arms (supplementary material SI-S1, Supplementary Material online). Most de novo transcripts were unique for their *D. melanogaster* sample (2,389/3,528) and only a few (38) were shared among all samples, suggesting a high birth/death rate of de novo transcripts (Grandchamp et al. 2024), Fig. 1b. This high birth/death rate of de novo transcripts is likely the result of gain/loss of transcription, as most de novo transcripts (14,058 identified homologs out of 18,903 potential homologs in a maximum of 6 samples) had a 'nontranscribed' homolog in at least one other *D. melanogaster* sample (Supplementary text, Supplementary Material online). Moreover, de novo transcripts show different patterns compared to annotated transcripts (both genes and noncoding RNAs), with de novo transcripts having lower expression level, GC content, exon number, and spliced length compared to annotated transcripts (GLMM: TPM: $P < 0.001$, GC content: $P < 0.001$, exon number: $P < 0.001$, spliced length: $P < 0.001$, (supplementary material SI-S2, Supplementary Material online)).

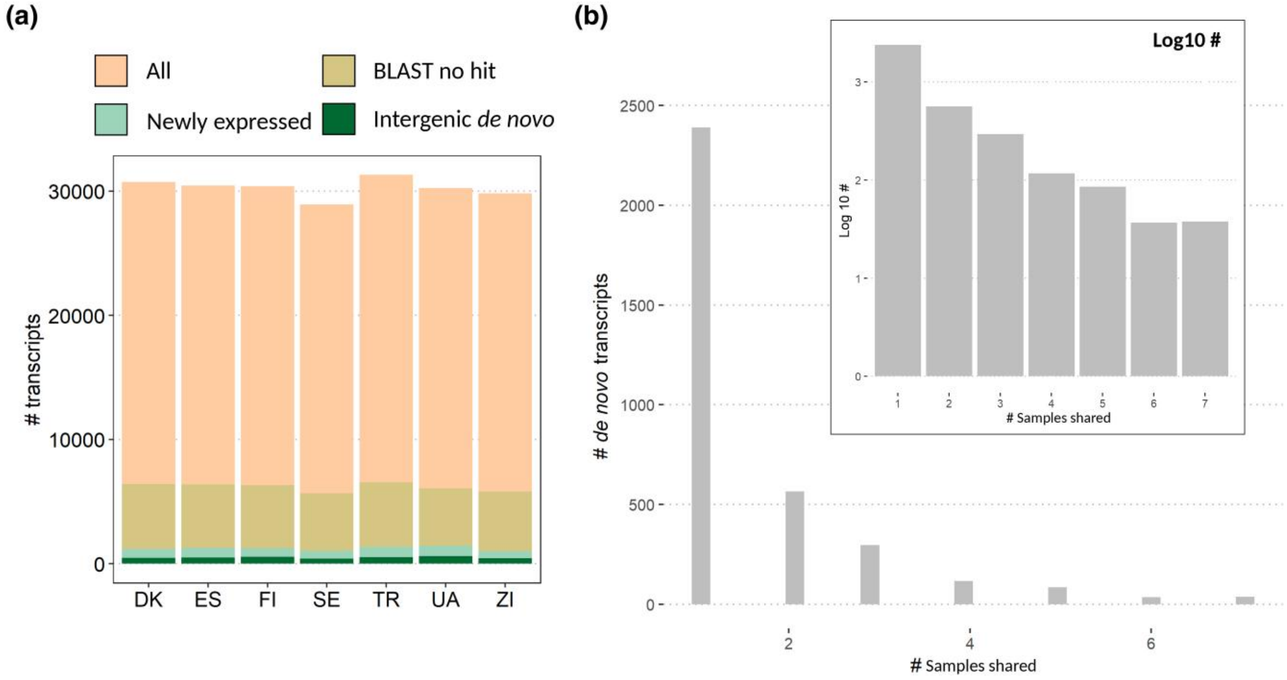


Fig. 1. De novo transcripts. a) Number of transcripts after filtering steps. The beige bar represents all transcripts detected with transcriptome assembly. The gray bar represents all transcripts without a BLAST hit. The green bar represents de novo transcripts after filtering for TPM and splicing. The dark green bar represents only the intergenic de novo transcripts after filtering out transcribed TE. Each color is a subset of the largest group corresponding to all transcripts. b) Number of de novo transcripts shared by samples. The insert plot shows the log transformed numbers.

DNA Transposon Insertions Favor The Gain of Transcription

For each genome, we performed a de novo annotation of TEs, using the TransposonUltimate pipeline (Riehl et al. 2022) (Method, Supplementary text, Supplementary Material online). To understand how TEs can favor the gain of transcription, we first assessed the relationship between these annotated TEs and de novo transcripts at the chromosome scale (Fig. 2). De novo transcripts were unevenly distributed along chromosomes. Specifically, de novo transcript densities were positively correlated with TE densities at a 100 kb scale (GLMM: $P < 0.001$). Moreover, annotated TEs and expressed TEs were in higher density in the telomere regions of chromosomes (GLMM: $P < 0.001$; Fig. 2a, supplementary material SI-S3, Supplementary Material online). Finally, de novo transcripts displayed low levels of GC (CpG: mean & median = 0.902 ± sd 0.222). Moreover, their CpGo/e values were negatively correlated with TE density (GLMM: $P < 0.001$; supplementary material SI-S4, Supplementary Material online)).

In addition to our chromosome scale analyses, we also investigated the impact of TE insertions on de novo transcripts by comparing the number of TEs overlapping with these transcripts, as well as their down- and upstream regions, using random intergenic regions as a control (see Materials and Methods). De novo transcripts displayed a higher amount of TE insertions compared to their up- and downstream sequences, however with a lower length of TE overlap (GLMM: $P < 0.001$; Fig. 3a,b, Supplementary text, Supplementary Material online).

Furthermore, to precisely pinpoint the role of TE insertions for the gain of transcription, we directly compared de novo transcripts with their 'nontranscribed' homolog sequences present in other *D. melanogaster* samples. Our analyses revealed that the average number of TE insertions did not differ between de novo transcripts and 'nontranscribed' homologs. However, de novo transcripts displayed shorter overlaps with TEs in their sequence and in their up- and downstream regions as well as a lower number of TE insertions compared to 'nontranscribed' homologs (GLMM, $P < 0.001$, supplementary material SI-S5, Supplementary Material online). We then compared the proportion of transcripts and homologs overlapping with either of the two TE classes (or both). RNA TEs were less abundant (present in 5.15% of all regions) in de novo transcripts compared to 'nontranscribed' homologs (present in 16.10%, GLMM, $P < 0.001$, Fig. 3c, supplementary material SI-S5, Supplementary Material online). On the contrary, DNA TEs were more abundant in de novo transcripts compared to 'non-transcribed' homologs (Proportion sequences with DNA TE overlap (all regions): 15.97% (homologs) vs 17.80% (transcripts), GLMM, $P < 0.001$).

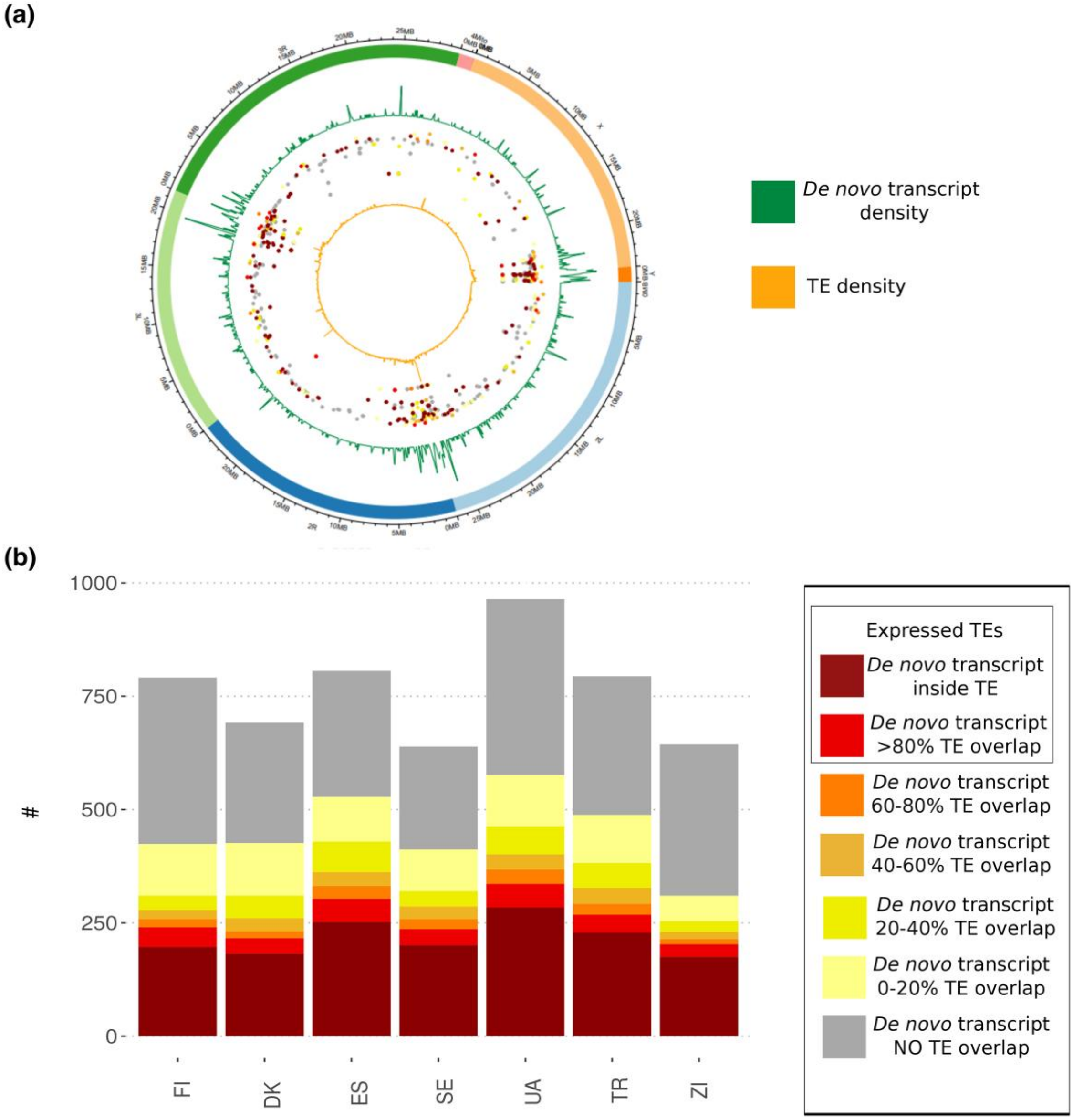


Fig. 2. De novo transcripts and TE density among the chromosomes. The circular plots represent the *D.melanogaster* sample collected in Denmark (DK). Plots with similar distributions can be found for all other samples in the supplementary material data, Supplementary Material online. The 8 chromosome arms are represented with specific colors. In the two circle plots, the green circles represent de novo transcripts, and the yellow lines represent TEs distributions. a) The colored dot represents expressed TEs and de novo transcripts distribution according to their relative overlap with TEs. b) de novo transcripts overlap with annotated TEs. Transcripts are colored based on their percentage of overlap with TEs. Transcripts with more than 80% overlap are considered to be expressed TEs.

Additionally, about 5.55% of the homolog and 3.08% of the transcript regions overlap at least one TE from each class with the remaining regions overlapping no TE. Our results thus highlight a higher impact of DNA TEs (DNA vs. RNA) on transcription gain.

Motif Enrichment

A major factor influencing gene expression is the presence of specific DNA motifs enabling the transcription machinery to bind to the DNA region. We therefore investigated the role of DNA binding motifs for the gain of transcription. We compared motif enrichment of TFBS motifs from enhancers and distal promoters, as well as core promotor motifs upstream of our de novo transcripts. As positive controls, we studied regions upstream of conserved *Drosophila melanogaster* genes and expressed TEs. As negative control we included random intergenic regions. Core motifs are located within 40 bp around the transcription start side (TSS) (Butler and Kadonaga 2002). To capture

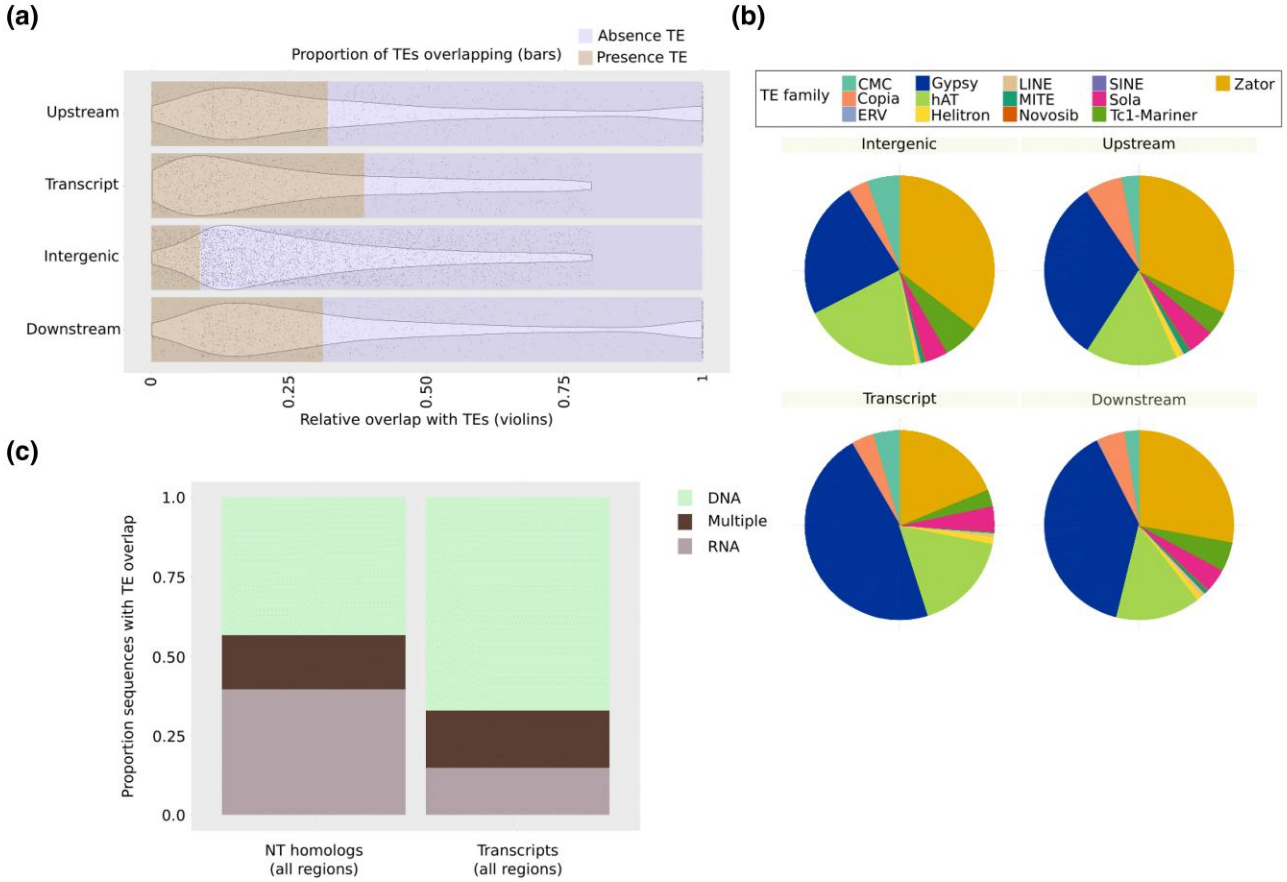


Fig. 3. TEs overlap. a) Relative sequence overlap with TEs and number of sequences overlapping with TEs into four datasets: Intergenic sequences, upstream sequences of de novo transcripts, downstream sequences of de novo transcripts, de novo transcripts. b) Normalized proportion of TEs overlapping de novo transcripts, their up-/downstream regions and intergenic sequences according to the TE category c) Major classes of TEs overlapping with de novo transcripts and their 'nontranscribed homologs'. NT homologs refer to the 'nontranscribed homologs' of the de novo transcripts.

them, the region between −200 bp upstream and +100 bp downstream of the TSS was searched. Proximal enhancers are commonly placed within 200 bp around the TSS. Distal enhancers can be located much further upstream (Butler and Kadonaga 2002). Therefore, TFBS motifs were searched in a set region between −1,000 bp upstream and +100 bp downstream of the TSS. Motif enrichments were further divided into two classes (high and low identity motifs), according to their thresholds of similarity to their PSSM matrix (Fig. 4, see Method). Low identity motifs (minimal motifs), were required to have a score of identity higher than 80% ID to the PSSM. For high identity motifs a score of 95% was required.

Our analysis of TFBS motifs revealed that TEs as well as de novo transcripts overlapping with TEs have higher number of low identity TFBS motifs compared to de novo transcripts without TE overlap, as well as intergenic sequences and genes (GLMM: $P < 0.001$). Moreover, de novo transcripts that do not overlap with TEs had higher numbers of core promoter motifs with a high identity score (mean: 2.98) than TEs (mean: 2.80) and de novo transcripts with TE overlap (mean: 2.96) (GLMM: $P < 0.001$; supplementary material SI-S6,S7, Supplementary Material online). Overall, genes and intergenic regions displayed a higher enrichment of both high and low identity core promoter motifs in the 300 nt regions and of TFBS motifs with high identity score in the 1,100 nt regions. On the other hand, TEs as well as de novo transcripts with TE overlap (mean: 1,029.83) and without TE overlap (mean: 996.33) displayed an enrichment of TFBS motifs of low identity score in the 1,100 nt regions (GLMM, $P < 0.001$, Fig. 4, supplementary material SI-S6,S7, Supplementary Material online).

When studying TFBS motifs in the 1,100 nt regions individually (supplementary material SI-S6,S7, Supplementary Material online), 13 motifs were enriched upstream de novo transcripts and TEs, compared to intergenic regions, with a high threshold (high identity score: relative score = 0.95; supp data). Four of those were also

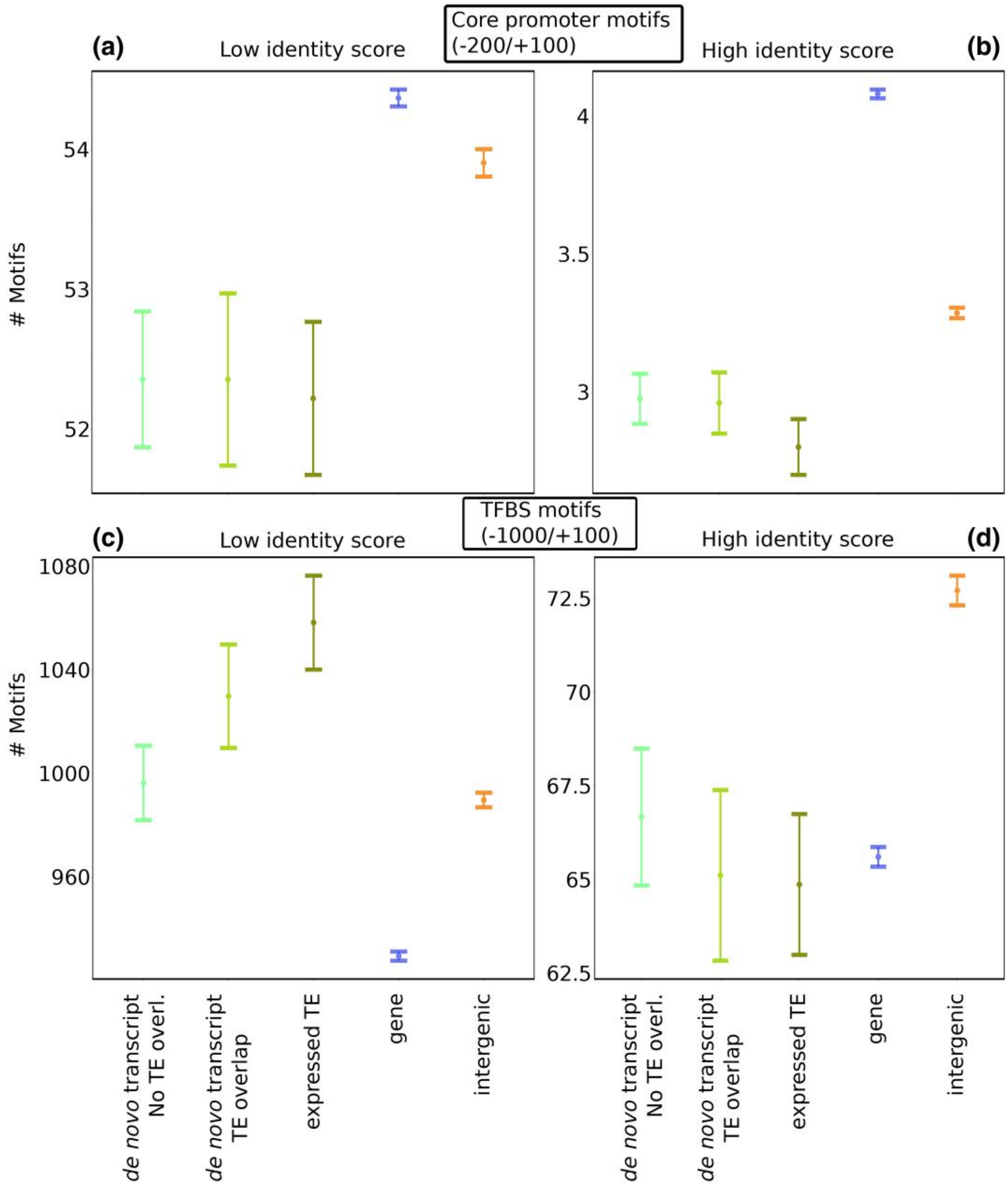


Fig. 4. Number of motifs detected upstream of five sequences datasets. a) Number of minimal Core promotor (0.8) motifs detected upstream i) de novo transcripts overlapping no TE, ii) de novo transcripts overlapping with TEs, iii) Newly expressed TEs, iv) conserved genes, v) randomly selected intergenic regions that are not transcribed. b) Number of high identity Core promotor motifs (0.95) detected upstream the aforementioned dataset of sequences c) Number of minimal TFBS motifs (0.80) detected upstream the aforementioned dataset of sequences. d) Number of high identity TFBS motifs (0.95) detected upstream the aforementioned dataset of sequences. The y axis of each plot has a different scale.

significantly enriched in upstream regions of conserved genes. Three of these 13 motifs were significantly enriched upstream de novo transcripts without TE overlap. Eleven of these 13 motifs were specific for homeo domain factors, one for a zinc finger factor (Supplementary text, Supplementary Material online). We then investiated if any of the ten most abundant motifs (ara, mirr, CG4328-RA, lbe, PHDP, H2.0, Deaf.1, caup, C15, lbl) from the dataset at the high treshold were enriched in de novo transcript upstream regions. Four were enriched in transcribed TEs and in TEs overlapping de novo transcripts.

When studying TFBS motifs in the 1,100 nt regions using a low threshold (low identity score: relative, score = 0.8, we found 78 TFBS motifs that were enriched upstream de novo transcripts and TEs, 13 of them being also significantly enriched in conserved genes. Only 18 of these 78 were

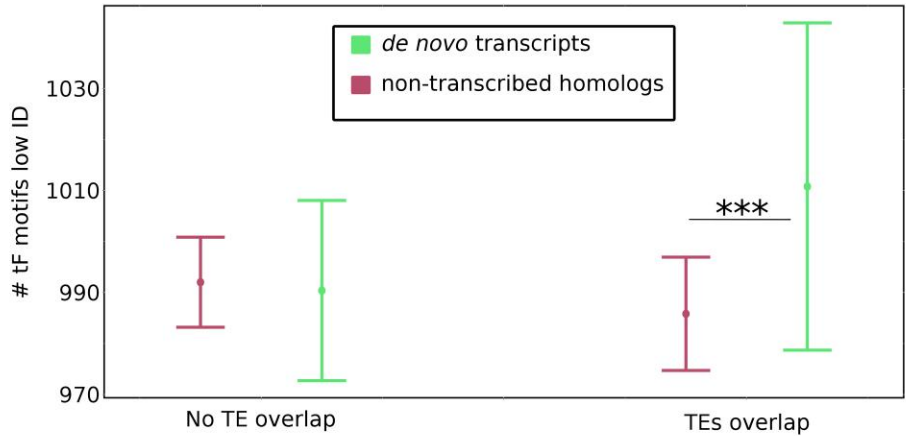


Fig. 5. Enrichment in low TFBS motifs upstream de novo transcripts and their nontranscribed homologs. The bars on the left represent de novo transcript and homologous sequences without TE overlap, while the bars on the right represent sequences with TE overlap.

enriched upstream of de novo transcripts that did not overlap any TE. Most of these 78 motifs were specific TFBS for homeo domain factors or zinc fingers, however they also included one motif for a high mobility group domain factor, one for a heat shock factor, two motifs for leucine zipper factors, two for paired box factors, one for a fork head/winged helix factor, and one each for a STAT and a TEA domain factor respectively. Among the ten most abundant motifs (CG4328-RA, br, H2.0, PHDP, C15, vvl, Dbx, ct, lbl, ara) from the dataset using a treshold of 0.8, seven were enriched in both de novo transcript categories as well as TEs, one of those also in conserved genes. Another two were enriched only in TEs and de novo transcripts overlapping TEs.

In addition, we directly compared the enrichment of binding motifs in upstream sequences of de novo transcripts and their ‘nontranscribed’ homologs. We observed no significant difference in motif enrichment between de novo transcripts and their ‘nontranscribed’ homologs. The best statistical model (based on AIC, BIC, ICOMP, and Cp) included the enrichment of low identity core promoters but it showed no significance (GLMM, $P = 0.136$, supplementary material SI-S8, Supplementary Material online). Furthermore, we implemented the impact of TE insertions along motif enrichment between de novo transcripts and their ‘nontranscribed’ homologs. De novo transcripts exhibit, when TE inserted, a higher density of TFBS motifs of low identity compared to their homologs. This suggests that TE insertions could enable transcription through low TFBS motif enrichment (Fig. 5), e.g. by providing TF binding sites sufficient for transcription. Finally, we investigated the different TE class (DNA vs. RNA) inserting among de novo transcripts and their ‘nontranscribed homologs’. While de novo transcripts overlap less with RNA transposons than their “nontranscribed” homologs, RNA transposons overlapping with de novo transcripts are highly enriched in low identity core promoter motifs. (GLMM, $P < 0.001$, supplementary material SI-S8, Supplementary Material online). Therefore, RNA transposons could play a role in the gain of transcription by providing core promotor motifs that, while less similar to the consensus, are still sufficient to enable transcription.

Discussion

Detection of de novo Transcripts

To understand how transcription can be gained in intergenic regions leading to the emergence of de novo genes, we searched for de novo intergenic transcripts that emerged in seven samples of *Drosophila melanogaster*. Our stringent definition led to the discovery of 3,799 transcripts over 7 *D. melanogaster* samples, with an average of 504 intergenic de novo transcripts per sample. This amount of de novo transcripts, while being lower than in a previous study of new transcripts emergence in samples (Everett et al. 2020), corresponds well to previous estimates (Huang et al. 2015; Camilleri-Robles et al. 2022), if we account only for intergenic de novo transcripts.

Moreover, the characteristics of our de novo transcripts corresponds well to those of previous studies, namely a lower expression, lower GC content, lower number of exons, and a shorter sequence than known genes. Finally, our estimation of de novo transcripts could have been underestimated by not accounting for transcripts with low level of expression or tissue- and life-stage specific expression, resulting in lower detection of de novo transcripts (Grandchamp et al. 2023).

Overlap with Transposable Elements

34% of the newly expressed transcripts corresponded to TEs due to an overlap of more than 80% of their sequence.

Moreover, TE insertions were divergent between samples when comparing transcripts to their 'nontranscribed homologs', indicating significant mobility of TEs within the species. Overall, TEs were predominantly located and transcribed in telomeric regions, consistent with previous reports (Kordyukova et al. 2018). TEs overlapped with newly expressed transcripts more frequently than expected by chance, primarily with small sub-sequences of the transcripts. This correlation between TE overlap and new transcription events suggests that TE insertion could have contributed to the emergence of new transcripts unrelated to TE expression.

When comparing de novo transcripts with their 'nontranscribed homologs', they did not differ in their proportion of TE insertions. However when focusing on TE classes de novo transcripts display a higher proportion of DNA transposons compared to their homologs.

These results suggest first that de novo transcripts emerge in regions that are prone to TE insertion, and are highly variable in TE proportions due to TEs activity. Second, given that DNA TEs are more associated with new transcription events, the insertion of DNA rather than RNA TEs seems to be the more likely event to initiate new transcription. Interestingly, the main difference in TE composition of de novo transcripts compared to intergenic sequences was the higher amount of overlap with retrotransposons, mainly LTR elements from the gypsy family. In *D. melanogaster* certain TEs, such as LTR retrotransposons are reported to be more active than others (Petrov et al. 2011; Kofler et al. 2015). High TE activity can also strongly scramble genomes. This could explain why 25% of de novo transcripts had no detected transcribed homolog when requiring a high degree (80% identity) of sequence similarity between transcript and homolog.

Taken together, our results corroborate that TEs are actively transposing in *D.melanogaster*, and that such activity is noticeable even between samples/ individuals. This aligns with previous studies reporting high activity of several TE families in *Drosophila* (Kofler et al. 2015; Bourque et al. 2018; Mérel et al. 2020; Lawlor et al. 2021). Additionally aligning with previous reports on high retrotransposon activity in *Drosophila* (Kofler et al. 2015), we find that the majority of the identified newly expressed TEs are LTR retrotransposons (supplementary material SI-S5, Supplementary Material online). Moreover, the significant overlap of active TEs with de novo transcripts strongly suggests that TE activity plays a role in initiating new transcription events in intergenic genome regions.

Minimal TFBS Motifs Enrichment Leads to Transcription Gain

Intergenic regions of genomes are known to contain a high proportion of (distal) enhancers which interact with very distant promoters (Small and Arnosti 2020). That was confirmed in our results, with random intergenic sequences being the most enriched in highly conserved TFBS motifs. However, when studying motifs with lower scores of similarity to annotated motifs (80% ID), de novo transcripts contained the highest amount of such motifs, compared to genes and intergenic sequences. Indeed, such low TFBS motifs, also called suboptimal transcription factor motifs, appear to be a significant factor for initiating new transcription in genomes. De novo transcripts showed lower expression levels than expressed genes, in line with the finding that transcription is initiated at low levels without the presence of canonical core motifs (Palazzo and Lee 2015).

While de novo transcripts showed high motif enrichment of minimum TFBS motifs, upstream regions of transcripts overlapping with TEs showed the highest amount of low identity TFBS motifs. Such enrichment was still lower than in TEs. Most TEs possess a machinery for transcription, which necessitates the presence of TFBS motifs in their sequence (Chuong et al. 2017). The enrichment of low TFBS motifs upstream of de novo transcripts overlapping with TEs opens two hypothesis. First, the insertion of new TEs in previously untranscribed genomic location could provide sufficient sequence disruption to mutate into TFBS motifs. TFBS motifs are usually shorter than 15 nucleotides, and several positions allow nucleotide variability without affecting the binding. Therefore, some of the motif emergences were likely due to mutations caused by TE insertions. As a second hypothesis, new transcripts could have benefited from the presence of TFBS motifs in TEs which are used by TEs to initiate new transcription events. While both hypotheses could find support in literature (Chuong et al. 2017; Moschetti et al. 2020), our data seem to give more credit to the second one. Indeed, TFBS enrichment with low similariry to the consensus was observed in de novo transcript s, compared to their'nontranscribed homologs', and only when a TE insertion within the transcript sequence was present. Furthermore, while de novo transcripts and their homologs shared similar proportions of TE insertions, the TE content of de novo transcripts and their homologs differ. De novo transcripts overlap more with DNA TEs, while 'nontranscribed homologs' overlaps more with RNA TEs. Intriguingly, the two TE classes do not carry the same TFBS motifs, as their insertion mechanisms differ. Indeed, our results tend to suggest that DNA TE insertion generates more new transcription events, and that this could be due to the recycling of the DNA TE's TFBS motifs.

Many different regulatory elements were shown to have been gained through a TE insertion (Moschetti et al. 2020), such as enhancers/enhancer-like elements (Chung et al. 2007), promoters (Batut et al. 2013), splice sites (Ding et al. 2016), cis-regulatory elements (González et al. 2008) and poly-A signals (Mateo et al. 2014). In noncoding regions, transcription can also be initiated through TEs

(Kapusta and Feschotte 2014). TEs have been shown to have the ability to induce a regulatory sequence through different mechanisms such as domestication (use of TEs for a new function), gene duplication, change of gene expression and ectopic recombination (Rizzon et al. 2002; Kapusta and Feschotte 2014; Moschetti et al. 2020). About 75% of human and 68% of the mouse lncRNAs include at least one (partial) retrotransposon insertion (Kapusta et al. 2013). In humans, TEs provided up to 23% of non redundant transcription start sites and about 30% of poly-A sites of lncRNAs. (Ganesh and Svoboda 2016). In *Drosophila*, TE content has been shown to be high in long noncoding RNAs (Ganesh and Svoboda 2016; Fort et al. 2021), compared to protein coding genes. Since the newly expressed transcripts are primarily sample-specific and likely not yet selected for coding functions, they appear to share a high frequency of TE insertions with lncRNAs.

Indeed, TEs (and especially DNA TEs), could have played a (partial) role in the gain of transcription of new transcripts, e.g. by inserting the motifs enabling the start of transcription. Our outcomes demonstrate that this gain of transcription through TEs is possible, and can occur independently in different samples from the same species. Determining how exactly the TEs lead to the transcription of these regions and which elements (poly-A, promoter, enhancer etc.) they contributed for insertion would need further investigation and more detailed comparisons between the transcript (and up- and downstream) sequences and their homologous regions in the outgroup samples.

In total, 60% of de novo transcripts emerged without overlapping with TEs. These transcripts showed higher enrichment in low similarity TFBS motif enrichment than control sequences, but the difference was less obvious than for transcripts overlapping with TEs. Such small enrichment could be explained by the emergence of low identity TFBS motifs by other mechanisms than TE insertions, like indels, or other sequence reshuffling that we did not investigate, e.g. genomic inversions or duplications. Furthermore, we only regarded TE insertion as a source of new transcript initiation directly linked to the insertion site. However, TE insertion could potentially generate new transcription regulation motifs upstream of a new transcription site. Therefore, the lack of TE overlap with a de novo transcript does not necessarily rule out the possibility that a TE insertion further upstream has influenced transcription initiation. Also, we found surprisingly low amounts of core promoter motifs upstream de novo transcripts. If such motif enrichment was suspected to be lower than in regions upstream genes, it was surprising to find them less enriched than in the intergenic control. One explanation could be that core promoter motifs are highly present upstream of conserved genes involved in developmental regulation (Sloutskin et al. 2021; Georgakopoulos-Soares et al. 2022), which might not be well represented in our dataset.

Conclusion

Overall, our study reveals the importance of TEs in transcription gain and loss. At a large scale, a high TE density seems to enable transcription, most likely through changes of chromatin organization (Lawson et al. 2023), as TE density was correlated with de novo transcripts density within 100 kb windows. At a finer scale, insertions of TEs seems to lead to different outcomes depending on their insertion patterns. Indeed, a singular insertion of DNA transposons shortly overlapping with the transcript sequence tends to favor the gain of transcription, most likely through enrichment of the upstream region with low identity TFBS motifs, although the alternative mechanism of TEs inserting into newly transcribed regions cannot yet be ruled out. Additionally, insertions of RNA transposons, while overlapping lower proportions of de novo transcripts compared to their homologs, are accompanied by an enrichment in minimal core promotor motifs in de novo transcripts upstream regions. This suggests that, while less frequent, they could also contribute to the gain of new transcription events but seem to provide different motifs.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Author Contributions

M.K.L.: Data Curation, Formal Analysis, Investigation, Methodology, Writing and Reviewing original draft; B.F.: Formal Analysis, Investigation, Methodology, Writing and Reviewing original draft; J.S.: Data Curation, Formal Analysis; E.B.B.: Funding Acquisition, Reviewing original draft; A.G.: Conceptualization, Funding Acquisition, Project Administration, Supervision, Validation, Writing and Reviewing original draft.

Funding

M.K.L. and A.G. acknowledge funding from the Deutsche Forschungsgemeinschaft priority program "Genomic Basis of Evolutionary Innovations" (SPP 2349), project BO 2544/20-1 awarded to E.B.B. A.G. and E.B.B. acknowledge funding from grant 3.3-1213745-Fra-HFST-P from the Alexander von Humboldt-Stiftung. B.F. was funded by the European Union REA through a EU-H2020 Marie Skłodowska-Curie Action, Grant Number 101024100 (TEEPI). We acknowledge support from the Open Access Publication Fund of the University of Münster.

Conflict of Interests

The authors declare no competing interests.

Data Availability

The files containing processed data is available in the Zenodo archive https://doi.org/10.5281/zenodo.8403184, and is referred in the main text as “Supplemental Deposit”. Supplemental figures, information, analyses and models are found in the Supplementary Information (SI). All programs are stored on GitHub (https://github.com/MarieLebh). The position frequency matices (PFM) of the studied motifs can be downloaded from https://jaspar2020.genereg.net/collection/POLII/ (Pol II database for core motifs) and https://jaspar2022.genereg.net/downloads/ (tFBS motifs, download the insect core non redundant database).

Literature Cited

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

Baalsrud HT, Tørresen OK, Solbakken MH, Salzburger W, Hanel R, Jakobsen KS, Jentoft S. De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. Mol Biol Evol. 2018:35(3):593–606. https://doi.org/10.1093/molbev/msx311.

Batut P, Dobin A, Plessy C, Carninci P, Gingeras TR. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. Genome Res. 2013:23(1):169–180. https://doi.org/10.1101/gr.139618.112.

Boeva V. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. Front Genet. 2016:7:24. https://doi.org/10.3389/fgene.2016.00024.

Bornberg-Bauer E, Hlouchova K, Lange A. Structure and function of naturally evolved de novo proteins. Curr Opin Struct Biol. 2021:68:175–183. https://doi.org/10.1016/j.sbi.2020.11.010.

Bornberg-Bauer E, Schmitz J, Heberlein M. Emergence of de novo proteins from ‘dark genomic matter’ by ‘grow slow and moult’. Biochem Soc Trans. 2015:43(5):867–873. https://doi.org/10.1042/BST20150089.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, et al. Ten things you should know about transposable elements. Genome Biol. 2018:19(1):1–12. https://doi.org/10.1186/s13059-018-1577-z.

Butler JE, Kadonaga JT. The RNA polymerase ii core promoter: a key component in the regulation of gene expression. Gene Dev. 2002:16(20):2583–2592. https://doi.org/10.1101/gad.1026202.

Camilleri-Robles C, Amador R, Klein CC, Guigó R, Corominas M, Ruiz-Romero M. Genomic and functional conservation of lncrnas: lessons from flies. Mamm Genome. 2022:33(2):328–342. https://doi.org/10.1007/s00335-021-09939-4.

Canty A, Ripley B. Package ‘boot’. *Bootstrap Functions. CRAN R Proj*.2017.

Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. Proto-genes and de novo gene birth. Nature. 2012:487(7407):370–374. https://doi.org/10.1038/nature11184.

Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N, et al. Jaspar 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2022:50(D1):D165–D173. https://doi.org/10.1093/nar/gkab1113.

Chung H, Bogwitz MR, McCart C, Andrianopoulos A, Ffrench-Constant RH, Batterham P, Daborn PJ. Cis-regulatory elements in the accord retrotransposon result in tissue-specific expression of the Drosophila melanogaster insecticide resistance gene Cyp6g1. Genetics. 2007:175(3):1071–1077. https://doi.org/10.1534/genetics.106.066597.

Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 2017:18(2):71–86. https://doi.org/10.1038/nrg.2016.139.

Chuong EB, Rumi M, Soares MJ, Baker JC. Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat Genet. 2013:45(3):325–329. https://doi.org/10.1038/ng.2553.

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009:25(11):1422–1423. https://doi.org/10.1093/bioinformatics/btp163.

Corà D, Di Cunto F, Provero P, Silengo L, Caselle M. Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrep-resented upstream motifs. BMC Bioinformatics. 2004:5(1):1–12. https://doi.org/10.1186/1471-2105-5-57.

Delihas N. An ancestral genomic sequence that serves as a nucleation site for de novo gene birth. PLoS ONE. 2022:17(5):e0267864. https://doi.org/10.1371/journal.pone.0267864.

Delprat A, Negre B, Puig M, Ruiz A. The transposon galileo generates natural chromosomal inversions in Drosophila by ectopic recombination. PLoS ONE. 2009:4(11):e7883. https://doi.org/10.1371/journal.pone.0007883.

Ding Y, Berrocal A, Morita T, Longden KD, Stern DL. Natural courtship song variation caused by an intronic retroelement in an ion channel gene. Nature. 2016:536(7616):329–332. https://doi.org/10.1038/nature19093.

Dixon G, Matz M. Changes in gene body methylation do not correlate with changes in gene expression in Anthozoa or Hexapoda. BMC Genomics. 2022:23:234. https://doi.org/10.1186/s12864-022-08474-z.

Dowle M, Srinivasan A. data. table: extension of “data. frame” [R package data]. table version 1.14. 2.2021.

Dunwell TL, Pfeifer GP. Drosophila genomic methylation: new evidence and new questions. Epigenomics. 2014:6(5):459–461. https://doi.org/10.2217/epi.14.46.

Durand É, Gagnon-Arsenault I, Hallin J, Hatin I, Dubé AK, Nielly-Thibault L, Namy O, Landry CR. Turnover of ribosome-associated transcripts from de novo orfs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. Genome Res. 2019:29(6):932–943. https://doi.org/10.1101/gr.239822.118.

Everett LJ, Huang W, Zhou S, Carbone MA, Lyman RF, Arya GH, Geisz MS, Ma J, Morgante F, Armour GS, et al. Gene expression networks in the Drosophila genetic reference panel. Genome Res. 2020:30(3):485–496. https://doi.org/10.1101/gr.257592.119.

Fornes O, Castro-Mondragon JA, Khan A, Van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranaši⍰ D, et al. Jaspar 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2020:48(D1):D87–D92. https://doi.org/10.1093/nar/gkz1001.

Fort V, Khelifi G, Hussein SM. Long non-coding RNAs and transposable elements: a functional relationship. Biochim Biophys Acta (BBA)-Mol Cell Res. 2021:1868(1):118837. https://doi.org/10.1016/j.bbamcr.2020.118837.

Ganesh S, Svoboda P. Retrotransposon-associated long non-coding RNAs in mice and men. Pflügers Archiv-Eur J Physiol. 2016:468(6):1049–1060. https://doi.org/10.1007/s00424-016-1818-5.

Georgakopoulos-Soares I, Victorino J, Parada GE, Agarwal V, Zhao J, Wong HY, Umar MI, Elor O, Muhwezi A, An J-Y, et al. High-throughput characterization of the role of non-b dna motifs on promoter function. Cell Genom. 2022:2(4):100111. https://doi.org/10.1016/j.xgen.2022.100111.

González J, Lenkov K, Lipatov M, Macpherson JM, Petrov DA. High rate of recent transposable element–induced adaptation in Drosophila melanogaster. PLoS Biol. 2008:6(10):e251. https://doi.org/10.1371/journal.pbio.0060251.

Grandchamp A, Kühl L, Lebherz M, Brüggemann K, Parsch J, Bornberg-Bauer E. Population genomics reveals mechanisms and dynamics of de novo expressed open reading frame emergence in Drosophila melanogaster. Genome Res. 2023:33(6):872–890. https://doi.org/10.1101/gr.277482.122.

Grandchamp A, Czuppon P, Bornberg-Bauer E. Quantification and modeling of turnover dynamics of de novo transcripts in Drosophila melanogaster. Nucleic Acids Res. 2024:52(1):274–287. https://doi.org/10.1093/nar/gkad1079.

Gubala AM, Schmitz JF, Kearns MJ, Vinh TT, Bornberg-Bauer E, Wolfner MF, Findlay GD. The goddard and saturn genes are essential for Drosophila male fertility and may have arisen de novo. Mol Biol Evol. 2017:34(5):1066–1082. https://doi.org/10.1093/molbev/msx057.

Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol. 2018:19(10):621–637. https://doi.org/10.1038/s41580-018-0028-8.

Huang W, Carbone MA, Magwire MM, Peiffer JA, Lyman RF, Stone EA, Anholt RR, Mackay TF. Genetic basis of transcriptome diversity in Drosophila melanogaster. Proc Natl Acad Sci USA. 2015:112(44):E6010–E6019. https://doi.org/10.1073/pnas.1519159112.

Iyengar BR, Bornberg-Bauer E. Neutral models of de novo gene emergence suggest that gene evolution has a preferred trajectory. Mol Biol Evol. 2023:40(4):msad079. https://doi.org/10.1093/molbev/msad079.

Kapusta A, Feschotte C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. Trends Genet. 2014:30(10):439–452. https://doi.org/10.1016/j.tig.2014.08.004.

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. 2013:9(4):e1003470. https://doi.org/10.1371/journal.pgen.1003470.

Kassambara A, Kassambara M. Package "ggpubr". R package Version 0.3. 5.2020.

Kim T-K, Shiekhattar R. Architectural and functional commonalities between enhancers and promoters. Cell. 2015:162(5):948–959. https://doi.org/10.1016/j.cell.2015.08.008.

Kofler R, Nolte V, Schlötterer C. Tempo and mode of transposable element activity in Drosophila. PLoS Genet. 2015:11(7):e1005406. https://doi.org/10.1371/journal.pgen.1005406.

Kordyukova M, Olovnikov I, Kalmykova A. Transposon control mechanisms in telomere biology. Curr Opin Genet Dev. 2018:49:56–62. https://doi.org/10.1016/j.gde.2018.03.002.

Kurafeiski JD, Pinto P, Bornberg-Bauer E. Evolutionary potential of cis-regulatory mutations to cause rapid changes in transcription factor binding. Genome Biol Evol. 2019:11(2):406–414. https://doi.org/10.1093/gbe/evy269.

Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010:11(3):204–220. https://doi.org/10.1038/nrg2719.

Lawlor MA, Cao W, Ellison CE. A transposon expression burst accompanies the activation of y-chromosome fertility genes during Drosophila spermatogenesis. Nat Commun. 2021:12(1):6854. https://doi.org/10.1038/s41467-021-27136-4.

Lawson HA, Liang Y, Wang T. Transposable elements in mammalian chromatin organization. Nat Rev Genet. 2023:24(10):712–723. https://doi.org/10.1038/s41576-023-00609-6.

Lis M, Walther D. The orientation of transcription factor binding site motifs in gene promoter regions: does it matter? BMC Genomics. 2016:17(S1):1–21. https://doi.org/10.1186/s12864-016-2549-x.

Lyko F, Ramsahoye BH, Jaenisch R. Dna methylation in Drosophila melanogaster. Nature. 2000:408(6812):538–540. https://doi.org/10.1038/35046205.

Magnusson A, Skaug H, Nielsen A, Berg C, Kristensen K, Maechler M, van Bentham K, Bolker B, Brooks M, Brooks MM. Package "glmmtmb". R Package Version 0.2. 0, 25.2017.

Majic P, Payne JL. Enhancers facilitate the birth of de novo genes and gene integration into regulatory networks. Mol Biol Evol. 2020:37(4):1165–1178. https://doi.org/10.1093/molbev/msz300.

Malik HS. Retroviruses push the envelope for mammalian placentation. Proc Natl Acad Sci USA. 2012:109(7):2184–2185. https://doi.org/10.1073/pnas.1121365109.

Mateo L, Ullastres A, González J. A transposable element insertion confers xenobiotic resistance in Drosophila. PLoS Genet. 2014:10(8):e1004560. https://doi.org/10.1371/journal.pgen.1004560.

McCullers TJ, Steiniger M. Transposable elements in Drosophila. Mob Genet Elements. 2017:7(3):1–18. https://doi.org/10.1080/2159256X.2017.1318201.

McLysaght A, Hurst LD. Open questions in the study of de novo genes: what, how and why. Nat Rev Genet. 2016:17(9):567–578. https://doi.org/10.1038/nrg.2016.78.

Meers MP, Adelman K, Duronio RJ, Strahl BD, McKay DJ, Matera AG. Transcription start site profiling uncovers divergent transcription and enhancer-associated RNAs in Drosophila melanogaster. BMC Genomics. 2018:19(1):1–20. https://doi.org/10.1186/s12864-018-4510-7.

Mérel V, Boulesteix M, Fablet M, Vieira C. Transposable elements in Drosophila. Mob DNA. 2020:11(1):1–20. https://doi.org/10.1186/s13100-020-00213-z.

Moschetti R, Palazzo A, Lorusso P, Viggiano L, Massimiliano Marsano R. "what you need, baby, i got it": transposable elements as suppliers of cis-operating sequences in Drosophila. Biology. 2020:9(2):25. https://doi.org/10.3390/biology9020025.

Neme R, Tautz D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. Elife. 2016:5:e09977. https://doi.org/10.7554/eLife.09977.

Palazzo AF, Lee ES. Non-coding RNA: what is functional and what is junk? Front Genet. 2015:6:2. https://doi.org/10.3389/fgene.2015.00002.

Peng FY, Hu Z, Yang R-C. Bioinformatic prediction of transcription factor binding sites at promoter regions of genes for photoperiod and vernalization responses in model and temperate cereal plants. BMC Genomics. 2016:17:1–16. https://doi.org/10.1186/s12864-016-2916-7.

Petrov DA, Fiston-Lavier A-S, Lipatov M, Lenkov K, González J. Population genomics of transposable elements in Drosophila melanogaster. Mol Biol Evol. 2011:28(5):1633–1644. https://doi.org/10.1093/molbev/msq337.

Quinlan AR, Hall IM. Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010:26(6):841–842. https://doi.org/10.1093/bioinformatics/btq033.

Reineke AR, Bornberg-Bauer E, Gu J. Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes. Nucleic Acids Res. 2011:39(14):6029–6043. https://doi.org/10.1093/nar/gkr179.

Riehl K, Riccio C, Miska EA, Hemberg M. Transposonultimate: software for transposon classification, annotation and detection. Nucleic Acids Res. 2022:50(11):e64–e64. https://doi.org/10.1093/nar/gkac136.

Rizzon C, Marais G, Gouy M, Biémont C. Recombination rate and the distribution of transposable elements in the Drosophila melanogaster genome. Genome Res. 2002:12(3):400–407. https://doi.org/10.1101/gr.210802.

Rödelsperger C, Prabh N, Sommer RJ. New gene origin and deep taxon phylogenomics: opportunities and challenges. Trends Genet. 2019:35(12):914–922. https://doi.org/10.1016/j.tig.2019.08.007.

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

Schmitz JF, Chain FJ, Bornberg-Bauer E. Evolution of novel genes in three-spined stickleback populations. Heredity. 2020:125(1-2):50–59. https://doi.org/10.1038/s41437-020-0319-7.

Schmitz JF, Ullrich KK, Bornberg-Bauer E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. Nat Ecol Evol. 2018:2(10):1626–1632. https://doi.org/10.1038/s41559-018-0639-7.

Scruggs BS, Gilchrist DA, Nechaev S, Muse GW, Burkholder A, Fargo DC, Adelman K. Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. Mol Cell. 2015:58(6):1101–1112. https://doi.org/10.1016/j.molcel.2015.04.006.

Sloutskin A, Shir-Shapira H, Freiman RN, Juven-Gershon T. The core promoter is a regulatory hub for developmental gene expression. Front Cell Dev Biol. 2021:9:666508. https://doi.org/10.3389/fcell.2021.666508.

Small S, Arnosti DN. Transcriptional enhancers in Drosophila. Genetics. 2020:216(1):1–26. https://doi.org/10.1534/genetics.120.301370.

Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nat Rev Genet. 2011:12(10):692–702. https://doi.org/10.1038/nrg3053.

Team RDC. A language and environment for statistical computing. http://www.R-project.org.2022.

Thybert D, Roller M, Navarro FC, Fiddes I, Streeter I, Feig C, Martin-Galvez D, Kolmogorov M, Janoušek V, Akanni W, et al. Repeat associated mechanisms of genome evolution and function revealed by the Mus caroli and Mus pahari genomes. Genome Res. 2018:28(4):448–459. https://doi.org/10.1101/gr.234096.117.

Van Oss SB, Carvunis A-R. De novo gene birth. PLoS Genet. 2019:15(5):e1008160. https://doi.org/10.1371/journal.pgen.1008160.

Wang Y-W, Hess J, Slot JC, Pringle A. De novo gene birth, horizontal gene transfer, and gene duplication as sources of new gene families associated with the origin of symbiosis in amanita. Genome Biol Evol. 2020b:12(11):2168–2182. https://doi.org/10.1093/gbe/evaa193.

Wang M, Wang D, Zhang K, Ngo V, Fan S, Wang W. Motto: representing motifs in consensus sequences with minimum information loss. Genetics. 2020a:216(2):353–358. https://doi.org/10.1534/genetics.120.303597.

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, et al. Welcome to the tidyverse. J Open Source Softw. 2019:4(43):1686. https://doi.org/10.21105/joss.

Wickham H, Chang W, Wickham MH. Package 'ggplot2': Create Elegant Data Visual Grammar Graph. Ver. 2016:2(1):1–189.

Wickham H, François R, Henry L, Müller K. 2022. Rstudio. (2021). dplyr: A grammar of data manipulation (1.0. 7).

Wingett SW, Andrews S. Fastq screen: a tool for multi-genome mapping and quality control. F1000Res. 2018:7:1338. https://doi.org/10.12688/f1000research.

Wolf T, Shelest V, Nath N, Shelest E. Cassis and smips: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. Bioinformatics. 2016:32(8):1138–1143. https://doi.org/10.1093/bioinformatics/btv713.

Ylla G, Nakamura T, Itoh T, Kajitani R, Toyoda A, Tomonari S, Bando T, Ishimaru Y, Watanabe T, Fuketa M, et al. Insights into the genomic evolution of insects from cricket genomes. Commun Biol. 2021:4(1):733. https://doi.org/10.1038/s42003-021-02197-9.

Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of de novo genes in Drosophila melanogaster populations. Science. 2014:343(6172):769–772. https://doi.org/10.1126/science.1248286.

Zhuang X, Cheng C-HC. Propagation of a de novo gene under natural selection: antifreeze glycoprotein genes and their evolutionary history in codfishes. Genes. 2021:12(11):1777. https://doi.org/10.3390/genes12111777.

Associate editor: Victor Luria