A quantitative, genome-wide analysis in *Drosophila* reveals transposable elements' influence on gene expression is species-specific.

Marie Fablet[1,2], Judit Salcez-Ortiz[1,6], Angelo Jacquet[1,7], Bianca F. Menezes[1,8], Corentin Dechaud[3], Philippe Veber[1], Camille Noûs[4], Rita Rebollo[5], Cristina Vieira[1]

[1] Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon; Université Lyon 1; CNRS; UMR 5558, Villeurbanne, France.

[2] Institut Universitaire de France (IUF)

[3] Institut de Génomique Fonctionnelle de Lyon, Univ Lyon, CNRS UMR 5242, Ecole Normale Supérieure de Lyon, Université Claude Bernard Lyon 1, 46 allée d'Italie, F-69364 Lyon, France

[4] Laboratoire Cogitamus

[5] Univ Lyon, INRAE, INSA-Lyon, BF2I, UMR 203, 69621 Villeurbanne, France.

[6] Current address: Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra) P° Marítimo de la Barceloneta, 37-49 08003 Barcelona, Spain

[7] Current address: Symbiotron; FR3728 Biodiversité, Eau, Environnement, Ville, Santé; Université Claude Bernard Lyon 1; Villeurbanne 69622, France

[8] Current address: Federal Institute of Rio de Janeiro (IFRJ), Pinheiral – RJ, Brazil

Corresponding authors: marie.fablet@univ-lyon1.fr; cristina.vieira@univ-lyon1.fr

Running title: TEs' impact on gene expression is species-specific

Abstract

Transposable elements (TEs) are parasite DNA sequences that are able to move and multiply along the chromosomes of all genomes. They are controlled by the host through the targeting of silencing epigenetic marks, which may affect the chromatin structure of neighboring sequences, including genes. In this study, we used transcriptomic and epigenomic high-throughput data produced from ovarian samples of *Drosophila melanogaster* and *Drosophila simulans* wild-type strains, in order to finely quantify the influence of TE insertions on gene RNA levels and histone marks (H3K4me3 and H3K9me3) enrichments. We find general, negative impacts of TE insertions on gene RNA levels and H3K4me3 enrichments, which intensity varies according to the localization of the TE insertion. We also uncover that TE insertions within exons are associated with a reduction in H3K9me3 enrichments, which we propose is associated with TE splicing out. Overall, we find that gene RNA and histone mark levels show contrasted dependencies on TE insertions regarding the species. The fold-decrease in TE-carrying gene proportions between extreme expression classes is 2.6 in *D. melanogaster* and 1.8 in *D. simulans*, while it is strongly contrasted between extreme classes of H3K4me3 enrichments: 4.4 in *D. melanogaster* and 19.2 in *D. simulans.* This provide a new light on the considerable natural variability provided by TEs, which may be associated with contrasted adaptive and evolutionary potentials.

Introduction

Transposable elements (TEs) are parasite DNA sequences that are able to move and multiply along the chromosomes of all genomes (Wells and Feschotte 2020). As such, they are source of mutations and even genome instability if uncontrolled (Biémont and Vieira 2006; Malone and Hannon 2009; Senti and Brennecke 2010). Such control generally consists in the targeting of silencing epigenetic marks at TE copies, which may spread to neighboring sequences and impact gene expression. In this regard, few attempts were made to finely analyze and quantify TEs' influence at the whole genome scale (Cridland et al. 2015; Hollister and Gaut 2009; Lee and Karpen 2017; Uzunović et al. 2019). However, TE contents, activities and

control pathways vary a lot in nature (Akkouche et al. 2012, 2013; Baduel et al. 2021; Mérel et al. 2021, 2020). The present study therefore sheds new light on this question integrating natural variability at the intra- and inter-specific levels in a *Drosophila* system. Specifically, we assessed the transcriptomic and epigenetic status of genes in the ovaries of wild-type strains of *D. melanogaster* and *D. simulans* in relation with TE contents.

Previous studies focusing on a few TE families in wild-type strains of *D. simulans* and *D. melanogaster* uncovered between-species differences in histone mark landscapes (Rebollo et al. 2012). Lee and Karpen (Lee and Karpen 2017) previously provided an analysis on the repressive histone mark H3K9me2 (Histone 3 Lysine 9 dimethylation) around TEs from two Drosophila Genetic Reference Panel (DGRP) strains (*D. melanogaster*), and concluded to pervasive epigenetic effects of TEs. Here, we focused on H3K9me3 (Histone 3 Lysine 9 trimethylation, repressive) and H3K4me3 (Histone 3 Lysine 4 trimethylation, permissive) marks, which proved to be particularly relevant in TE epigenetic dynamics (Ozata et al. 2019), and we produced genome-wide data from four wild-type strains of *D. melanogaster* and four wild-type strains of *D. simulans*, which allows us to provide a quantitative assessment of the contribution of TEs to gene expression, in a comparative genomics perspective (Fig. 1).

Previous research already described the effects of TE insertions on gene expression using collections of strains of *D. melanogaster* (Cridland et al. 2015; Everett et al. 2020; Osada et al. 2017; Zhang et al. 2020). The novelties of ours is 1) to bring together TE insertion polymorphism, RNA-seq, ChIP-seq on two histone marks, and small RNA-seq data on the exact same strains, 2) to focus on ovaries, which is the organ where strong selective pressures act on TE control (and not on whole bodies or mix of tissues), 3) to provide a new, quantitative approach based on linear models, and 4) to reach a larger evolutionary scale thanks to the inclusion of data not only from *D. melanogaster* but also from its sister species *D. simulans*. We investigated the impacts of TEs on gene expression at the whole genome scale using eight previously characterized, wild-type strains of *D. melanogaster* and *D. simulans* (Mohamed et al. 2020). These strains derived from samples collected in France and Brazil, two strains per location and per species. Using the Oxford Nanopore long read sequencing technology, we previously produced high quality genome assemblies at the chromosome resolution for each of the strains (Mohamed et al. 2020).

Henceforth, we reveal that TE insertions within genes (exons and introns) have stronger negative impacts on gene RNA levels and H3K4me3 enrichments compared to TE insertions in the gene flanking regions, and that the proportion of genes carrying TE insertions decreases when RNA levels and H3K4me3 increase, and increases with H3K9me3 enrichments. We also uncover a surprinsing result regarding TE insertions within exons: they are associated with a reduction in H3K9me3 enrichments. Considering that H3K9me3 is involved in the definition of splice sites, we speculate that TE insertions within exons are not purged by natural selection provided they are skipped from the corresponding transcripts, which may be achieved through reductions in H3K9me3 marks. Regarding orthologous genes across strains, the more genes display different RNA and histone mark levels, the more likely they have different numbers of TE insertions (especially regarding RNA levels in *D. melanogaster*, and H3K4me3 enrichments in *D. simulans*). These results participate in the accurate, quantitative understanding of TEs' impacts in genomes. Moreover, our major finding in this research work is that, across genes, RNA and histone mark levels show contrasted dependencies on TE insertions regarding the species, *D. melanogaster* or *D. simulans*. This provide a new light on the considerable natural variability provided by TEs, which may be associated with contrasted adaptive and evolutionary potentials, all the more sensible in a rapidly changing environment (Baduel et al. 2021; Fablet and Vieira 2011; Mérel et al. 2021).

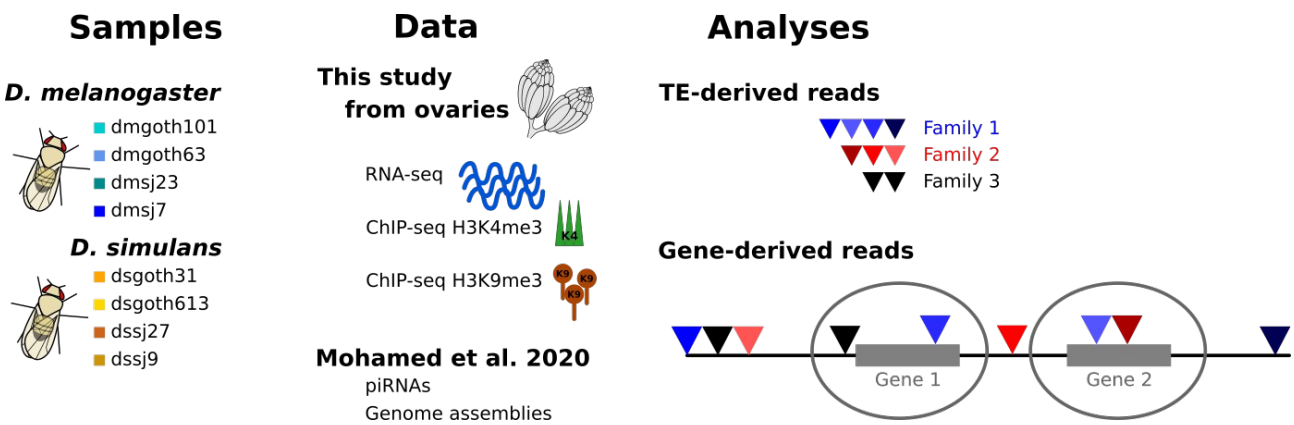


Figure 1. Graphic summary of the study.

Eight wild-type strains from *D. melanogaster* and *D. simulans* were included in the study. The present datasets are RNA-seq and ChIP-seq for H3K4me3 and H3K9me3 marks, and were prepared from ovarian

samples. They were analyzed in parallel with already published data produced from the same drosophila strains: ovarian piRNA repertoires and genome assemblies based on Oxford Nanopore long read sequencing (Mohamed et al. 2020). For RNA-seq and ChIP-seq, TE-derived reads were analyzed at the TE family level, and gene-derived reads were analyzed in relation to TE insertions inside or nearby the genes (therefore restricted to the TE insertions included within the gray bubbles).

Results

TE expression and epigenetic targeting in Drosophila ovaries

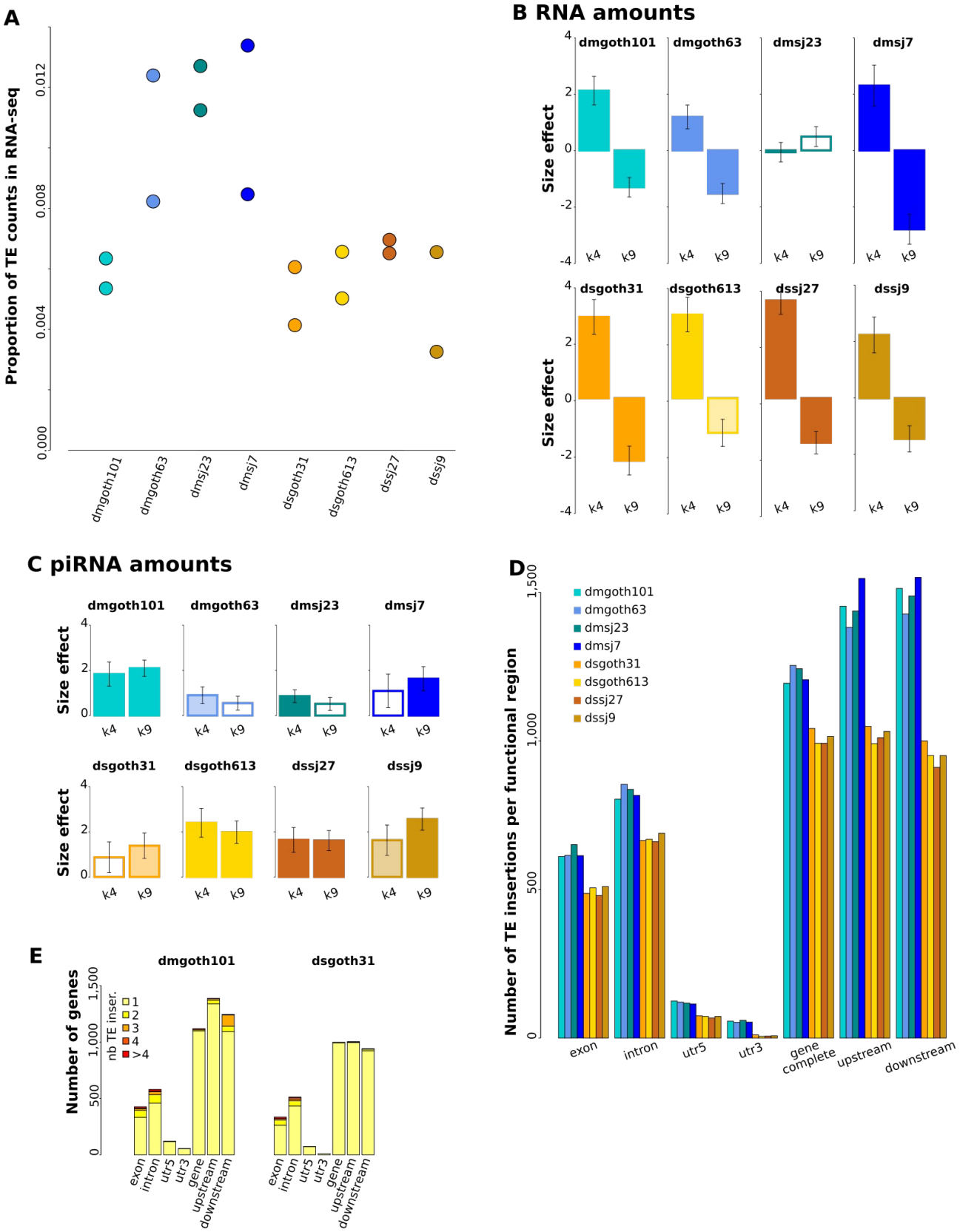
We first considered TE-derived RNA-seq reads from all our sequence files, which we analyzed at the TE-family level (Fig. 1). TEs account for 0.6% to 1.2%, and 0.5% to 0.7%, of read counts corresponding to annotated sequences (genes and TEs) within the ovarian transcriptomes of our strains of *D. melanogaster* and *D. simulans*, respectively (Fig. 2A). We find significant positive correlations between per TE family RNA counts and family sequence occupancy (total number of bp spanned by each TE family along the genome) (Spearman correlations, rho = 0.20 to 0.31, and 0.28 to 0.35, in *D. melanogaster* and *D. simulans*, respectively; Supplemental_Fig_S2A). We also find significant positive correlations between per TE family RNA counts and piRNA counts (Spearman correlations, rho = 0.65 to 0.70, and 0.64 to 0.68, in *D. melanogaster* and *D. simulans*, respectively; Supplemental_Fig_S2B), as already described for wild-type strains (Lerat et al. 2017).

We assessed the contribution of histone mark enrichments to TE RNA amounts considering the following linear model on log-transformed normalized read counts: RNA ~ H3K4me3 + H3K9me3 + input. These models led to adjusted r^2 as high as 0.48 to 0.64, and 0.45 to 0.60, in *D. melanogaster* and *D. simulans* strains, respectively, suggesting that this approach captures a significant portion of TE RNA amount variation. We find that TE RNA amounts are positively correlated with H3K4me3 and negatively correlated with H3K9me3 amounts (Fig. 2B), as expected considering that H3K4me3 and H3K9me3 are activating and silencing marks, respectively. We used a similar approach to analyze piRNA amounts, and considered the following linear model on log-transformed read counts: piRNA ~ H3K4me3 + H3K9me3 + input. We

obtained even higher adjusted r^2 values, from 0.71 to 0.75, and 0.64 to 0.68, in *D. melanogaster* and *D. simulans* strains, respectively. We find that TE-derived piRNA amounts are positively correlated both with permissive H3K4me3 and repressive H3K9me9 enrichments (Fig. 2C).

TE insertions within or nearby genes

In the following sections, we focus on gene-derived reads from all our sequence files, which we analyzed in relation with the presence of TE insertions within or nearby genes. Based on gene annotations, we distinguished the different functional regions of genes, *i.e.* exons, introns, 5' UTRs, 3' UTRs, upstream, or downstream sequences (5 kb flanking regions). In the present study, we considered a set of 17,417 annotated genes for *D. melanogaster*, and 15,251 for *D. simulans* (see Material and Methods). We quantified the number of TE insertions within genes (Fig. 2D), and found that they only account for ~10% of the total number of TE insertions per genome. The majority of genes (93%) are devoid of TE insertions, and 6% have exactly one TE insertion (Fig. 2E, Supplemental_Fig_S1). 92% to 94% of genes are devoid of TE insertions in their upstream flanking regions (over 5 kb), and 6% to 8% display exactly one insertion. The proportions are the same regarding the downstream flanking regions.



138

139 **Figure 2.** (*A*) Proportions of TE read counts in RNA-seq data over read counts corresponding to genes and

140 TEs. For each strain, two biological replicates are shown. (*B*) Contributions of H3K4me3 and H3K9me3

enrichments to TE-derived RNA read counts (according to the model RNA ~ H3K4me3 + H3K9me3 + input calculated on log transformed read count numbers, at the TE family level). Plain-colored bars: p-values < 0.01, light-colored bars: p-values in [0.01; 0.05], empty bars: p-values > 0.05. (*C*) Contributions of H3K4me3 and H3K9me3 enrichments to TE-derived piRNA read counts (according to the model piRNA ~ H3K4me3 + H3K9me3 + input calculated on log transformed read count numbers, at the TE family level). Plain-colored bars: p-values < 0.01, light-colored bars: p-values in [0.01; 0.05], empty bars: p-values > 0.05. (*D*) Numbers of TE insertions per functional region per strain. Upstream and downstream regions are 5 kb sequences directly flanking transcription units 5' and 3', respectively. (*E*) Numbers of genes for each value of TE insertion numbers. dmgoth101 and dsgoth31 are shown as examples; all strains can be found in Supplemental_Fig_S1.

TEs contribute to RNA level variability across genes within genomes

One of the objectives of the present work was to quantify the association of TE insertions with the differences in expression levels that we observe across the set of genes of a genome. We focused on the subset of genes that we found expressed in the ovaries (see Material and Methods). They account for 8,028 to 8,209 genes depending on the strains, and 6,149 to 6,431 genes, in *D. melanogaster* and *D. simulans*, respectively. We first quantified TE insertion contribution to gene RNA levels using the following linear models built on log-transformed TPM (Transcript Per Million): TPM ~ gene + upstream + downstream, where these variables correspond to the numbers of TE insertions in gene body, 5 kb upstream, and 5 kb downstream, respectively. We found that the strongest effect on gene RNA amounts is a negative correlation with the number of TE insertions in gene bodies. TE insertions that are located upstream of genes are not correlated to gene transcript amounts on average. On the contrary, TE insertions that are located downstream of genes are associated with increased gene transcript amounts, however only in *D. simulans* strains (Fig. 3A).

We investigated the impacts of TE insertions within genes into more details using the following model: log-transformed TPM ~ exon + intron + utr5 + utr3 + upstream + downstream, which allowed us to refine the

correlations between expression and each category of functional region. We found that TE insertions in all internal regions display a general tendency to reduce gene transcript amounts (Supplemental_Fig_S3A).

Then, considering each genome individually, we defined 10 TPM classes, and calculated the proportion of genes with at least one TE insertion in each class. This procedure allowed us to bring out clear negative correlations between the average TPM and the proportion of such genes (Fig. 3B). This indicates that the more the genes are transcribed, the lower the probability that they carry TE insertions. Depending on the strain, we observe a 2.2 to 2.9-fold decrease in the proportions of such genes between the two extreme expression classes in *D. melanogaster*, and 1.5 to 2.2 in *D. simulans*. Using the model including more detailed characterization of functional regions (exon, intron, 5' UTR, 3' UTR, upstream, downstream), we found similar negative correlations in *D. melanogaster*, especially for exons and introns, and also, but more mildly, for 5' UTRs and 3' UTRs. These trends are much less pronounced in *D. simulans*. In addition, in both species, the expression classes do not shown differences in the proportions of TE-carrying genes within the 5 kb upstream or downstream flanking regions (Supplemental_Fig_S4).

TEs contribute to histone enrichment variability across genes within genomes

We used a similar approach to analyze H3K4me3 and H3K9me3 enrichments (*i.e.* we aligned ChIP-seq reads against whole gene sequences and computed corresponding read counts). Similar to what we found for RNA levels, TE insertions within genes are significantly negatively correlated with H3K4me3 amounts (Fig. 3C). On the contrary, TE insertions in the flanking downstream regions have a mild, positive impact, although detected in *D. melanogaster* only. We could not detect impacts from insertions in upstream regions.

When we detailed the functional regions within genes, we detected a robust, negative impact of TE insertions within introns on gene H3K4me3 levels in *D. melanogaster* (Supplemental_Fig_S3B). Similarly, we find mild, negative impacts of TE insertions within 5' UTRs.

Considering H3K9me3 amounts, the impacts of TE insertions are more significant and more widespread across strains. Using the model H3K9me3 ~ input + gene + upstream + downstream, we detect significant,

positive contributions of TE insertions in all regions (gene, upstream, downstream) (Fig. 3E). These results are in agreement with TEs being the preferential targets for H3K9me3 deposition, which then spreads to neighboring regions. Using the detailed model H3K9me3 ~ input + exon + intron + utr5 + utr3 + upstream + downstream, we find negative contributions of TE insertions within introns and flanking regions (both upstream and downstream). Unexpectedly however, TE insertions within exons have a clear, negative impact on gene H3K9me3 amounts (Supplemental_Fig_S3C).

Then, similar to the approach we used for RNA level analyses, we classified genes according to their histone mark enrichment levels, and computed the proportions of TE-carrying genes for each of these classes (Fig. 2D,F). We find significant, negative correlations between these proportions and H3K4me3 enrichment levels, which are stronger for *D. simulans* strains compared to *D. melanogaster*. The reduction in the proportion of TE-carrying genes is 3.1 to 5.7-fold, and 17.1 to 20.5-fold between the extreme histone classes, in *D. melanogaster* and *D. simulans*, respectively. When considering detailed functional regions, we found that this effect holds on for exons, introns, and 5' UTRs (Supplemental_Fig_S5). No significant effects are detected for 3' UTRs, and flanking regions.

Regarding H3K9me3, we found significant, strong positive correlations between the proportion of TE-carrying genes and histone levels in *D. simulans* genomes: the proportion increases 8.3 to 11.7-fold between the extreme histone classes. Surprisingly, correlations are weaker in *D. melanogaster* genomes. However, we detected strong, positive correlations between the proportions of genes carrying TE insertions within flanking regions (both upstream and downstream) and H3K9me3 amounts, consistently across all eight genomes.

Private and shared TE insertions are associated with similar patterns

It has already been described that TE impacts differ across young (*i.e.* polymorphic) and old (*i.e.* fixed) TE copies; indeed, the pool of old TE insertions has been purged from deleterious insertions by natural selection. Indeed, Uzunovic *et al.* (Uzunović et al. 2019) showed in the plant *Capsella* that young TE insertions had a

negative effect on gene expression while old insertions were more likely to increase expression. In this view, we performed the same analyses as above separately on TE insertions that are unique to one genome (private) and those that are shared by all four strains of the species (common). The majority of the TE insertions (>80%) that we included in the analysis in the first place fall in this last category, leading to differences in statistical power between both sets of insertions. Despite this bias, the observed patterns are highly similar (Supplemental_Table_S1). Since our approach is gene-centered, it is very likely that our primary set of TE insertions is already biased: when deleterious, insertions within or near genes have such a negative impact that we are not able to catch them from natural samples. Therefore, our primary set of TE insertions may already correspond to copies that have passed the filter of natural selection.

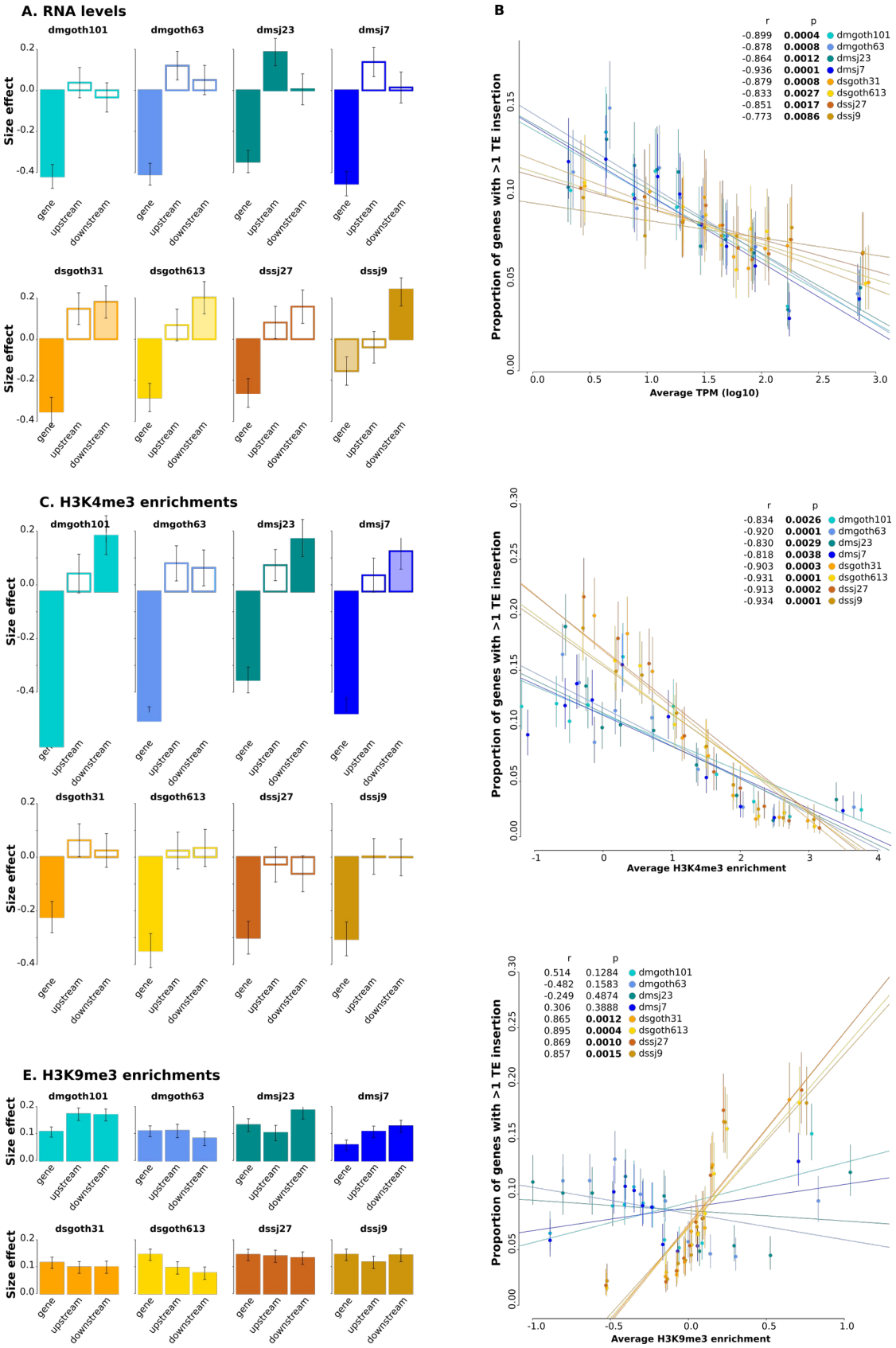


Figure 3. (*A*) Contribution of TE insertion numbers (within genes or in the 5 kb upstream or downstream sequences) to gene-derived RNA read counts. Plain-colored bars: p-values < 0.01, light-colored bars: p-

values in [0.01; 0.05], empty bars: p-values > 0.05. (*B*) Proportions of genes carrying at least one TE insertion according to gene expression classes. We plotted confidence intervals for the proportion of TE-carrying genes and we fitted a linear regression line with mean TPM per class (Pearson correlation coefficients and the corresponding p-values are indicated). (*C*) Contribution of TE insertion numbers (within genes or in the 5 kb upstream or downstream sequences) to gene H3K4me3 enrichments. Plain-colored bars: p-values < 0.01, light-colored bars: p-values in [0.01; 0.05], empty bars: p-values > 0.05. (*D*) Proportions of genes carrying at least one TE insertion according to gene H3K4me3 enrichment classes. We plotted confidence intervals for the proportion of TE-carrying genes and we fitted a linear regression line with mean histone enrichment per class (Pearson correlation coefficients and the corresponding p-values are indicated). (*E*) Contribution of TE insertion numbers (within genes or in the 5 kb upstream or downstream sequences) to gene H3K9me3 enrichments. Plain-colored bars: p-values < 0.01, light-colored bars: p-values in [0.01; 0.05], empty bars: p-values > 0.05. (*F*) Proportions of genes carrying at least one TE insertion according to gene H3K9me3 enrichment classes. We plotted confidence intervals for the proportion of TE-carrying genes and we fitted a linear regression line with mean histone enrichment per class (Pearson correlation coefficients and the corresponding p-values are indicated).

TE insertions are associated with orthologous gene expression variability across genomes

Next, we used our experimental dataset to infer the contribution of TE insertions at the inter-genomic level, *i.e.* we compared expression levels of orthologous genes across genomes. We performed pairwise comparisons and calculated the coefficient of variation (CV, square root of the variance divided by the mean) of TPM for genes that are transcribed in our ovarian samples. We defined 10 classes based on CV values, and computed the proportion of genes that displayed different numbers of TE insertions within each class. The general trend is positive correlation between these proportions and the mean CVs of each class, and the adjustment is generally better using the square root of mean CV. This indicates that the more the genes show differences in expression levels, the greater the probability that they have different numbers of TE insertions

(Fig. 4A, Supplemental_Fig_S7, Supplemental_Fig_S8). The effect appears to be stronger in *D. melanogaster* compared to *D. simulans*.

We applied the same approach to ChIP-seq data: for each gene and each pair of strains, we considered the model H3K*i*me3 (*i* stands for 4 or 9) ~ input + strain, and retrieved the absolute value of the effect of the strain factor. We defined 10 classes of genes based on the values of this variable (square root transformed), and, within each class, we computed the proportion of genes that carried different numbers of TE insertions across strains. Considering H3K4me3, results are not significant for *D. melanogaster*, but there are clear positive correlations between these proportions and the difference in histone mark enrichment for *D. simulans* (Fig. 4B). In other terms, the more the genes show differences in H3K4me3 enrichments, the greater the probability that they carry different numbers of TE insertions. This trend is essentially pronounced for TE insertions located in introns: we observed a 2.0 to 4.0-fold increase in the proportions of genes with different numbers of TE insertions, depending on the considered pairs of strains (Supplemental_Fig_S9, Supplemental_Fig_S10).

Considering H3K9me3, we find the same trends, *i.e.* positive correlations between the proportions of genes that display different numbers of TE insertions and differences in H3K9me3 enrichments, and essentially due to TE insertions in introns (Fig. 4C, Supplemental_Fig_S11, Supplemental_Fig_S12). However, the majority of these correlations are not significant at the 0.05 threshold.

Even if we observe substantial variability across pairwise comparisons, altogether, these results indicate that differences in TE insertion numbers between strains lead to clear differences in RNA levels, and weak differences in histone mark enrichments, at least for those that we tested here.

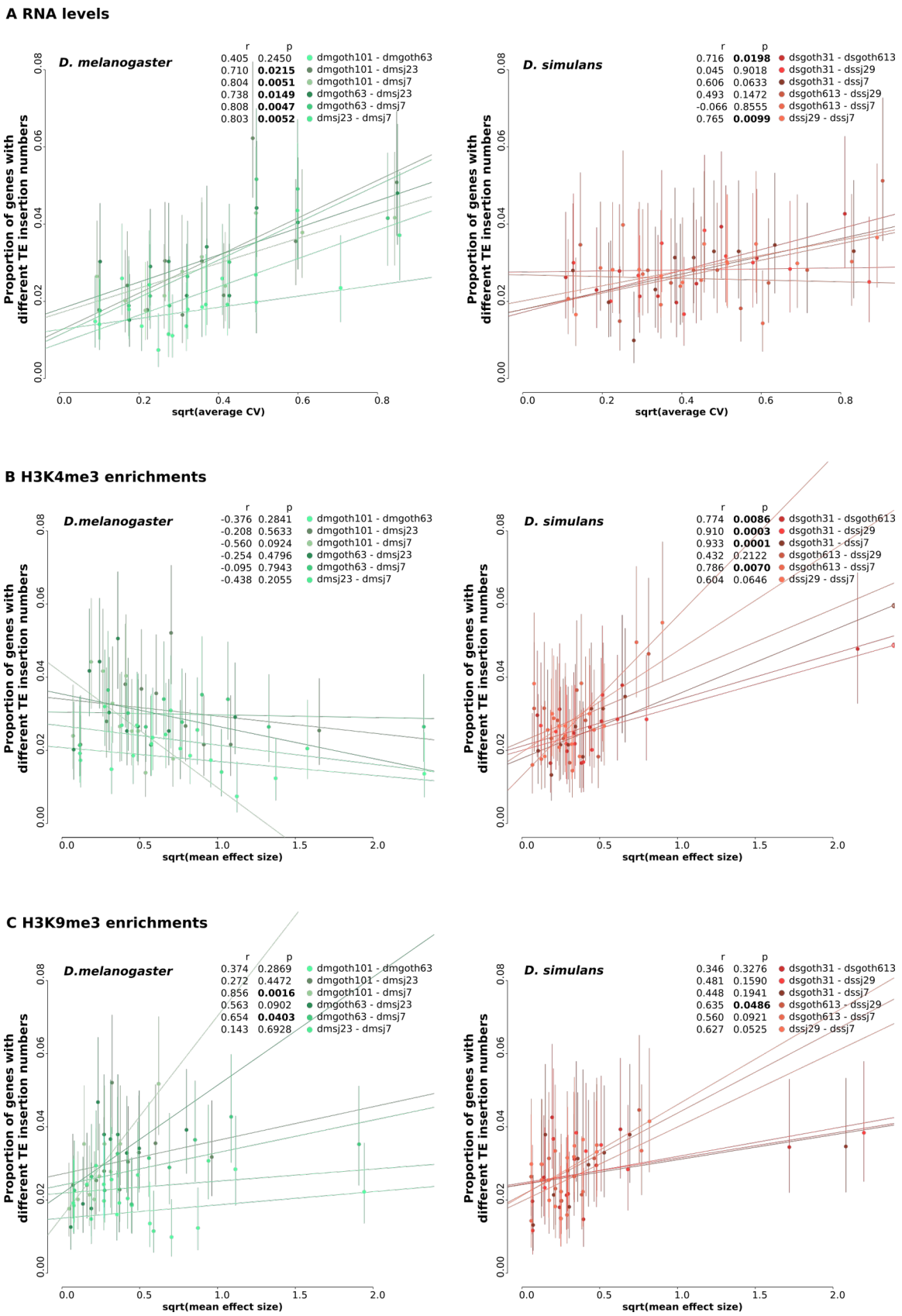


Figure 4. Pairwise comparisons across strains of TE insertion numbers and RNA levels or histone mark enrichments. For each pair of strains, proportions of orthologous genes carrying different numbers of

TE insertions according to (*A*) differences in gene RNA levels (computed based on variation coefficient), (*B*) size of the strain effect for H3K4me3 enrichments, (*C*) and H3K9me3 enrichments. We plotted confidence intervals for the proportion of genes carrying different numbers of TEs and we fitted a linear regression line with square root of mean CV per class (A) or square root of mean histone enrichment effect per class (B, C) (Pearson correlation coefficients and the corresponding p-values are indicated).

Discussion

The common-held view is that, as parasites that are fought against by the genomes, TEs have a general negative impact on gene expression (Cridland et al., 2015; Lee, 2015; Lee & Karpen, 2017). Our present findings are in agreement with this idea. However, the originality of this research work is to provide an unprecedented quantitative view, which allows to precisely decipher TE impacts, integrating intra- and inter-species data from two closely related drosophila species. In addition, this study combines genomic, transcriptomic, and epigenetic high-throughput sequencing data, all produced from the same strains and tissues from natural strains, which allows to draw actual connections between the different nucleic repertoires of the cells. Moreover, the chosen tissue is ovary, which is known to be the place of the strongest epigenetic control of TE sequences, through the piRNA pathway (Malone and Hannon 2009; Senti and Brennecke 2010). This is therefore where we are to expect the strongest impacts of TEs on genes.

Expression and epigenetic marks of TE sequences

Our results uncovered a lower contribution of TEs to the *D. simulans* transcriptome as compared to *D. melanogaster*. This is in agreement with the lowest contribution of TEs in the genomes of *D. simulans*, in terms of sequence occupancy and copy numbers (Mohamed et al. 2020; Vieira et al. 1999). In both species, we found that H3K9me3 marks on TE sequences were associated with a decrease in TE-derived RNA amounts, and the opposite for H3K4me3 marks. On the contrary, we observed that both histone marks are positively correlated with TE-derived piRNA amounts, which is congruent with the piRNA-targeted deposition of H3K9me3 marks at transcriptionaly active TE copies (Czech et al. 2018).

TE impact on gene expression and histone marks : intragenomic approach

As expected due to the spreading of TE epigenetic marks to flanking sequences, our analyses revealed that TE insertions within genes induce a general decrease in gene RNA levels and H3K4me3 enrichments, and an increase in H3K9me3 enrichments. While we found that TE insertions in flanking regions equally induce H3K9me3 enrichments, we found that the impacts of TE insertions on RNA levels and H3K4me3 enrichments are stronger for TE insertions within genes compared to TE insertions in flanking regions. This had already been observed by Cridland *et al.* (Cridland et al. 2015), who found a stronger correlation with extreme gene expression for TEs inserted into genes compared to TEs upstream of genes. This may be due to differences in deposition patterns for these marks, which result in difference in spreading abilities. Indeed, contrary to H3K9me3, H3K4me3 marks are known to appear as peaks. In addition, we observe that TE insertions in flanking regions are associated with increases in gene RNA levels and H3K4me3 enrichments. This may be related to TEs bringing transcription factor binding sites or behaving as enhancers, as was already described (Sundaram et al. 2014; Villanueva-Cañas et al. 2019; Moschetti et al. 2020).

While the detailed analyses of gene functional compartments (exon, intron, 5' UTR, 3' UTR) generally revealed the same patterns as the complete gene sequence, we were surprised to uncover a negative influence of TE insertions in exons on H3K9me3 enrichments. This finding may be explained by the influence of H3K9me3 on the splicing machinery: it was shown in mammals that H3K9me3 favors the inclusion of alternative exons while demethylation is associated with exon skipping (Bieberstein et al. 2016; Saint-André et al. 2011). We propose that the TE insertions in exons that are maintained and not purged by natural selection, are those that are spliced out from transcripts, so that the coding sequence of the gene is not impaired. This specific splicing would be possible for TE insertions that are associated with a reduction in H3K9 methylation, which allows the splicing out of the TE.

TE impacts on genes are different between *D. melanogaster* and *D. simulans*

Although less abundant in the genome, the epigenetic impacts of TEs are more pronounced in *D. simulans*. Indeed, we found that the probability that a gene carries at least one TE insertion is more strongly correlated with its histone enrichment in *D. simulans* compared to *D. melanogaster* (Fig. 5). Similarly, although using a different, more indirect genome-wide approach, Lee and Karpen (Lee and Karpen 2017) also came to the conclusion that, at the same time, *D. simulans* displays a lower TE content and stronger epigenetic effects of TE control, compared to *D. melanogaster*. Again, it sustains the idea that *D. simulans* epigenetic machinery is less efficient since a strong pattern in H3K9me3 and/or H3K4me3 results in a weak pattern at the RNA levels, or in other words, since weak histone patterns in *D. melanogaster* are enough to produce a strong RNA response. Alternatively, upstream of the regulatory effectors, it may suggest that the targeting of TE sequences for histone mark deposition is more efficient in *D. simulans*, maybe because the colonization of the *D. simulans* genome by TEs started more recently (as suggested by our previous results (Mohamed et al. 2020)). We expect that the strength of the epigenetic control of TEs should decrease after the colonization, as the equilibrium between the genome and its TEs is reached. These contrasted impacts of TE insertions on genes through epigenetic marks across the species provide an additional demonstration of the considerable natural variability due to TEs. We predict that this leads to contrasted adaptive and evolutionary potentials, all the more sensible in a rapidly changing environment (Baduel et al. 2021; Fablet and Vieira 2011; Mérel et al. 2021).

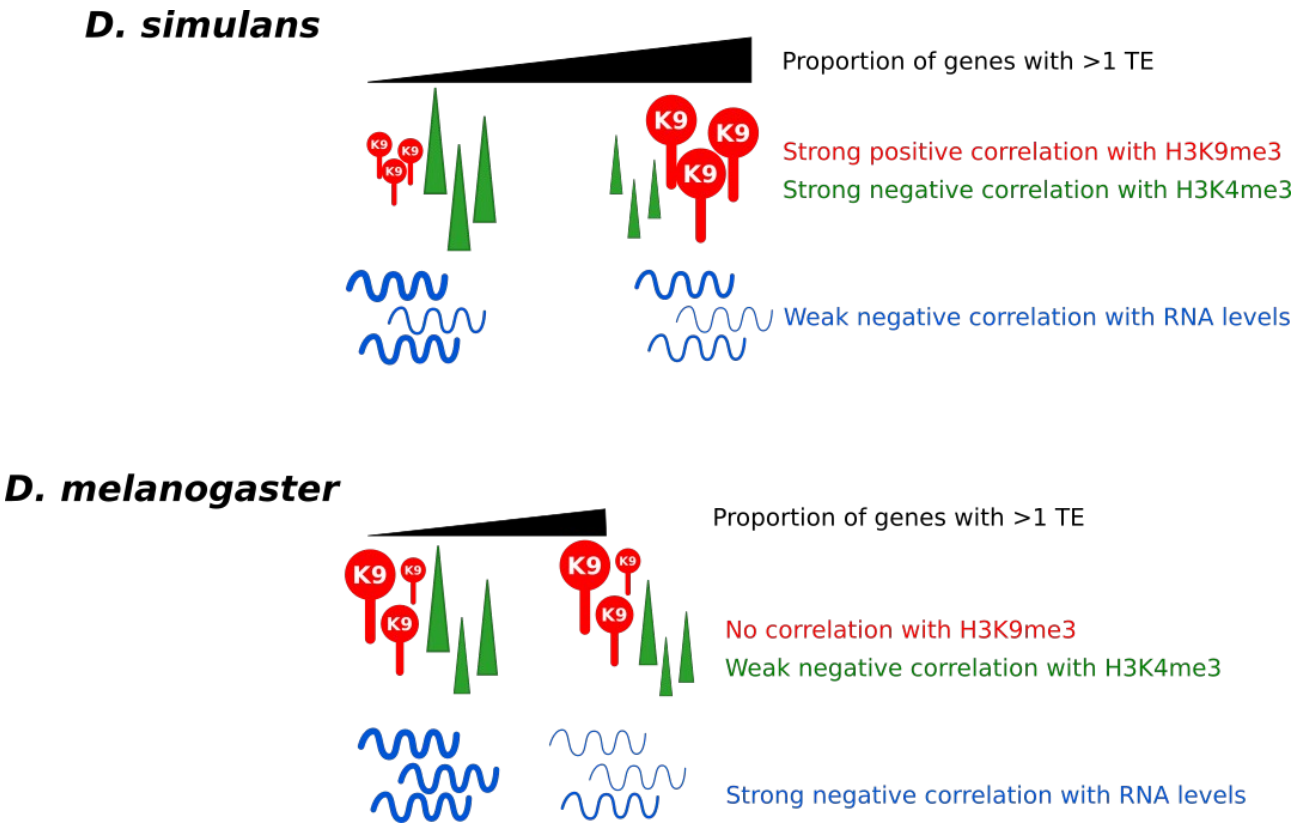


Figure 5. Graphic summary. The probability that a gene carries at least one TE insertion is more strongly correlated with its histone enrichment in *D. simulans* compared to *D. melanogaster*. On the contrary, this probability is more strongly correlated with gene RNA levels in *D. melanogaster*.

Methods

Drosophila strains

The strains under study in the present work were previously described in Mohamed et al. (Mohamed et al. 2020). The eight samples of *D. melanogaster* and *D. simulans* wild-type strains were collected using fruit baits in France (Gotheron, 44°56'0"N 04°53'30"E - "goth" strains) and Brazil (Saõ Jose do Rio Preto 20°41'04.3"S 49°21'26.1"W – "sj" strains) in June 2014. Two isofemale lines per species and geographical origin were established directly from gravid females from the field (French *D. melanogaster*: dmgoth63, dmgoth101; Brazilian *D. melanogaster*: dmsj23, dmsj7; French *D. simulans*: dsgoth613, dsgoth31; Brazilian *D. simulans*: dssj27, dssj9). Brothers and sisters were then mated for 30 generations to obtain inbred strains with very low intra-line genetic variability. Strains were kept at 24°C in standard laboratory conditions on cornmeal–sugar–yeast–agar medium.

Genome annotation

Genome assemblies were produced in (Mohamed et al. 2020) and have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB50024 (https://www.ebi.ac.uk/ena/browser/view/PRJEB50024).

We used the genome assemblies produced for each of the eight strains using the Oxford Nanopore sequencing technology, and previously published (Mohamed et al. 2020). Throughout the present analysis, we kept scaffolds corresponding to complete chromosomes 2L, 2R, 3L, 3R, 4, and X.

TE annotation: We used RepeatMasker 4.1.0 (http://repeatmasker.org/) -species Drosophila in order to identify TE sequences in the assemblies, followed by OneCodeToFindThemAll (Bailly-Bechet et al. 2014) with default parameters, in order to parse RepeatMasker results.

Gene annotation: For *D. melanogaster*, we retrieved dmel-all-exon-r6.16.fasta, dmel-all-intron-r6.16.fasta, dmel-all-five_prime_UTR-r6.16.fasta, dmel-all-five_prime_UTR-r6.16.fasta, and dmel-all-gene-r6.16.fasta, which are available at ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.16_FB2017_03/fasta/. For *D. simulans*, we retrieved dsim-all-exon-r2.02.fasta, dsim-all-intron-r2.02.fasta, dsim-all-five_prime_UTR-

r2.02.fasta, dsim-all-three_prime_UTR-r2.02.fasta, and dsim-all-gene-r2.02.fasta, which are available at ftp.flybase.net/genomes/Drosophila_simulans/dsim_r2.02_FB2017_04/fasta/. These different sequence sets were aligned to our assemblies using BLAT (Kent 2002) and pslReps from the BLAT suite, in order to keep the best hit (-singleHit). We filtered out genes with 0 hit and genes with multiple hits. For each gene, we also checked that the positions of the hits for each of the internal regions (*i.e.* exon, intron, 5' UTR, and 3' UTR) were included in the positions of the complete gene hit. We excluded regions that did not fulfill these criteria. We also excluded hits that were not on the same chromosome as the corresponding gene on the reference genome. Then, we used the GenomicRanges R package (version 1.38.0) (Lawrence et al. 2013) and the subsetByOverlaps function to cross gene and TE annotations.

TE genomic sequence occupancy (bp) was computed using OneCodeToFindThemAll (Bailly-Bechet et al. 2014). Data corresponding to the *DNAREP1* transposon (also known as *INE-1*) were removed because it is a highly abundant element displaying mainly fixed insertions in the *melanogaster* complex of species (Chakraborty et al. 2021).

In order to determine which TE insertions were common (shared) to the four strains of a species or unique (private) to one strain, we performed pairwise comparisons of TE gff using Liftoff (Shumate and Salzberg 2020). Private insertions to one strain are those that appear in the unmapped output of the three other strains. Common insertions are those that are found in all pairwise comparisons.

Small RNA extraction, sequencing, and analyses

Small RNA extraction, sequencing, and analyses had already been performed and described in Mohamed et al. (Mohamed et al. 2020). Sequence files had been deposited in NCBI SRA under the accession number PRJNA644327.

RNA-seq preparation

RNA was extracted from ovaries of 30 3-to-5 day-old females. Two replicates per strain were produced. RNA extraction was carried out using RNeasy Plus (Qiagen) kit following manufacturer's instructions. After DNAse treatment (Ambion), quality control was performed using an Agilent Bioanalyzer. Libraries were

constructed from mRNA using the Illumina TruSeq RNA Sample Prep Kit following manufacturer's recommendations. Libraries were sequenced on Illumina HiSeq 3000 with paired-end 150 nt reads.

RNA-seq analysis

TE read counts were computed at the family level using the TEcount module of TEtools (Lerat et al. 2017) and the list of TE sequences available at ftp://pbil.univ-lyon1.fr/pub/datasets/Roy2019/. In agreement with other research works (Chakraborty et al. 2021), we removed the *DNAREP1* family from the analyses because it is a highly abundant element displaying mainly fixed insertions in the *melanogaster* complex of species (it is also known as *INE-1*, and corresponding to DNAREP1_DM, DNAREP1_DSim, and DNAREP2_DSim entries from our TE sequence list).

Genome and gene annotations: The genomes of *D. melanogaster* and *D. simulans* were downloaded from flybase (http://flybase.org/) along with GTF gene annotations. *D. melanogaster*: *f*tp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.16_FB2017_03/fasta/dmel-all-chromosome-r6.16.fasta.gz; ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.16_FB2017_03/gtf/dmel-all-r6.16.gtf.gz. *D. simulans*: ftp://ftp.flybase.net/releases/FB2017_03/dsim_r2.02/fasta/dsim-all-chromosome-r2.02.fasta.gz; ftp://ftp.flybase.net/releases/FB2017_03/dsim_r2.02/gtf/dsim-all-r2.02.gtf.gz.

Reads mapping: The alignment of all read samples on their corresponding reference genome was performed with Hisat2 version 2.1.0 (Kim et al. 2019) using the following command line options: hisat2-build $Genome $GenomeIndex to build an index of the genome; hisat2 -p 7 -k 2 --dta -x $GenomeIndex -1 $Sample_1.fq.gz -2 $Sample_2.fq.gz -S $Sample.sam to align each sample on the reference genome. The --dta option is used as recommended in the New tuxedo pipeline to perform an assembly of the transcripts (Pertea et al. 2016). The -k 2 option allows to report a maximum of two alignments per sequencing read. As we wanted to exclude potentially expressed TEs from the assembly, we discarded multimapped reads using the following command line: samtools view -h -@ 7 -f 0x2 $Sample.bam |awk '{if(substr($1, 0, 1)=="@" || $5==60){print $0}}' >$Sample.sam. The option -f 0x2 extracts the properly aligned paired reads from the bam. The awk command also keeps the header section of the SAM file as well as the reads with a MAPQ=60. In Hisat2, a MAPQ=60 means that the read has been uniquely aligned regardless of the

mismatches / indels number. The filtered SAM file is finally converted to BAM and sorted: samtools view -h -@ 7 -b -S $Sample.sam >$Sample_filtered.bam && samtools sort -@ 7 $Sample_filtered.bam sorted.bam && mv sorted.bam $Sample_filtered.bam.

Transcripts assembly: We used StringTie (Pertea et al. 2015) to assemble the transcripts using the genomic coordinates of the aligned reads, for each sample: stringtie -p 7 -o $Sample_assemb.gtf -j 2 $Sample_filtered.bam && echo "$Sample$_assemb.gtf" >> mergelist.txt. The -j option specifies that at least two reads are needed to support an exon-exon junction and create a new transcript. This is thus more stringent than the default value of 1 used by StringTie. One assembly per sample is created; we used the following command line to merge these assemblies: stringtie --merge -o merged.gtf mergelist.txt.

Transcript expression quantification: Stringtie was also used to quantify transcript expression in each sample: stringtie -e -B -p 7 -G merged.gtf -o $Sample/$Sample.gtf $Sample_filtered.bam. -e option is used to estimate the expression of transcripts of the GTF file given to the -G option. -B option is used to create an output needed to run ballgown (Frazee et al. 2015) in the next step.

Count matrix generation: We used the prepDE.py (https://github.com/gpertea/stringtie/blob/master/prepDE.py) script from StringTie (Pertea et al. 2015) to generate the count matrix from the output of the previous step.

Gene expression level quantification: We used the ballgown R package (Frazee et al. 2015) to estimate the TPM (Transcript Per Million) expression of each gene or transcript. Data are loaded with ballgown from the output of StringTie -e -B. Transcripts with low expression are filtered out as recommended in the new tuxedo procedure (Pertea et al. 2016) using: subset(data,"rowMeans(texpr(data))>0.5", genomesubset=TRUE). FPKM were then calculated using texpr function from ballgown. TPM where finally computed from this table by normalizing the sum of each sample to 1e6.

Comparison to gene annotation: Genes and transcripts reconstructed by StringTie were compared to the published gene annotation of the genome using gffcompare (https://ccb.jhu.edu/software/stringtie/gffcompare.shtml). Each new transcript was assigned a class code depending on its similarity to a reference transcript, as described in gffcompare documentation.

Quantification of the effects of TE insertions on gene transcript levels: considering only genes expressed in ovaries, we computed mean TPMs across replicates and used the following linear models after log transformation: TPM ~ gene + upstream + downstream, where “gene”, “upstream”, and “downstream” are the numbers of TE insertions in gene body, 5 kb upstream sequences, and 5 kb downstream sequences, respectively. Size effects for each of these factors were then recorded. In order to go into more details regarding the locations of TE insertions, we also used TPM ~ exon + intron + utr5 + utr3 + upstream + downstream. In parallel, for each strain, we defined 10 classes of genes based on mean TPM values, and computed the proportion of genes carrying at least one TE insertion in each of the 10 classes.

For the intergenomic analysis, we considered all pairs of strains within species, and for each pair of strains and each gene, we computed the difference in TE insertion numbers on one hand and the coefficient of variation (CV, square root of variance divided by mean) for TPM on the other hand. We then defined 10 classes of genes based on CV values and, for each of these classes, we computed the proportions of genes carrying different numbers of TEs.

ChIP-seq preparation

Chromatin immunoprecipitation was performed using 50 ovary pairs dissected from three to five day old females. Ovaries were re-suspended in A1 buffer containing 60mM KCl, 15mM NaCl, 15mM Hepes, 0,5% Triton and 10mM Sodium butyrate. Formaldehyde (Sigma) was added to a final concentration of 1.8% for secondary cross-linking for 10 min at room temperature. Formaldehyde was quenched using glycine (0.125 M). Cross-linked cells were washed and pelleted twice with buffer A1, once with cell lysis buffer (140mM NaCl, 15mM Hepes, 1mM EDTA, 0.5mM EGTA, 1% Triton X100, 0.1% Sodium deoxycholate, 10mM sodium butyrate) , followed by lysis in buffer containing ,140mM NaCl, 15mM Hepes, 1mM EDTA, 0.5mMEGTA, 1% Triton X100, 0.5% SDS, 0.5% N-Laurosylsarcosine, 0.1% sodium deoxycholate, 10mM sodium butyrate for 120 min at 4ºC. Lysates were sonicated in Bioruptor sonicator to reach a fragment size window of 200-600 bp.

Chromatin was incubated overnight at 4ºC with the following antibodies: for H3K9me3 ChIP using α-H3K9me3 (actif motif #39161, 3µg/IP) and for H3K4me3 using α-H3K4me3 (millipore #07-473, 3µg/IP)

antibodies. The Magna ChIP™ A/G Chromatin Immunoprecipitation Kit (cat# 17-10085) was used following manufacturer's instructions. Final DNA recovery was perfomed by classic phenol/chloroform DNA precipitation method using MAxtract high density tubes to maximize DNA recovery.

DNA fragments were then sequenced on an Illumina HiSeq 4000 apparatus, with paired-end 100 nt reads. Due to technical issues, only one replicate could be used for dsgoth31 input.

ChIP-seq quality check: Validation of H3K4me3 enrichment around promoters.

Raw reads were trimmed using trim_galore (https://zenodo.org/record/5127899#.YbnMs73MLDc) with default parameters along with --paired, --clip_R1 9, --clip_R2 9, and --max_n 0. Mapping was performed using Bowtie2 ((Langmead and Salzberg 2012)) with --sensitive-local against the *D. melanogaster* r6.16 and *D. simulans* r2.02 genomes. Samtools was used to convert SAM to coordinated sorted BAM files, while sambamba ((Tarasov et al. 2015)) was used to filter for uniquely mapping reads and to remove duplicates (sambamba view -h -t 2 -f bam -F "[XS] == null and not unmapped and not duplicate"). For *D. melanogaster* datasets, we filtered available blacklisted regions ((Amemiya et al. 2019)) with bedtools. Finally coverage files containing reads per genome coverage (RPGC) were obtained with DeepTools (Ramírez et al. 2016) bamCoverage with --extendReads, --effectiveGenomeSize 129789873 for *D. melanogaster* available from the Deeptools suite, and --effectiveGenomeSize 121102921 computed with unique-kmers.py from khmer (https://github.com/dib-lab/khmer). Promoter regions were obtained with gencode_regions (https://github.com/saketkc/gencode_regions) and along with coverage files were used in DeepTools computeMatrix and plotProfile to build the average coverage of H3K4me3 and H3K9me3 around transcription start sites in both species. The corresponding profiles looked as expected (Supplemental_Fig_S13).

ChIP-seq analysis

For each of the immunoprecipitated samples (H3K4me3, H3K9me3, input), TE read counts were computed at the family level using the TEcount module of TEtools (Lerat et al., 2017) and the list of TE sequences available at ftp://pbil.univ-lyon1.fr/pub/datasets/Roy2019/.

ChIP-seq counts were normalized across samples of the same species using the counts(normalize=T) function of DESeq2 1.26.0 (Love et al. 2014). This was done independently for each of the immunoprecipitated samples (H3K4me3, H3K9me3, input). We then performed a log-transformation using the rlogTransformation function of DESeq2, and subsequently considered mean values across replicates. We only kept genes expressed in ovaries.

In order to quantify the effects of TE insertions on histone marks enrichments, we used the following linear models on log transformed read counts: histone mark (either H3K4me3 or H3K4me3) ~ input + gene + upstream + downstream, where “gene”, “upstream”, and “downstream” are the numbers of TE insertions in gene body, 5 kb upstream sequences, and 5 kb downstream sequences, respectively. Size effects for each of these three factors were then recorded. In order to go into more details regarding the locations of TE insertions, we also used histone mark (either H3K4me3 or H3K4me3) ~ input + exon + intron + utr5 + utr3 + upstream + downstream.

In parallel, for each strain, we defined 10 classes of genes based on histone mark enrichment, built upon the values of {log transformed histone mark read counts (either H3K4me3 or H3K9me3) – log transformed input read counts}, calculated for each gene. We computed the proportion of genes carrying at least one TE insertion in each of the 10 classes.

For the intergenomic analysis, for each pair of strains and each ovary-expressed gene, we considered the model H3Kime3 (i stands for 4 or 9) ~ input + strain, and retrieved the absolute value of the effect of the strain factor. We defined 10 classes of genes based on the values of this variable (square root transformed), and computed the proportion of genes that carried different numbers of TE insertions within each class.

Data access

The RNA-seq data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA795668. The ChIP-seq data generated in this study have been submitted to the NCBI BioProject database under accession number PRJNA796157. TE and gene annotations have been deposited to Zenodo doi: 10.5281/zenodo.5844405.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Francois Sabot, Matthieu Boulesteix, and Vincent Mérel for useful discussions and technical help. We thank Gladys Mialdea, Justine Picarle, Sonia Janillon, and Nelly Burlet for technical help. This work was performed using the computing facilities of the CC LBBE/PRABI. Sequencing was performed by the GenomEast platform, a member of the “France Génomique”consortium (ANR-10-INBS-0009). This work was supported by Fondation pour la Recherche Médicale (grant DEP20131128536) and Agence Nationale de la Recherche (grant ExHyb ANR-14-CE19-0016-01).

554 # References

Akkouche A, Grentzinger T, Fablet M, Armenise C, Burlet N, Braman V, Chambeyron S, Vieira C. 2013. Maternally deposited germline piRNAs silence the tirant retrotransposon in somatic cells. *EMBO Rep* **14**: 458–464.

Akkouche A, Rebollo R, Burlet N, Esnault C, Martinez S, Viginier B, Terzian C, Vieira C, Fablet M. 2012. tirant, a newly discovered active endogenous retrovirus in Drosophila simulans. *J Virol* **86**: 3675–3681.

Amemiya HM, Kundaje A, Boyle AP. 2019. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**. https://pubmed-ncbi-nlm-nih-gov.inee.bib.cnrs.fr/31249361/ (Accessed January 4, 2022).

Baduel P, Leduque B, Ignace A, Gy I, Gil JJ, Loudet O, Colot V, Quadrana L. 2021. Genetic and environmental modulation of transposition shapes the evolutionary potential of Arabidopsis thaliana. *Genome Biol* **22**. https://pubmed-ncbi-nlm-nih-gov.inee.bib.cnrs.fr/33957946/ (Accessed January 4, 2022).

Bailly-Bechet M, Haudry A, Lerat E. 2014. “One code to find them all”: a perl tool to conveniently parse RepeatMasker output files. *Mob DNA* **5**: 13.

Bieberstein NI, Kozáková E, Huranová M, Thakur PK, Krchňáková Z, Krausová M, Carrillo Oesterreich F, Staněk D. 2016. TALE-directed local modulation of H3K9 methylation shapes exon recognition. *Sci Rep* **6**: 29961.

Biémont C, Vieira C. 2006. Genetics: junk DNA as an evolutionary force. *Nature* **443**: 521–524.

Chakraborty M, Chang C-H, Khost DE, Vedanayagam J, Adrion JR, Liao Y, Montooth KL, Meiklejohn CD, Larracuente AM, Emerson JJ. 2021. Evolution of genome structure in the Drosophila simulans species complex. *Genome Res* **31**: 380–396.

Cridland JM, Thornton KR, Long AD. 2015. Gene expression variation in Drosophila melanogaster due to rare transposable element insertion alleles of large effect. *Genetics* **199**: 85–93.

Czech B, Munafò M, Ciabrelli F, Eastwood EL, Fabry MH, Kneuss E, Hannon GJ. 2018. piRNA-Guided Genome Defense: From Biogenesis to Silencing. *Annu Rev Genet* **52**: 131–157.

Everett LJ, Huang W, Zhou S, Carbone MA, Lyman RF, Arya GH, Geisz MS, Ma J, Morgante F, St Armour G, et al. 2020. Gene expression networks in the Drosophila Genetic Reference Panel. *Genome Res* **30**: 485–496.

Fablet M, Vieira C. 2011. Evolvability, epigenetics and transposable elements. *BioMol Concepts* **2**: 333–341.

Frazee AC, Pertea G, Jaffe AE, Langmead B, Salzberg SL, Leek JT. 2015. Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat Biotechnol* **33**: 243–246.

Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res* **19**: 1419–1428.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656–664.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**: 907–915.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9**: e1003118.

Lee YCG. 2015. The Role of piRNA-Mediated Epigenetic Silencing in the Population Dynamics of Transposable Elements in Drosophila melanogaster. *PLoS Genet* **11**: e1005269.

Lee YCG, Karpen GH. 2017. Pervasive epigenetic effects of Drosophila euchromatic transposable elements impact their evolution. *eLife* **6**.

Lerat E, Fablet M, Modolo L, Lopez-Maestre H, Vieira C. 2017. TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids Res* **45**: e17.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Malone CD, Hannon GJ. 2009. Small RNAs as guardians of the genome. *Cell* **136**: 656–668.

Mérel V, Boulesteix M, Fablet M, Vieira C. 2020. Transposable elements in Drosophila. *Mob DNA* **11**: 23.

Mérel V, Gibert P, Buch I, Rodriguez Rada V, Estoup A, Gautier M, Fablet M, Boulesteix M, Vieira C. 2021. The Worldwide Invasion of Drosophila suzukii Is Accompanied by a Large Increase of Transposable Element Load and a Small Number of Putatively Adaptive Insertions. *Mol Biol Evol* **38**. https://pubmed-ncbi-nlm-nih-gov.inee.bib.cnrs.fr/34021759/ (Accessed January 4, 2022).

Mohamed M, Dang NT-M, Ogyama Y, Burlet N, Mugat B, Boulesteix M, Mérel V, Veber P, Salces-Ortiz J, Severac D, et al. 2020. A Transposon Story: From TE Content to TE Dynamic Invasion of Drosophila Genomes Using the Single-Molecule Sequencing Technology from Oxford Nanopore. *Cells* **9**.

Moschetti R, Palazzo A, Lorusso P, Viggiano L, Marsano RM. 2020. “What You Need, Baby, I Got It”: Transposable Elements as Suppliers of Cis-Operating Sequences in Drosophila. *Biology* **9**: E25.

Osada N, Miyagi R, Takahashi A. 2017. Cis- and Trans-regulatory Effects on Gene Expression in a Natural Population of Drosophila melanogaster. *Genetics* **206**: 2139–2148.

Ozata DM, Gainetdinov I, Zoch A, O’Carroll D, Zamore PD. 2019. PIWI-interacting RNAs: small RNAs with big functions. *Nat Rev Genet* **20**: 89–108.

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* **11**: 1650–1667.

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

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**: W160–W165.

Rebollo R, Horard B, Begeot F, Delattre M, Gilson E, Vieira C. 2012. A snapshot of histone modifications within transposable elements in Drosophila wild type strains. *PloS One* **7**: e44253.

Saint-André V, Batsché E, Rachez C, Muchardt C. 2011. Histone H3 lysine 9 trimethylation and HP1γ favor inclusion of alternative exons. *Nat Struct Mol Biol* **18**: 337–344.

Senti K-A, Brennecke J. 2010. The piRNA pathway: a fly's perspective on the guardian of the genome. *Trends Genet TIG* **26**: 499–509.

Shumate A, Salzberg SL. 2020. Liftoff: accurate mapping of gene annotations. *Bioinforma Oxf Engl* btaa1016.

Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. 2014. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res* **24**: 1963–1976.

Tarasov A, Villeva AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: fast processing of NGS alignment formats. *Bioinforma Oxf Engl* **31**. https://pubmed-ncbi-nlm-nih-gov.inee.bib.cnrs.fr/25697820/ (Accessed January 4, 2022).

Uzunović J, Josephs EB, Stinchcombe JR, Wright SI. 2019. Transposable Elements Are Important Contributors to Standing Variation in Gene Expression in Capsella Grandiflora. *Mol Biol Evol* **36**: 1734–1745.

Vieira C, Lepetit D, Dumont S, Biémont C. 1999. Wake up of transposable elements following Drosophila simulans worldwide colonization. *Mol Biol Evol* **16**: 1251–1255.

Villanueva-Cañas JL, Horvath V, Aguilera L, González J. 2019. Diverse families of transposable elements affect the transcriptional regulation of stress-response genes in Drosophila melanogaster. *Nucleic Acids Res* **47**: 6842–6857.

Wells JN, Feschotte C. 2020. A Field Guide to Eukaryotic Transposable Elements. *Annu Rev Genet*. https://www-annualreviews-org.inee.bib.cnrs.fr/doi/10.1146/annurev-genet-040620-022145 (Accessed November 4, 2020).

Zhang X, Zhao M, McCarty DR, Lisch D. 2020. Transposable elements employ distinct integration strategies with respect to transcriptional landscapes in eukaryotic genomes. *Nucleic Acids Res* **48**: 6685–6698.

555