ChimeraTE: A pipeline to detect chimeric transcripts derived from genes and transposable elements

Daniel S. Oliveira[1,2]; Marie Fablet[2,3]; Anaïs Larue[2]; Agnès Vallier[4], Claudia M. A. Carareto[1]; Rita Rebollo[4,*]; Cristina Vieira[2,*]

[1] Institute of Biosciences, Humanities and Exact Sciences, São Paulo State University (Unesp), São José do Rio Preto, São Paulo, 15054-000, Brazil.
[2] Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, CNRS, UMR5558, Villeurbanne, Rhone-Alpes, 69100, France.
[3] Institut Universitaire de France (IUF), Paris, Île-de-France , F-75231, France.
[4] Univ Lyon, INRAE, INSA-Lyon, BF2I, UMR 203, 69621 Villeurbanne, France.

* To whom correspondence should be addressed. Tel: +33 (0)4 72 44 82 16; Email: cristina.vieira@univ-lyon1.fr; and Tel: +33 (0)4 72 43 62 46; Email: rita.rebollo@inrae.fr

ABSTRACT Transposable elements (TEs) are structural variants considered an important source of genetic diversity, which may arise in the transcriptome when TEs are transcribed in the same RNA molecule as genes, producing what we hereafter call chimeric transcripts. The presence of chimeric transcripts has been associated with adaptive traits in several species, but their identification remains hindered due to the lack of tools to detect them on a transcriptome-wide scale. Previous bioinformatics tools were developed to identify chimeric transcripts derived from TEs present in a reference genome. Nevertheless, different individuals/cells/strains might harbor different TE insertions generating such chimeric transcripts. Therefore, we have developed ChimeraTE, a pipeline that uses paired-end RNA-seq reads to identify chimeric transcripts with or without a reference genome, in a transcriptome-wide manner. ChimeraTE has two Modes: Mode 1 is a genome-guided approach that employs the canonical method of genome alignment, whereas Mode 2 identifies chimeric transcripts without a reference genome, being able to predict chimeras derived from fixed or polymorphic TEs. We have used both Modes with Illumina RNA-seq reads from ovarian tissues of *Drosophila melanogaster* wild-type strains, and found that ~3% of all genes generate chimeric transcripts. Approximately ~9% of all detected chimeras were absent from the *D. melanogaster*'s reference genome, corresponding to polymorphic insertions in the wild-type strains. ChimeraTE is the first pipeline with the ability to automatically uncover chimeric transcripts without a reference genome.

INTRODUCTION

Transposable elements (TEs) are mobile DNA sequences that comprise a large fraction of eukaryotic genomes, from 15% in *Drosophila melanogaster* (1), 45% in humans (2), and 85% in maize (3). Many TE copies have lost their ability to transpose as a result of accumulated mutations and recombination throughout evolution (4). Despite their lack of mobility, such ancient TE insertions may still harbor functional protein domains, alternative splice sites, and *cis*-acting regulatory sequences, as transcription factor binding sites (TFBSs) and polyadenylation (PolyA) sites. Therefore, TEs are a major source of genetic diversity, not only due to their mobilization, but also because they donate protein domains to gene functions (5–8) and regulatory sequences that modify the expression of nearby genes (9–13). When one of these processes is integrated into the biology of the host, this evolutionary process is called exaptation (14).

Chimeric transcripts are RNAs stemming from two sequences from different origins (15). Hereafter we assume chimeric transcripts as mature transcripts that have both gene and TE-derived sequences. These transcripts can be divided into three types: (1) TE-initiated transcripts: chimeric transcripts with a TE transcription start site (TSS) (16, 17); (2) TE-exonized transcripts: TE sequences are incorporated into the transcript either partially or as full-length exons (18–20); and (3) TE-terminated transcripts: chimeric transcript with a TE transcription termination site (21, 22). TE-initiated and TE-terminated transcripts might modulate gene expression levels either by the presence of TFBSs, PolyA sites, or chromatin changes; while TE-exonized transcripts may alter the protein sequence of coding genes and have a direct effect on the protein function. Regardless of the TE position, such events of TE exaptation and domestication have been associated with many biological roles and are widespread among eukaryotic species (23). In *D. melanogaster*, the *CHKov1* gene generates a chimeric transcript with a truncated mRNA resulting in resistance to insecticide and viral infection (24). In humans, the *SETMAR* gene produces a chimeric transcript containing a *Hsmar1* copy, involved in non-homologous end-joining DNA repair (25). In cancer, TEs become active due to a global hypomethylation state (26) and such activation may generate new chimeric transcripts with detrimental outcomes (27), a process called onco-exaptation (9). For example, in large B-cell lymphoma, the *FABP7* gene has an endogenous retrovirus LTR co-opted as a promoter, generating a novel protein involved in abnormal cell proliferation (28). Therefore, chimeric transcripts have a large impact on host biology, but their study remains hindered by the ubiquitous repetitive nature of TE copies.

Previous studies with Cap Analysis Gene Expression (CAGE) revealed a significant percentage of genes producing TE-initiated transcripts, ranging from 3-14% in humans and mice, depending upon the tissue (29). More specifically, in human pluripotent stem cells, chimeric transcripts comprise 26% of coding and 65% of noncoding transcripts (30). In *D. melanogaster*, a study with expressed sequence tags (ESTs) has shown that the proportion of genes with chimeric transcripts is reduced to ~1%, slowing down chimeric transcript searches in the *Drosophila* species (31). More recently, a tissue-specific study has shown that 264 genes produce chimeric transcripts in the midbrain transcriptome of *D. melanogaster*, corresponding to ~1.5% of all genes (32). Several bioinformatics methods have been developed to take advantage of RNA sequencing (RNA-seq) to identify chimeric transcripts, as CLIFinder (33) and LIONS (34). The former is designed to identify chimeric transcripts derived from

LINE in the human genome, whereas the latter is able to identify only TE-initiated transcripts. Both methods need a reference genome and they only detect chimeric transcripts derived from TE insertions present in the reference. Therefore, it is not possible to identify chimeric transcripts derived from polymorphic TE insertions that may exist in other populations, strains, or individuals. Finally, the latest addition to chimeric transcript detection, TEchim (32), is able to detect chimeric transcripts without a genome annotation in *D. melanogaster*, but it is not a pipeline designed to run automatically with any other genome.

In this study, we have developed ChimeraTE, a pipeline that uses paired-end RNA-seq reads to identify chimeric transcripts. The pipeline has two Modes: Mode 1 can predict chimeric transcripts through genome alignment, whereas Mode 2 performs chimeric transcript searches without a reference genome, being able to identify chimeras derived from fixed or polymorphic TE insertions. In order to benchmark the pipeline, we have used RNA-seq from ovaries of four *D. melanogaster* wild-type strains, for which we have assembled and annotated genomes. We found that ~3% of genes have chimeric transcripts in the ovarian transcriptome, of which 56.23% are TE-exonized transcripts. Our results also reveal that *roo* is the most frequent exonized TE family. In addition, with Mode 2, we found 11 polymorphic chimeric transcripts deriving from TE insertions that are absent from the *D. melanogaster* reference genome. Therefore, this work provides a new strategy to identify chimeric transcripts with or without the reference genome, in a transcriptome-wide manner.

MATERIAL AND METHODS

ChimeraTE: the pipeline

ChimeraTE was developed to detect chimeric transcripts with paired-end RNA-seq reads. It is developed in BASH scripting that is able to fully automate the process in only one command-line. The pipeline has two detection Modes: (1) genome-guided, the reference genome is provided and chimeric transcripts are detected aligning reads against it; and (2) genome-blind, the reference genome is not provided and chimeric transcripts are predicted for fixed or polymorphic TEs. These Modes have different approaches that may be used for different purposes. In Mode 1, chimeric transcripts will be detected considering the genomic location of TE insertions and exons. Chimeras from this Mode can be classified as TE-initiated upstream, TE-initiated 5'UTR, TE-exonized, TE-terminated 3'UTR, and TE-terminated downstream. In addition, results from Mode 1 can be visualized in genome browsers, which allows a manual curation of chimeric transcripts in the reference genome. Mode 1 does not detect chimeric transcripts derived from TE insertions absent from the provided reference genome. Mode 2 predicts chimeric transcripts considering singletons and concordant read mappings against transcripts and TE insertions, in addition to a transcriptome assembly (user optional). Hence, Mode 2 detects chimeric transcripts from *de novo* TE insertions and an assembled genome is not necessary. In this Mode, two alignments are performed: (1) transcript alignment and (2) TE alignment. Then, based on both alignments, the pipeline identifies chimeric reads that support chimeric transcripts, regardless of the TE genomic location. In Mode 2, since there is no alignment against an annotated genome, it is not possible to classify chimeric transcripts considering the TE position, as in Mode 1.

Both ChimeraTE Modes use chimeric read pairs as evidence of chimeric transcripts. This method has been widely demonstrated by other authors as a potential source of artifactual reads, mainly due to the occurrence of mixed clusters on the Illumina's flow cell that may be too close to each other, generating read pairs that are connecting two cDNA portions that are not actually joined in the sample (35–38). Indeed, it has been shown that up to 1.56% of all reads produced by Illumina multiplexed approaches may generate chimeric reads (39), including cases that may support chimeric transcripts derived from different genes. These artifactual reads originate more likely from highly expressed genes since there are more molecules on the Illumina's flow cell. Conversely, because TE-derived sequences might comprise a low proportion of the transcriptome, artifactual reads from TEs should be produced at a low frequency. Furthermore, it is unlikely to produce artefactual reads from the same gene and TE family among RNA-seq replicates. Therefore, in order to avoid including false chimeric reads, both Modes of ChimeraTE only call chimeric transcripts that are detected in at least two RNA-seq replicates.

ChimeraTE Mode 1: genome-guided approach

In ChimeraTE Mode 1, paired-end RNA-seq reads, genome, and its respective gene/TE annotation are used to predict chimeric transcripts (Figure 1A). The genome alignment is performed with bowtie2 (40) (Figure 1B) and transcript expression is assessed with cufflinks (41). We consider FPKM >= 1 as an expressed gene by default, but it can be changed by the user with the *--fpkm* parameter. The alignment is converted to BED format with samtools (42) and bedtools (43) and reads aligned into the forward and reverse strands are separated with samtools. The IDs from reads that have aligned against genes are identified with bedtools (Figure 1B) and separated according to the gene region. The 5' UTR and 3' UTR are selected from the GFF/GTF file. Exon regions are extracted from GFF/GTF file corresponding to "CDS" for protein-coding genes and "exon" for non-coding genes. This is an important step to predict TE-exonized transcripts without counting TEs incorporated in the UTRs since "exon" into GTF/GFF files corresponds to both CDSs and UTRs. Next, reads with at least 50% of their length (*--overlap* parameter) aligned against TE copies have their IDs selected, and TE copies without aligned reads are removed from the downstream analysis. Then, expressed genes that harbor TE copies or have TE copies in their vicinity (3 Kb default but adjustable with *--window* parameter) are identified with bedtools (43). In order to identify reads where one mate has aligned against the TE copy and the other aligned into the gene regions (CDSs/UTRs), *chim_search.sh* performs several rounds of matching tests between the lists of read IDs from transcripts and TEs, generating a raw list of chimeric transcripts (Figure 1C). Then, chimeric transcript classification is performed based upon the gene feature with chimeric reads: 5' UTR: TE-initiated; exon: TE-exonized, and 3' UTR: TE-terminated. TEs located in introns are also considered in these classifications, depending on which gene feature the chimeric reads have aligned (Figure 1D). TE-initiated upstream and TE-terminated downstream are assigned depending upon the TE location regarding gene compartments. These steps are repeated for all RNA-seq replicates provided in the input. Next, the raw results from replicates are compared, and all chimeric transcripts that have been identified with at least >= 2 chimeric reads and were found in >= 2 replicates, are considered as true chimeric transcripts (Figure 1C). These thresholds may be changed by the user with *--cutoff* and *--replicate* parameters. Mode 1 output is a table with a list of predicted chimeric transcripts categorized

by TE position, with gene ID, TE family, chimeric read coverage, TE location, gene location, and gene expression (Figure 1D).

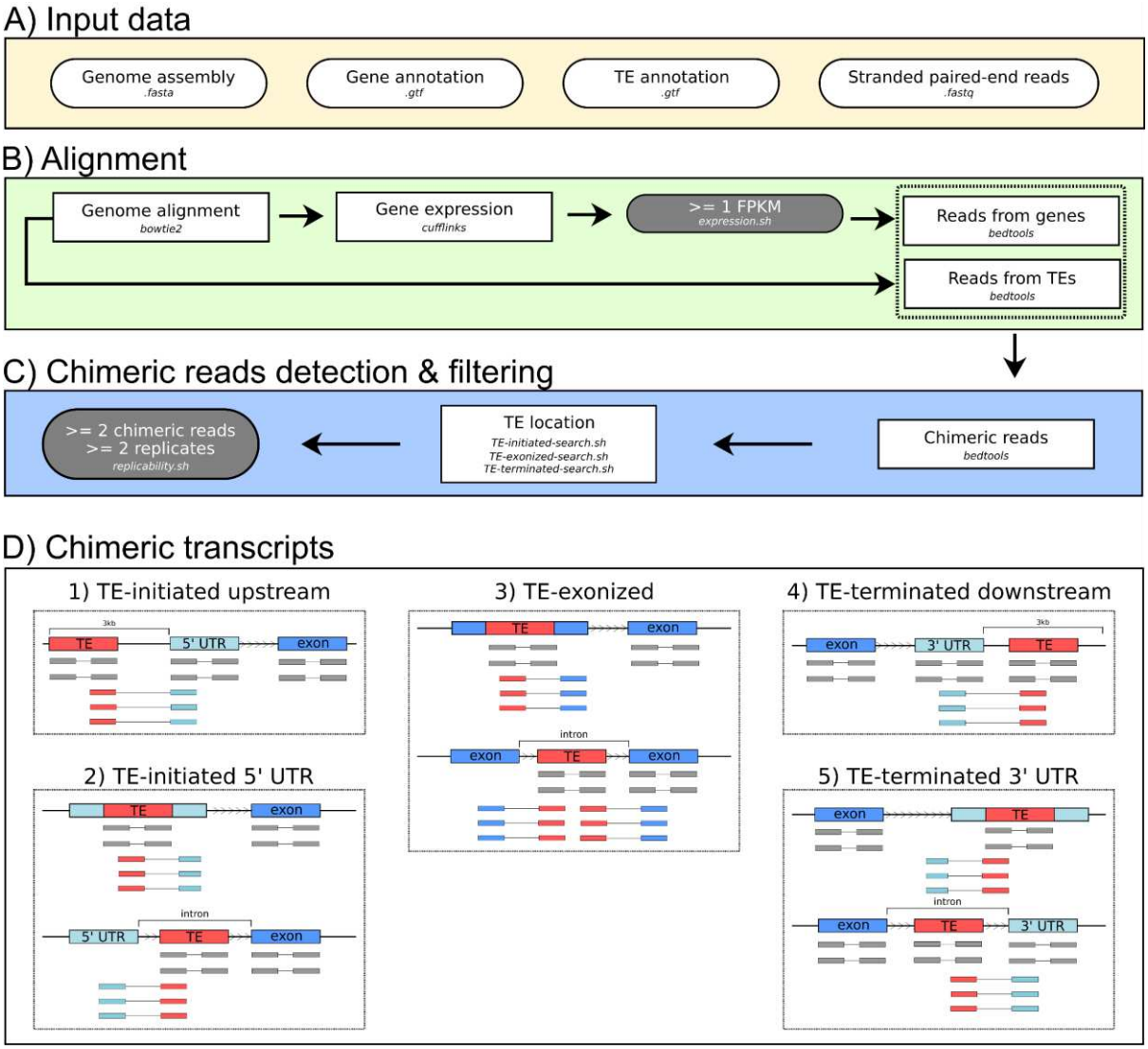

Figure 1: ChimeraTE Mode 1 (genome-guided) workflow. Round white boxes: input data; square boxes: pipeline step; round gray boxes: thresholds that can be modified. **A)** Input data: fasta file with the genome assembly, gtf files with gene and TE annotations, as well as stranded paired-end reads from RNA-seq (fastq). **B)** Alignment: The genome alignment is used to calculate gene expression levels. Genes with FPKM =< 1 are removed from downstream analyses. A subsequent list of reads that have aligned against genes or TE insertions is created. **C)** Chimeric reads detection & filtering: Both read lists are then compared and read pairs that have common reads between the two lists are named chimeric reads, *i. e.*, paired-end reads mapping to a gene and a TE copy. The sum of these reads is used as chimeric reads coverage for each putative chimeric transcript. All putative chimeras are then processed with three ChimeraTE scripts to categorize them into TE-initiated, TE-exonized, and TE-terminated transcripts. These steps are run for all RNA-seq replicates. Finally, all chimeric transcripts present in at least 2 replicates and with at least 2 chimeric reads as support are maintained. **D)** Chimeric transcripts: Five predictions obtained from Mode 1. “Exon” blue boxes are representing CDS regions and exon regions from ncRNA genes. Head arrow in between TE, UTR, and exon boxes: transcription sense; gray boxes linked by a line: non-chimeric paired-end reads; blue and red boxes linked by a line: chimeric paired-end reads. The ChimeraTE mode 1 output is divided into five predictions: (1) TE-initiated upstream: the TE insertion is located upstream of the gene region; (2) TE-initiated 5' UTR: the TE insertion may be located either inside the 5' UTR, or in an intron, but chimeric

paired-end reads involving the 5' UTR are present; (3) TE-exonized: the TE insertion is present within exons, or introns, while chimeric paired-end reads involving the exon are present; (4) TE-terminated downstream: the TE insertion is located downstream of the gene; (5) TE-terminated 3' UTR: the TE insertion is either located inside the 3' UTR, or at intron while chimeric paired-end reads involving the 3' UTR are present.

ChimeraTE Mode 2: a genome-blind approach to uncover chimeric transcripts

ChimeraTE Mode 2 is the genome-blind approach of the pipeline. The input data are stranded RNA-seq reads, reference TE insertions, and gene transcripts (Figure 2A). The data will be used to perform two alignments with *bowtie2* (44), one against all transcripts and another against all TE sequences, both with parameters: *-D 20 -R 3 -N 1 -L 20 -i S,1,0.50* (Figure 2A). In order to avoid very low-expressed transcripts predicted as chimeras, as well as decrease processing time, the SAM alignment is converted to BAM with *samtools* (42), and FPKMs are computed for the reference transcripts provided in the input using *eXpress* (45). Then, all genes with average FPKM < 1 are removed from the downstream analysis (Figure 2B). In order to identify chimeric reads between TEs and gene transcripts, both alignments are converted to BED with *bedtools* (43). Among all aligned paired-end reads, the pipeline considers as chimeric transcripts the ones that have at least one read aligned to the TE sequence (singleton mapping) and its mate to the gene transcript, or when both reads have aligned (concordant mapping) to the TE and gene transcript. In order to identify these reads, the TE alignment output is used to create a list with all read 1 IDs that have aligned against TEs, and another list with all read 2 IDs, regardless if their mates have also aligned (concordant mappings or singleton mappings). The same lists are created by using the transcript bed file: (1) read 1 IDs of transcript mapping reads and (2) mate 2 IDs of all mate 2 reads, regardless of mate mapping. All mate 2 IDs that have a TE-aligned read 1 are searched in the list of transcript-aligned mate 2. The same is performed in the opposite direction (TE-aligned read 2, transcript-aligned mate 1). These read pairs will therefore be comprised of two mates from the same pair that were singletons in the alignments, *i.e.*, pairs comprised of one read that has aligned against a TE, and its mate against a gene transcript. The cases in which the chimeric transcript does not have the TE insertion in the reference transcript, it will be supported only by these singleton chimeric reads. For cases in which the TE insertion is present inside the reference transcript, chimeric reads supporting it may either be singleton or concordant reads. Therefore, chimeric reads can be concordant reads in both alignments (TEs and genes), or they may be concordant only in the gene transcript alignment and singleton in the TE alignment. Due to the repetitive nature of TEs, short-read alignment methods provide very few unique aligned reads against loci-specific TE copies as most reads align ambiguously between similar TE insertions. Therefore, when a chimeric transcript has been identified involving more than one TE family, the TE family with the highest coverage of chimeric reads is maintained. Subsequently, ChimeraTE uses two chimeric reads as a threshold for calling a chimeric transcript, that can be modified by the user with the *--cutoff* parameter. Such value does not represent transcript expression nor TE expression, but it represents the coverage supporting the junction between a gene transcript (CDSs/UTRs) and a TE sequence. Finally, the output tables show the list of genes and the respective TE families detected as chimeras, reference transcript ID, and the total coverage of chimeric reads supporting it.

The support of chimeric transcripts performed by ChimeraTE Mode 2 is from chimeric reads aligned by an *end-to-end* approach. Such alignment may reduce alignment sensitivity, since exon/TE junctions may be covered by split reads. In order to mitigate the loss of detection power due to the alignment method with Mode 2, alongside with chimeric read detection using alignments against transcripts and TEs, there is an option to run Mode 2 with a transcriptome assembly approach, which can be activated with *--assembly* parameter (Figure 2C). This approach will use RNA-seq reads to perform a *de novo* transcriptome assembly with Trinity v2.8.5 (46). In order to identify assembled transcripts that may have TE-derived sequences, masking is performed with RepeatMasker v4.1.2 (47), providing *--ref_TEs*, a custom TE library, or pre-defined TE consensuses from Dfam v3.3 (48), according to the taxonomy level, i.e: flies, mouse, humans. Then, RNA-seq reads are aligned with bowtie2 (44) against the assembled transcripts. Subsequently, the alignment is used to identify whether transcripts containing TE-derived sequences have chimeric reads, considering those from split reads. All assembled transcripts with chimeric transcripts are selected as candidate chimeric transcripts. Next, these candidates are submitted to a homology analysis with Blastn v2.11.0+ (49) with reference transcripts. Finally, all assembled transcripts with masked TEs that have at least 80% of similarity with reference transcripts across 80% of its length (can be modified with *--min_length* parameter) are considered as chimeric transcripts. All these steps are repeated for all RNA-seq replicates provided in the input. Finally, the list of chimeric transcripts obtained from all replicates with the transcriptome assembly approach is compared, and all chimeras that have been identified with at least >= 2 chimeric reads and were found in >= 2 replicates, are considered as true chimeric transcripts. By activating the *--assembly* option in Mode 2, the output table will provide chimeric transcripts that have been predicted based on different evidences (Figure 2D): (1) only based on chimeric reads; (2) only based on transcript assembly; (3) based on chimeric reads and transcript assembly.

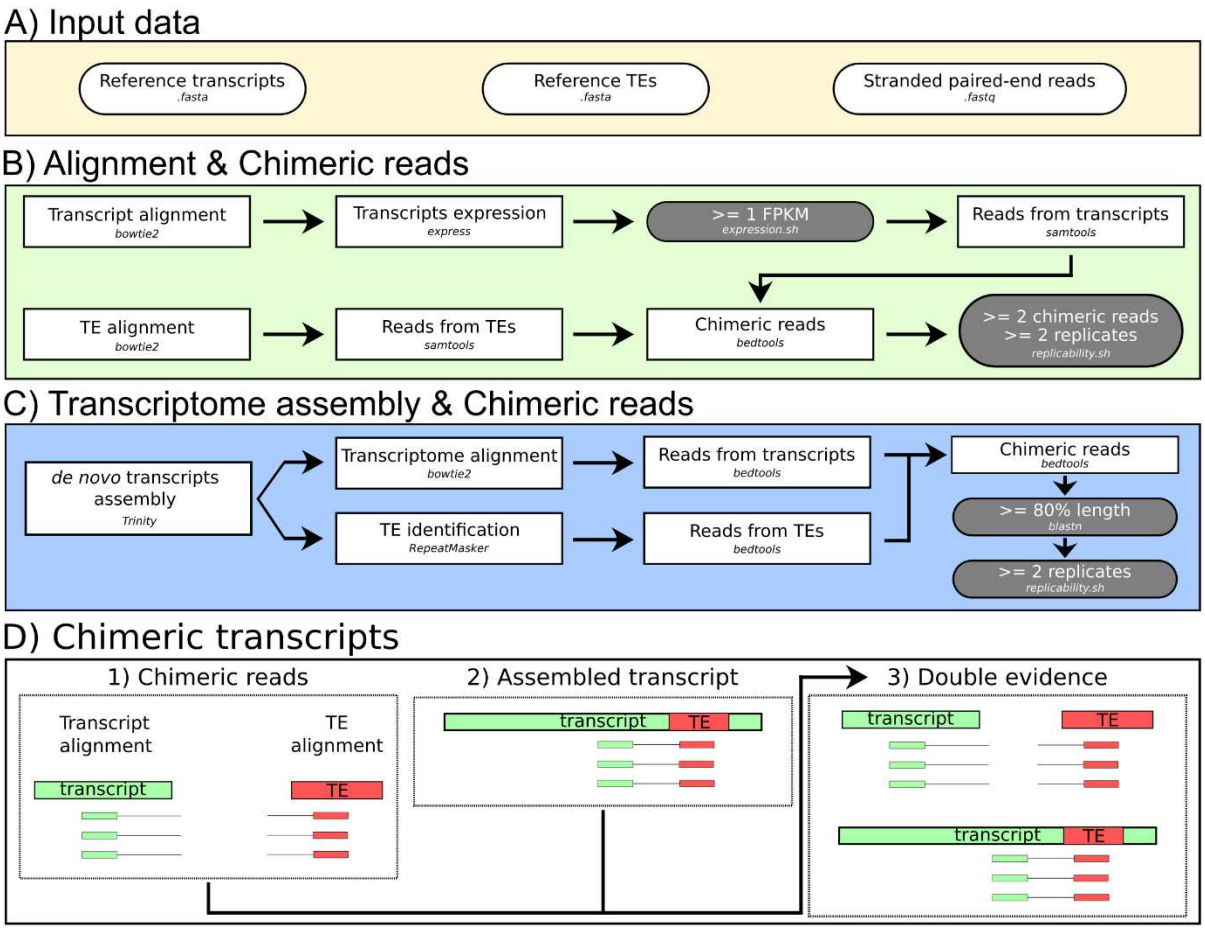

Figure 2: ChimeraTE Mode 2 (genome-blinded) workflow. Round white boxes: input data; square boxes: pipeline step; round gray boxes: thresholds that can be modified**. A**) Input data: two fasta files containing reference transcripts and TE insertions, as well as stranded paired-end reads from RNA-seq (fastq). **B**) Alignment and chimeric reads: The alignment against transcripts is performed and their expression is calculated. Transcripts with FPKM =< 1 are removed from the downstream analysis. Next, a list of reads aligned against transcripts is created. Through the alignment of reads against TE insertions, a second list with reads stemming from TEs is also created. Then, mapped paired-end reads and singletons are identified, generating the list of chimeric reads, for all replicates. All chimeric transcripts that have an average of chimeric reads >= 2 and are present in >= 2 replicates are maintained as true chimeras. **C**) Transcriptome assembly and chimeric reads: The *de novo* transcriptome assembly is a non-default option of ChimeraTE Mode 2. It performs a transcriptome assembly and aligns reads against the assembled transcripts. Then, TE insertions in the assembled transcripts are identified with RepeatMasker and the TE reads are recovered. Using the two lists of reads (transcripts and TEs), the chimeric read list is generated and the putative assembled chimeric transcripts are predicted. Next, it is performed a blastn between these transcripts and the reference transcripts provided in the input. All transcripts with length >= 80% are selected. The process is repeated for all RNA-seq replicates and chimeric transcripts assembled >= 2 replicates are maintained as true chimeras. **D**) Chimeric transcripts If the assembly is activated, ChimeraTE mode 2 provides three outputs: (1) Chimeric reads: These chimeric transcripts were predicted only based on the method demonstrated in -B-; (2) Assembled transcripts: Chimeric transcripts predicted only based on the transcriptome assembly method demonstrated in –C-; and (3) Double evidence: Chimeric transcripts predicted by both methods -B and C-.

D. melanogaster wild-type strains: genome assemblies and RNA-seq

In order to assess ChimeraTE's performance, as well as the efficiency in the identification of chimeric transcripts derived from polymorphic TE insertions, we have mined previously available RNA-seq data from four *D. melanogaster* wild-type strains (50). Two from France, Gotheron (44_56'0"N 04_53'30"E), named dmgoth101 and dmgoth63; and two from Brazil, São José do Rio Preto (20_41'04.3"S 49_21'26.1"W), named dmsj23 and dmsj7. The RNA-seq from ovaries was sequenced on an Illumina Hiseq (125 bp reads), with two biological replicates. All RNA-seq libraries were trimmed by quality, and adapters were removed with Trimmomatic (51). Each strain had also its genome previously sequenced by Nanopore long reads and assembled (52). The high-quality assemblies allowed us to manually check whether chimeric transcripts predicted by both ChimeraTE Modes have the predicted TE insertions inside/near genes, as well as manually curate the presence of chimeric reads.

Running ChimeraTE with *D. melanogaster* data

To run ChimeraTE Mode 1 on the available RNA-seq data, we performed gene annotation in the four *D. melanogaster* genomes with Liftoff (53) using default parameters and the *D. melanogaster* genome (*dm6* strain) available in Flybase r6.46 as reference (dmel-all-chromosome-r6.46.fasta.gz). TE annotation was performed with RepeatMasker v4.1.2 (47), with parameters: *-nolow*; *-norna*; *-s;* and *-lib* with the TE sequence library for *D. melanogaster* provided by Bergman's lab (https://github.com/bergmanlab/drosophila-transposons/blob/master/current/D_mel_transposon_sequence_set.fa). The annotation from RepeatMasker was then parsed with One Code to Find Them All (54), in order to merge LTRs and Internal regions from the same TE family and also merge fragments of the same TE family that are up to 50 bp (*--insert* 50) close to each other. We used ChimeraTE Mode 1 with default parameters and the mode *--utr* activated, which enables the search for chimeric transcripts with 5' and 3' UTRs. For ChimeraTE Mode 2, we aimed to demonstrate its potential in detecting chimeric transcripts derived from TE insertions that are not present in a reference genome, even though the transcript sequences and TE copies provided as input were generated from the *dm6* reference genome. Therefore, we have used ChimeraTE mode 2 with RNA-seq from the four wild-type strains with reference transcripts from *D. melanogaster* (*dm6* strain) available in Flybase r6.43 (http://ftp.flybase.net/genomes/Drosophila_melanogaster/current/fasta/dmel-all-transcript-r6.46.fasta.gz) (55). The TE annotation for the *dm6* genome was assessed with the same protocol used on the four wild-type strains, through RepeatMasker (47) and One Code to Find Them All (54).

Benchmarking polymorphic chimeric transcripts with Nanopore genomes

Once chimeric transcripts were identified, we used the high-quality Nanopore assemblies for dmgoth101, dmgoth63, dmsj23, and dmsj7 previously published (52) to confirm whether genes predicted by Mode 1 as chimeric transcripts have indeed the respective TE insertion located near or within them. To do so, we have used an *ad-hoc* bash script (*create-up-down-BED.sh*) to create three bed files from genes: 3 Kb upstream; 3 Kb downstream, and gene region. Then, we used *bedtools intersect* (43) to identify genes with TEs located in the three regions. For Mode 1 we have randomly sampled 100 chimeric transcripts of each wild-type strain to visualize the alignments performed by Mode

1 on the IGV genome browser (56). For Mode 2, all genes not found by *bedtools intersect* with the predicted TE insertion were visualized in IGV. In both manual curations, we considered as false positives those cases in which we did not find the TE insertion, or we found the TE insertion, but without chimeric reads.

In order to assess the number of chimeric transcripts found by Mode 2 in wild-type strains derived from TE insertions absent in the *dm6* genome, we also used the *ad-hoc* bash script (*create-up-down-BED.sh*) to create the bed files with 3 Kb +/- and the gene regions for *dm6*. Then, we used *bedtools intersect* (43) with TE annotation and the gene regions. By using this method, we generated a list of genes with TEs located 3 Kb upstream, inside genes (introns/exons), and 3 Kb downstream for the *dm6* genome. Then, the polymorphic chimeric transcripts were identified with the comparison of genes with TEs inside/nearby in *dm6* and the list of chimeric transcripts in the wild-type strains. In addition, all chimeric transcripts derived from TEs that were not found in *dm6* were manually curated with the IGV genome browser (56).

Functional enrichment analysis for genes generating chimeric transcripts

The genes generating chimeric transcripts in the four wild-type strains were submitted to functional enrichment analysis with DAVID (57), selecting for biological processes. Only gene ontology terms with p-value (*Bonferroni correction*) < 0.05 were selected. Redundant terms were removed with REVIGO, with 0.5 of reduction size.

Sequence and protein analysis of TE-exonized *roo* elements

The sequences of TE-exonized *roo* elements were extracted from wild-type genomes with *bedtools getfasta* (43), parameter *-s,* using BED files created by ChimeraTE Mode 1. In order to evaluate that these insertions are not repetitive DNA widespread across the genomes, we used Blastn v2.11.0+ to search them, with the following parameters: *-dust no*; *-soft_masking false*; *-qcov_hsp_perc 90*; *-perc_identity 90*. Then, because the TE reading frame incorporated into the gene transcript is unknown, all insertions were translated with *EMBOSS v6.6.0 transeq* in the three coding frames. Next, the protein domains in these sequences were assessed with Batch CD-Search (58), with default parameters. The conservation of TE-exonized *roo* copies was assessed with multiple sequence alignment, performed separately for each strain, with MUSCLE (59), implemented by MEGA X (60), with default parameters. High extension gaps caused by less than ~5% of TE insertions were removed manually with Aliview v1.28 (61). The alignments were plotted with MITOS (62), using *the Plots* package.

RESULTS AND DISCUSSION

ChimeraTE predicts chimeric transcripts derived from genes and TEs using two different strategies. Mode 1 is a genome-guided approach that will predict chimeric transcripts from paired-end RNA-seq through chimeric read pair detection. The main advantages of Mode 1 in comparison to Mode 2 is that the first one is able to detect split reads between gene regions (CDSs/UTRs) and TEs, capture chimeric transcripts with low coverage/expression and classify chimeric transcripts according to the TE position: TE-initiated upstream, TE-initiated 5' UTR, TE-exonized, TE-terminated 3'UTR, and TE-terminated downstream transcripts. However, Mode 1 misses chimeric transcripts derived from TE insertions that

are absent from a reference genome. Mode 2, which is a genome-blind approach, performs two alignments against reference transcripts and TE copies and then, similar to Mode 1, predicts chimeric read pairs between these two alignments. In addition, Mode 2 can optionally perform a *de novo* transcriptome assembly able to detect chimeric transcripts and chimeric read pairs through split read alignment, improving the sensitivity.

Setting up the datasets for ChimeraTE

Each ChimeraTE Mode requires different input datasets. To run Mode 1 (genome-guided), gene and TE annotations, along with a genome fasta file are necessary. We took advantage of available paired-end RNA-seq datasets from ovaries of four wild-type strains of *D. melanogaster,* for which high-quality genome assemblies were also available (52). We performed gene and TE annotation in the new assemblies using *D. melanogaster*'s genome (*dm6*) from Flybase r6.46 as a reference, and obtained ~17,357 genes, of which only ~64 were partially annotated (Sup. Table 1). Regarding TE annotation, we found ~10.29% of TE content in the four wild-type strains, similar to our previous estimations for these strains (52). In total, 128 TE families in the four genomes were uncovered, comprising the mean of ~20,754 TE insertions (*standard deviation* = 892). In all genomes the TE content in bp is higher for LTR, then LINE elements, followed by DNA and Helitron families (Sup. Table 1). In order to run Mode 2 (genome-blind), we used reference transcripts from Flybase r6.46 and performed TE annotation on the reference *dm6* genome. We have obtained 27,131 TE insertions, representing 16.14% of the genome content, following the same proportions as seen for the four wild-type genomes (Sup. Table 1).

ChimeraTE Mode 1 reveals that ~3% of genes produce chimeric transcripts in *D. melanogaster* wild-type strains

ChimeraTE mode 1 was run on the four wild-type strain genomes and their respective ovarian RNA-seq data (63). Across all strains, we found 506 genes producing chimeric transcripts, representing 2.83% of the total genes in *D. melanogaster* and 6.15% of the expressed genes (FPKM >1). In order to verify whether ChimeraTE identified the correct TE family for each chimera, we have compared the genomic coordinates of TEs and genes (3 Kb upstream/downstream and inside genes) with *bedtools intersect* (43) and predicted genes with TE copies in these regions. All chimeric transcripts in the four wild-type strains had at least one TE insertion in the expected position from the predicted TE family (Sup. Table 2-5). In addition, we randomly selected 100 chimeric transcripts in each wild-type strain to visualize in IGV (56) and confirm the presence of chimeric reads as expected (Fig. 3A). All the 400 manually inspected chimeras were correctly found in the genome browser. Among all chimeric transcripts, 56.23% correspond to TE-exonized, 21.78% to TE-3' UTR, 12.53% to TE-5' UTR, 8.06% to TE-downstream and 1.38% to TE-upstream transcripts (Fig. 3B). TE-exonized transcripts are derived from TE insertions that may be inside exons or introns. However, the high prevalence of these chimeras might be associated with potential cases of TE copies inside genes generating chimeras where the TE provides TSSs (TE-initiated transcripts) or the PolyA sites (TE-terminated transcript), but due to the evidence of chimeric reads from TEs and exons, ChimeraTE classify them as TE-exonized.

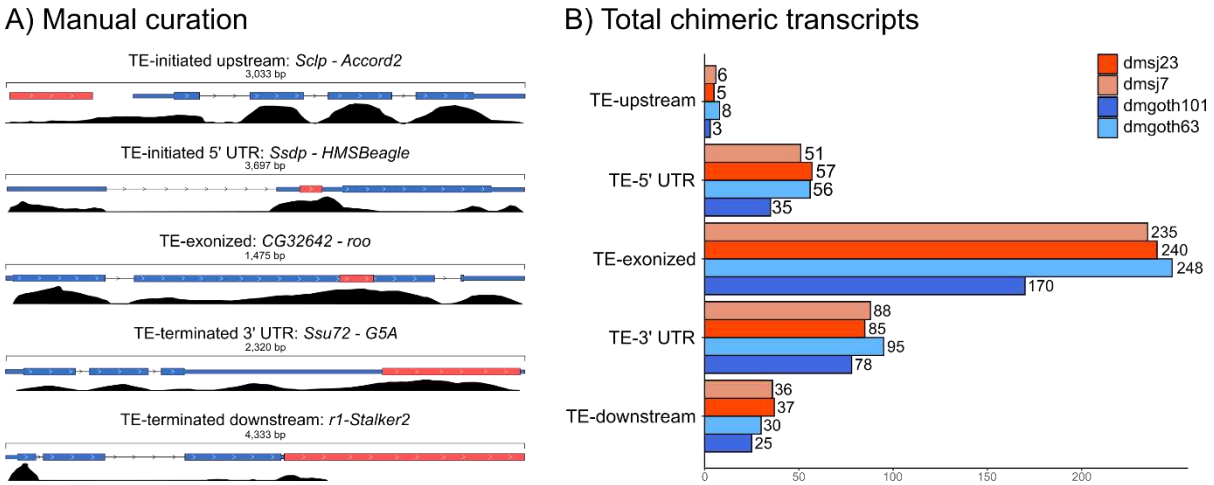

Figure 3: General results from Chimera Mode 1. **A**) Five examples of chimeric transcripts manually curated with IGV genome browser. Pink boxes: TE insertion; blue boxes: exons and UTRs; black density graphs: coverage of RNA-seq reads; head arrows: transcription sense. **B**) Total chimeric transcripts: Number of chimeric transcripts following the TE position classification in the four wild-type strains.

Among all chimeric transcripts predicted by ChimeraTE Mode 1, 16 chimeras have been previously described in *D. melanogaster* as TE-initiated transcripts, with CAGE data (31). From these, three chimeras were found in all strains: *Ssdp* (gene) - *HMS-Beagle* (TE family); *Agpat1-1360*; *anne-1360*. There are additional five chimeric transcripts also found in all strains, but with other TE families than the previously published: *PlexA-INE-1*; *Atf6-INE-1*; *CG2162-roo*; *CR43361-1360*, and *Hcf-INE-1*, whereas the previously observed TE families generating these chimeras were *Tc1*, *1360*, *S*, *invader4* and *1360*, respectively. We manually checked these chimeric transcripts in the IGV genome browser and we confirmed the presence of the TE families predicted by ChimeraTE instead of the TEs described previously (Sup. Table 6). Finally, the other eight chimeric transcripts harbor the TE family described in previous studies (31), but are not present in all strains: *ctp-HMSBeagle*; *CG6191-jockey*; *udd-297*; *Bmcp-Doc*; *Rnf11-Stalker2*; *CHKov1-Doc*; *Sumo-mdg1* and *Svil-roo* (Sup. Table 6). In addition, we have also found the *Kmn1-pogo* TE-terminated 3'UTR transcript in dmgoth63 and dmsj23, which has been previously described as an adaptive copy in *D. melanogaster*, increasing the resistance to insecticides (12). In *D. melanogaster* midbrain, a study has shown by single-cell RNA-seq that 264 genes produce chimeric transcripts (38). Despite the differences between tissues and methods, we have found 19 chimeric transcripts identified by this study (Sup. Table 7), of which four were previously found by CAGE (31). From these 19, six chimeric transcripts are derived from *roo* elements, which has been characterized as a TE family providing splice donor and acceptor sites, creating new isoforms during alternative splicing (32). Taken together, the genome-wide analysis performed by ChimeraTE Mode 1 has uncovered 483 genes with chimeric transcripts in the *D. melanogaster* ovarian tissue for the first time.

ChimeraTE Mode 2: a method to uncover chimeric transcripts without genome alignment

D. melanogaster has a high rate of TE insertion polymorphism across worldwide populations (64–67). In order to test the ability of ChimeraTE Mode 2 in detecting chimeric transcripts derived from TE

insertions absent from a reference genome, we used RNA-seq from the four wild-type strains with transcript and TE insertion annotations from the *D. melanogaster* reference genome (*dm6*). ChimeraTE Mode 2 may predict chimeric transcripts based on two evidences: either using chimeric reads only, or by taking advantage of *de novo* transcriptome assembly. Chimeric transcripts with both evidences are named as “double-evidence” (chimeric reads and transcript assembly - Figure 2B and C). Among the four wild-type strains, ChimeraTE Mode 2 has identified 378 genes (Sup. Table 8-11) producing chimeric transcripts (Figure 4A), representing 2.11% of the total *D. melanogaster* genes available in our analysis. Comparing the “chimeric read” approach with the “transcript assembly” one, the latter method has found twice as many chimeric transcripts than the detection by chimeric reads (Figure 4A). This is probably due to the possibility of aligning chimeric reads in the junctions of TEs and exons, which is not possible in the chimeric read approach (Figure 2B) because the alignment of TEs and transcripts are performed separately. Indeed, we found that chimeric transcripts detected only by the transcriptome approach had more chimeric read coverage than those detected only by alignments with reference transcripts and TEs (Figure 4B). Furthermore, transcripts with low expression (FPKM < 5) were also more efficiently detected by the transcriptome approach. Conversely, chimeric transcripts with high expression were mostly detected by chimeric reads evidence. In total, ~30.04% of all chimeric transcripts were found using both methods.

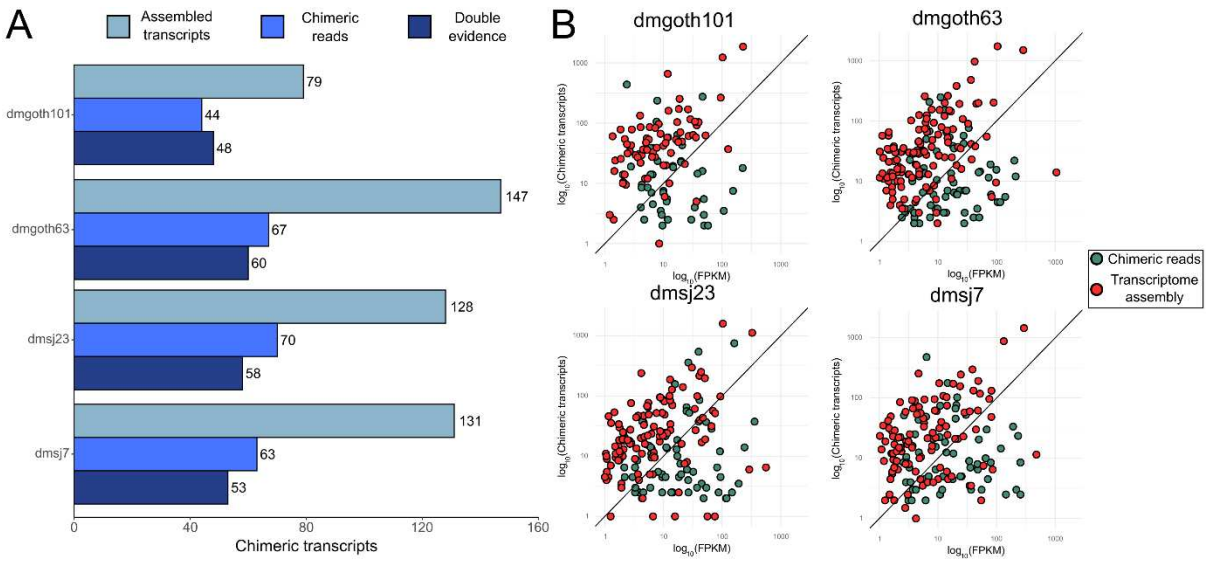

Figure 4: **A**) Total number of chimeric transcripts found by ChimeraTE mode 2. “Assembled transcripts”: chimeric transcripts detected only by the method of transcriptome assembly (Figure 2C). “Chimeric reads”: chimeric transcripts detected only by the method of chimeric read pairs (Figure 2B). “Double evidence”: chimeric transcripts detected by both methods. **B**) Comparison between chimeric transcripts found only by "transcriptome assembly" and "chimeric reads". In all strains, chimeric transcripts with the highest coverage of chimeric reads were detected by transcriptome, and chimeras found only by chimeric reads are those with high expression.

As each chimeric transcript was detected based on reference transcripts and TE insertions, we have used the Nanopore assemblies to manually inspect the presence of the predicted TE family inside and near (+/- 3 Kb) genes, with the intersection of genomic coordinates from genes and TEs with *bedtools intersect* (43). We considered as true chimeric transcripts the cases in which we found the presence of

the predicted TE insertion inside/near the gene. The alignment of RNA-seq libraries against the genome sequence was also used to confirm the presence of the TE insertion, as well as the presence of chimeric read pairs, with the IGV genome browser (56). The manual curation was performed with the three groups of results from Mode 2 ("chimeric reads", "transcriptome assembly" and "double-evidence"). We found that 96.30% of all chimeric transcripts predicted by double evidence were true, whereas we observed 90.59% from transcriptome assembly and 73.13% from chimeric reads. Therefore, taken together, ChimeraTE Mode 2 has provided a reliable inference with 86.68% of sensitivity, based upon genomic manual curation.

The main difference of ChimeraTE Mode 2 from the previous pipelines is its ability in detecting chimeric transcripts derived from TEs that are absent in a reference genome, using RNA-seq from non-reference individuals/cells/strains. We first identified in the *dm6* reference genome the genes with TEs located upstream, inside, and downstream genes. We found that 2,239 genes in the *dm6* genome have TEs located 3 Kb upstream, 1,863 inside (introns and exons), and 2,320 downstream. These genes were selected as potential chimeras in the *dm6* genome and then compared with the list of chimeric transcripts generated by ChimeraTE Mode 2 in the four wild-type strains. In addition to the comparison with *dm6* genes harboring TEs inside/near, we have manually curated the TE insertions and the presence of chimeric transcripts with the IGV genome browser (68). For instance, the *Mps1-FB* chimera is an interesting case, since it is the only chimeric transcript specific to French populations, as we found it in both dmgoth101 and dmgoth63. In the *dm6* reference genome, *Mps1* has an overlap between its 3' UTR and the 3' UTR of the *alt* gene, located in the other strand (Figure 5A). The same distribution of both genes is found in the Brazilian strains, dmsj23, and dmsj7 (Figure 5B). Conversely, in dmgoth101 and dmgoth63, there is a gap of ~9,500 bp between *Mps1* and *alt*, with four small *FB* copies with ~412 bp of length (Figure 5C), indicating that it is an old insertion. Furthermore, one of the *FB* copies is located downstream of the *alt* gene, which also has been identified as TE-terminated downstream in dmgoth63. This result shows that ChimeraTE is able to detect chimeric transcripts derived from TEs that are not present in the reference genome.

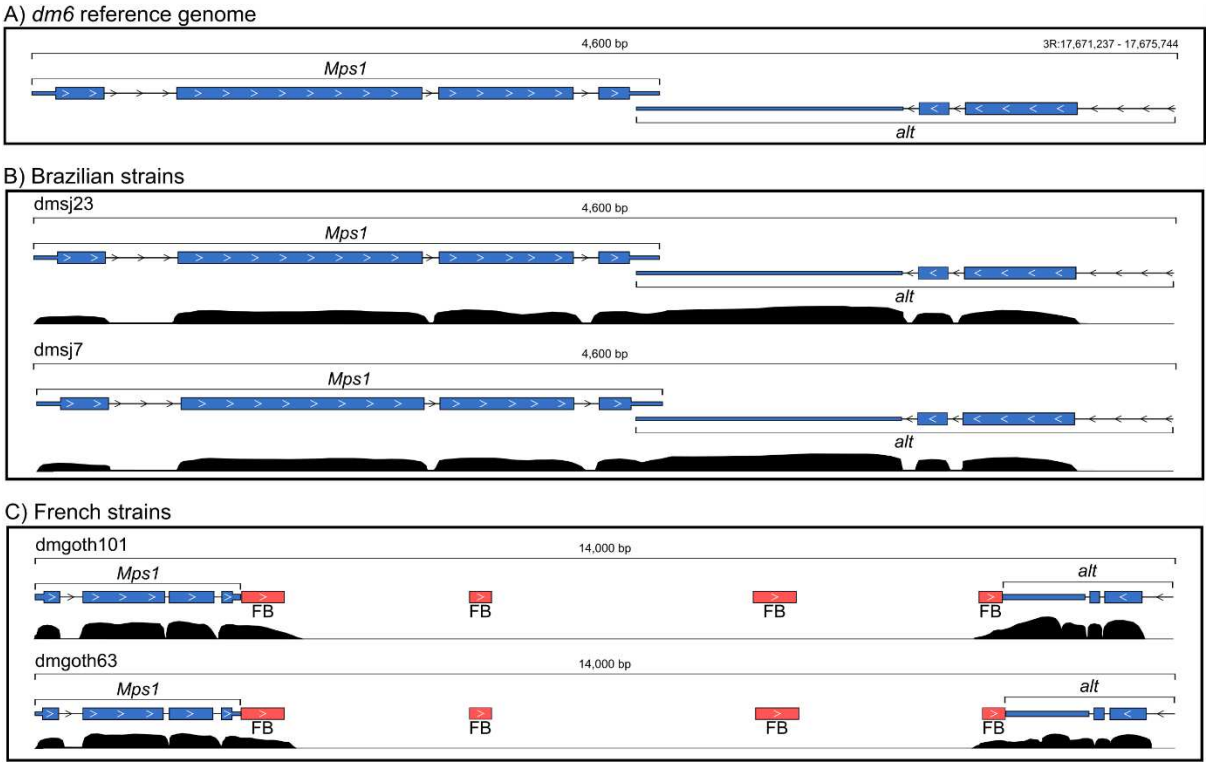

Figure 5: *Mps1* gene and its downstream region in the four wild-type strains genome. **A**) *Mps1* in the *dm6* reference genome and the *alt* gene located downstream to it, in opposite strands and with overlapped 3' UTRs. **B)** In the dmsj23 and dmsj7, *Mps1* and *alt* are distributed as found in the reference genome. **C**) In dmgoth101 and dmgoth63, there is a *FB* insertion located downstream to *Mps1*, which has chimeric read pairs supporting a TE-terminated downstream in both strains.

Taking into account all chimeric transcripts detected by Mode 2, we found 11 genes generating chimeric transcripts derived from TEs that are absent in the *dm6* genome (Table 1). There are specific chimeras from French strains: *Mps1-FB*, *CG1358-S,* and *CG46280-POGON1* genes, but only *Mps1* had chimeric transcripts in both French dmgoth63 and dmgoth101 strains, whereas *CG1358* was found as chimera only in the dmgoth101, and *CG46280* only in dmgoth63. The *Ythdc1-roo* chimera was observed with a strain-specific polymorphic *roo* element inside an exon in the dmgoth63 strain. Regarding the Brazilian strains, we found two TE insertions present in both dmsj23 and dmsj7, from the genes *cic* and *TrpRS-m*, but they were found as chimeras only in the dmsj23 strain. We also found *rb*, *r-1*, and *ArfGAP1* with dmsj23-specific TE insertions, whereas in the dmsj7 we found *caps* and *Ubi-p5E* with specific TE insertions giving rise to chimeric transcripts.

Gene	Function	TE	TE location	Strains			
				dmgoth101	dmgoth63	dmsj23	dmsj7
Mps1	meiotic and mitotic spindle assembly checkpoints	FB4	Downstream	▲	▲		
rb	lipid storage, eye pigment biogenesis; *Notch* receptor	MDG3	Downstream			▲	
r-l	pyrimidine biosynthesis	Stalker2	Downstream			▲	
ArfGAP1	Enables GTPase activator activity	POGO	Downstream			▲	
caps	axon guidance; morphogenesis	POGON1	Exon				▲
Ythdc1	regulation of alternative splicing; sex determination	ROO	Exon		▲		
CG1358	heme export	S	Downstream	▲	△		
TrpRS-m	catalyze the ligation of tryptophan to its cognate tRNA	PROTOP_A	Exon			▲	△
Ubi-p5E	protein ubiquitination	Gypsy	Upstream				▲
cic	transcriptional repressor	POGON1	Exon			▲	△
CG46280	unknown	POGON1	Exon	△	▲		

Table 1: 11 polymorphic chimeric transcripts from TE insertions that are not present in the *dm6* reference genome. The red triangle in a line represents the presence of a chimeric transcript; the white triangle in a line represents the presence of the TE insertion, but without a chimeric transcript; the line without triangles represents the absence of the TE insertion.

Differences between Mode 1 and Mode 2

ChimeraTE Mode 1 and Mode 2 use different alignment strategies and downstream approaches to detect chimeric transcripts (see Methods). In order to test whether these differences can lead to different outputs, we compared the chimeras from Mode 1 and Mode 2 by using the same RNA-seq libraries from four *D. melanogaster* wild-type strains. Taken all strains together, Mode 2 has uncovered 31.57% of all TE-initiated upstream cases detected by Mode 1; 57.69% of TE-initiated 5' UTR; 68.25% of TE-exonized; 80% of TE-terminated 3' UTR, and 35.82% of TE-terminated downstream (Figure 6A). These results indicate that ChimeraTE Mode 2 had low efficiency (~33.69%) to detect chimeric transcripts from TE insertions near genes. In addition, for chimeras derived from TE inside genes, most of them were detected by the transcriptome assembly approach in Mode 2, showing the relevance of performing this optional analysis. However, it must be considered that *–assembly* performed by Mode 2 is time-consuming, as well as hardware-consuming (Sup. Table 12).

In both ChimeraTE Modes, the main evidence used to detect chimeric transcripts is the presence of chimeric reads, which are paired-end reads spanning between TE and gene sequences. In Mode 1, at least 50% of one read (default parameter) from the read pair must align against the TE insertion, whereas in Mode 2 the whole read must align against a TE copy. Therefore, the alignment method performed by Mode 2 does not allow the detection of chimeric transcripts derived from TE copies shorter than the read size. Hence, we investigated whether chimeric transcripts found with Mode 1, but not with Mode 2, are generated by TE insertions shorter than the sequenced reads. From 286 chimeric transcripts found only by Mode 1, we found that 136 (47.55%) are shorter than the reads, making it impossible to detect them with Mode 2 (Figure 6-B). In addition, it is important to highlight that TEs longer than reads may have splice sites, generating chimeric transcripts with a small TE-derived fragment, not being detected by Mode 2 as well. Furthermore, we observed that 93 (32.51%) chimeric transcripts from TEs longer than our reads have coverage lower than 10 chimeric reads. We

hypothesized that Mode 1 may have substantially more TE-aligned reads than Mode 2, due to the use of strain-specific TE insertions and by counting split read alignment.

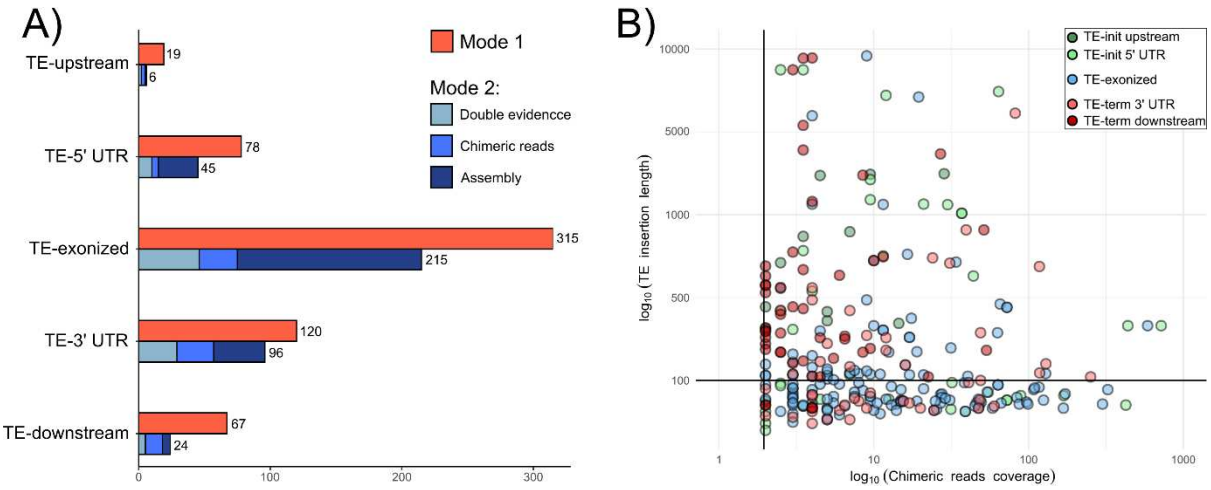

Figure 6: **A**) Total chimeric transcripts detected by ChimeraTE mode 1 and ChimeraTE mode 2. The three blue boxes mean the type of evidence used by Mode 2 to support the chimeric transcripts (see Methods). Mode 2 had more efficiency to detect chimeric transcripts derived from TEs inside genes (TE-5' UTR, TE-exonized, and TE-3' UTR) than near genes. **B**) Chimeric transcripts found by mode 1, but not by mode 2. 47.55% of all chimeric transcripts detected only by Mode 1 have TEs shorter than the read length, and 32.51% of chimeras with TEs longer than reads have low chimeric reads coverage. These factors explain the differences between results found by both Modes.

ChimeraTE Mode 1 is dependent on a reliable genome annotation for both genes and TEs, contrary to Mode 2. Indeed, we found in total seven chimeric transcripts detected only with Mode 2 (Sup. Table 14), which are derived from genes that were not annotated in the wild-type genomes. We compared them with the *dm6* genome and we found all seven have the predicted TE family inserted near/inside the gene, indicating that these predictions are probably true (Sup. Figure 1). For instance, the chimeric transcript *CG3164-McClintock* was detected by Mode 2 with double evidence in the four wild-type strains, and it is not annotated in any of the four genomes, perhaps due to its location in the telomeric region of the 2L chromosome (Sup. Figure 1G).

ChimeraTE Mode 1 and Mode 2 detected 506 and 378 genes producing chimeric transcripts, of which 275 were found by both methods. In Mode 2, chimeras with TE-derived sequences smaller than the read length are not detected. However, Mode 2 is able to detect TEs that are absent from the reference genome, along with low-frequency TE insertions from individuals of the wild-type strains that are not present in the assembled Nanopore genome. Since low-frequency Nanopore reads are discarded during the genome assembly (52), Mode 1 is not able to detect them, whereas Mode 2 can. Furthermore, Mode 1 detects chimeric transcripts derived from TEs inside/near genes, but TEs generating chimeras might not be located inside/near genes. Such chimeras have been reported from TEs skipping one or more genes in *D. melanogaster* (31), as well as TEs acting as distal *cis*-regulatory elements (69, 70). Therefore, although we have considered as false positives all cases in which the TE was not found inside/near a gene in the manual curation, we speculate that part of these cases found

only by Mode 2 might be either from low-frequency TEs in the pool of individuals sequenced, or from TEs located far from genes.

Replicability and coverage of chimeric transcripts in both ChimeraTE modes

The ability to detect TE expression is an important factor to identify chimeric reads. The different strategies of alignment to identify chimeric reads in Mode 1 and Mode 2 cause differences in the sensitivity of chimeric transcript detection (Figure 6A). We investigated whether such differences might be associated with the detection of TE-derived reads. We found that Mode 1 is more efficient than Mode 2 to detect reads aligned against TE insertions (Sup. Figure 2), as well as for chimeric read detection (Figure 7A). Indeed, the proportion of chimeric reads in Mode 1 and Mode 2 is ~0.39% and ~0.04% of the total library sizes, respectively. The power of chimeric read detection from the two Modes is different because of the alignment strategies used. Despite the differences, both modes show significant positive correlations between the library size and the number of chimeric reads, as well as the number of TE-aligned reads and chimeric reads (Sup. Figure 3).

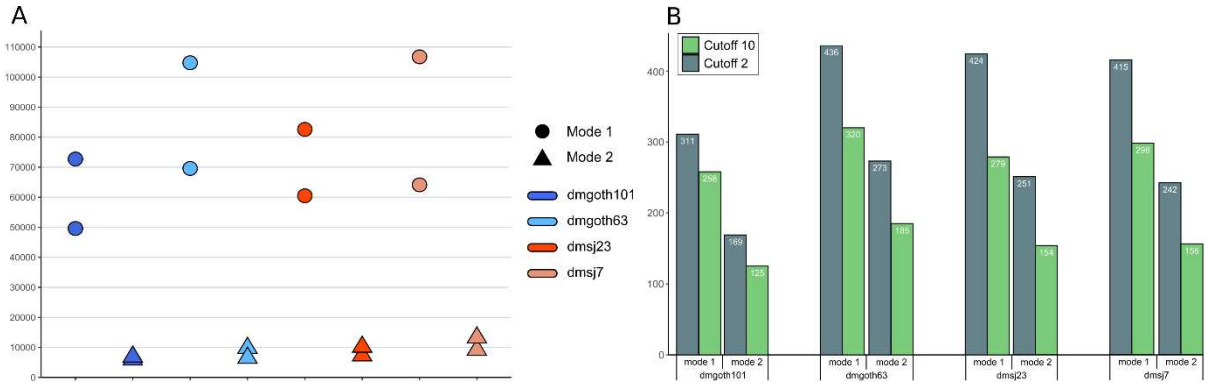

Figure 7: **A**) Total chimeric reads found in each RNA-seq replicate. **B**) Chimeric transcripts detected by both modes of ChimeraTE. The bars represent chimeras found in both RNA-seq replicates in the four *D. melanogaster* wild-type strains, applying two chimeric read cutoffs: 2 and 10.

To quantify ChimeraTE replicability between RNA-seq samples in both modes, as well as the impact of changes in the coverage thresholds, we performed a comparative analysis of chimeric transcripts using two thresholds of chimeric reads, 2 and 10. Overall, Mode 1 finds ~396 and ~288 genes with chimeric transcripts in both replicates (Figure 7B), using thresholds 2 and 10 respectively, whereas Mode 2 obtains ~233 and ~155. These results show that by increasing the chimeric read thresholds from 2 to 10, there is a decrease of 27.27% and 33.48% in the number of detected chimeric transcripts in Mode 1 and Mode 2. Therefore, even when found in both replicates, a substantial amount of chimeric transcripts is detected with low coverage of chimeric reads by both modes.

Artifactual chimeric reads

In both Mode 1 and Mode 2, only chimeric transcripts found in both RNA-seq replicates were considered true chimeras. Chimeric transcripts found in only one replicate may exist due to the lack of read coverage in one of the replicates to predict it, or they could be artifacts of PCR amplification during paired-end Illumina sequencing (35, 37). To quantify cases that may be artifacts, we aligned the RNA-

seq from the four wild-type strains against their genomes to identify paired-end reads where each mate maps to genes from different chromosomes, and therefore are artifacts. In replicate 1 and replicate 2, we found 55,152 and 38,484 cases representing fusions of genes that are in different chromosomes with at least one chimeric read. However, only 130 (0.27%) of these artifacts were found in both replicates with >= 2 chimeric reads (Sup. Table 14). In addition, 58.45% of these genes generating artifacts were highly expressed (FPKM > 100). Therefore, artifactual chimeric reads present in more than one replicate and derived from genes and TEs, might have a lower prevalence, since the proportion of RNA-seq reads derived from TEs is very low in comparison to genes. Thus, in order to reduce significantly false positive chimeric transcript calls, it is strongly recommended to use RNA-seq replicates, since artifactual chimeric reads exist in high frequency in highly expressed genes but only in one RNA-seq replicate. The proportion of the same chimeras found in more than one RNA-seq replicate, with >= 2 chimeric reads, is, on the contrary, very low.

Genes generating chimeric transcripts

Chimeric transcripts have been found in several species, being associated with genes holding different functions (12, 31, 32). In order to investigate whether genes generating chimeric transcripts in *D. melanogaster* have enriched biological processes, we have performed gene ontology enrichment analysis with a dataset of 566 genes, corresponding to all genes detected with Mode 1 (506 genes) and genes detected only with Mode 2, but confirmed with manual genomic curation (60 genes). We found 16 enriched ontology terms corresponding to 287 (50.7%) genes, associated with different functions, mainly transcription regulation and organ/tissue development (Fig. 8). More specifically, there are genes associated with tissue development and differentiation, such as imaginal disc-derived wing and leg morphogenesis, compound eye development, R3/R4 cell fate and regulation of organ growth. There are also ovary-specific biological processes such as ovary follicle cell development and regulation of border follicle cell migration. Finally, we found enrichment with two biological processes related to the nervous system: axon guidance and long-term memory. Altogether, these results indicate that genes producing chimeric transcripts in ovarian transcriptomes of *D. melanogaster* are associated with different functions. We then investigated whether specific biological processes may have specific TE families generating chimeric transcripts. Interestingly, for all biological processes, we found at least 75% of the chimeras involved a *roo* copy. For instance, all genes associated with positive regulation of border follicle cell migration (9 genes), and R3/R4 cell fate commitment (6 genes) have chimeras involving *roo*. Despite the global functional results obtained with enrichment analysis comprising 50,7% of all genes with chimeric transcripts, the presence of only the *roo* family (out of 128 families) in more than 75% of all genes from 16 biological processes is an unprecedented result.

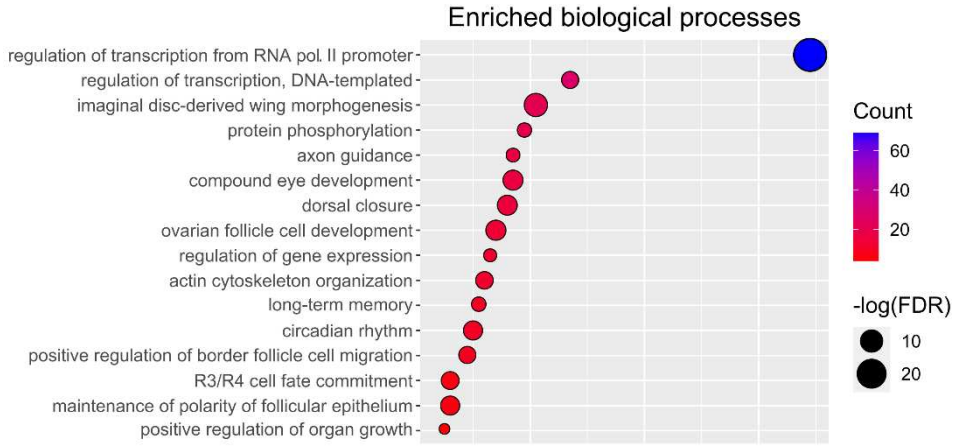

Figure 8: Gene ontology analysis with the enriched biological process with FDR < 0.05. The enrichment accounts with 16 biological processes from 287 genes, for which the most representative is “regulation of transcription from RNA pol. II promoter", with more than 60 genes. The other GO terms have several functions, for which "ovary follicle cell development" and "regulation of border follicle cell migration" are ovary-specific.

Roo element is the main TE family associated with chimeric transcripts

Several TE families have been associated with chimeric transcripts in *D. melanogaster* (31, 38). Here, we found 62 TE families producing at least one chimeric transcript in the four wild-type strains, representing 48.44% of all TE families annotated in the genomes. We then checked whether the frequency of TE families in chimeric transcripts is associated with their abundances in the *D. melanogaster* genomes. We found positive correlation between them (*Pearson; r*=0.39; $p < 2.2^{-16}$). In the TE-initiated upstream chimeras, we found *S2*, *accord2*, and *412* with the highest frequencies (Fig. 9). TE-initiated 5’ UTR is dominated by *roo* elements, representing up to 64.10% of all cases, followed by *HMS-Beagle* with 6.41%. In TE-exonized transcripts, the prevalence of *roo* elements is more pronounced, comprising more than 80% of all chimeras, followed by *INE-1* elements, with 1.90%. TE-terminated 3’ UTR transcripts also have more *roo* elements in comparison to other families, but in these chimeras, the frequency of *roo* is 33.33%, followed by *INE-1* elements with 32.50% and *1360* with 9.16%. Finally, in TE-terminated downstream chimeras, *INE-1* is the most frequent TE, with 28.35% of all cases, followed by *1360* and *roo*, with 10.44% and 4.47%, respectively. These results show that *roo* elements are the most frequent TE family involved in chimeric transcripts in *D. melanogaster*, in agreement with results obtained in the enrichment analysis. Nevertheless, this high frequency is only observed in chimeric transcripts that have TE insertions inside genes, such as TE-initiated 5' UTR, TE-exonized, and TE-terminated 3' UTR. We then investigated the correlation between the prevalence of TE insertions inside genes and the frequency of chimeric transcripts for each TE family, and we found a positive correlation (*Pearson*; *r=0.65*; $p < 2.2^{-16}$).

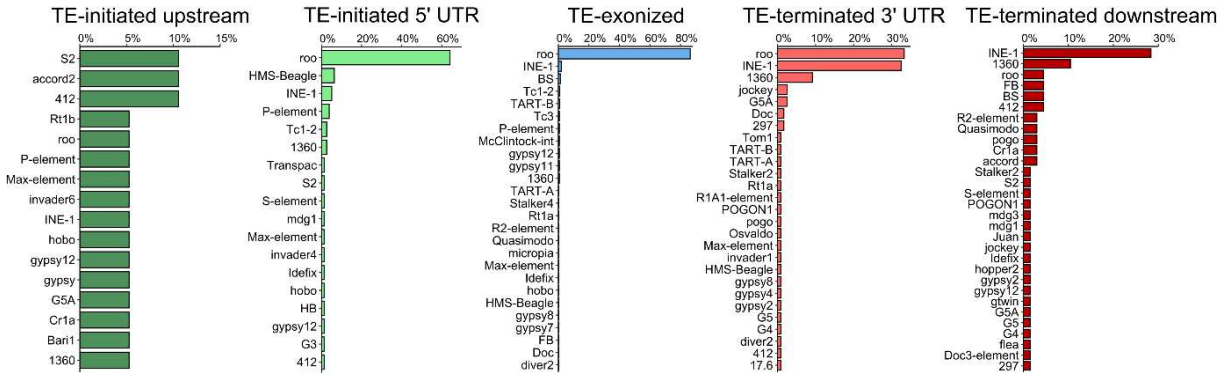

Figure 9: The frequency of the 64 TE families generating chimeric transcripts in the four wild-type strains. In chimeric transcripts derived from TEs near genes, *INE-1* was the most frequent (~30%) in TE-terminated downstream, whereas for TE-initiated upstream, *S2, accord2,* and *412* had the same frequency (~10%). Regarding TEs inside genes, *the roo* element has the highest frequency, with ~60%, ~80%, and ~30% to TE initiated 5' UTR, TE-exonized, and TE-terminated 3' UTR, respectively.

TE insertions disrupting coding regions and promoters are more likely to be deleterious, being consequently eliminated by purifying selection (71), although several exaptation events have been documented (24, 72). The presence of *roo* elements in more than 80% of all TE-exonized transcripts suggests the neutral or adaptive role of this family when incorporated into gene transcripts. *Roo* is the most abundant euchromatic TE family of *D. melanogaster* (1, 73, 74). *Roo* insertions have been associated with modifications in the expression level of stress response genes due to the presence of TFBSs (75). They have also low enrichment of repressive histone marks (76, 77), potentially explaining their high transposition rates (78, 79). *Roo* is an LTR retrotransposon, encoding three proteins: *gag*, *pol*, and *env*, that have been through domestication events from retroelements in many species, including *Drosophila* (8, 80, 81). We then investigated whether exonized *roo* insertions could contribute to protein domains. Through our search with CDD (58), we did not find any protein domain encoded by the chimeric *roo* elements. The length of these insertions is ~110 bp, indicating that they are old insertions since the full-length consensus sequence is 9,250 bp (Figure 10A). Subsequently, we analyzed whether these exonized *roo* insertions are donors of preferential motifs to the chimeric transcripts. All exonized *roo* sequences stem from a specific region, between the 5' UTR and the beginning of the *roo* open-reading frame (Figure 10B). Despite the low nucleotide diversity of *roo* insertions in *the D. melanogaster* genome, the 5' UTR region has a hypervariable region, including deletions and repeats, with several copies missing a tandem repeat of 99 bp (82, 83). It has been proposed that this region may have a role in *roo* transposition, by heterochromatinization, recruitment of RNA pol II, and interaction with other enzymes (83). Curiously, a study assessed the nucleotide diversity between *roo* insertions, and characterized the same region as a deletion hotspot (84). This is the first time that this region has been observed as part of TE-exonized transcripts in a transcriptome-wide manner in *D. melanogaster*. In order to evaluate whether these sequences identified as *roo* insertions were not widespread repeats across the genomes, we performed a blastn search. We found that 97.45% of these exonized *roo* insertions have only one hit. Across the four wild-type genomes, only 20 insertions had mappings at multiple loci, with an average of ~6 matches (Sup. Figure 4). Why this

region has been maintained through evolution, and whether it could be adaptive or neutral, remains unclear.

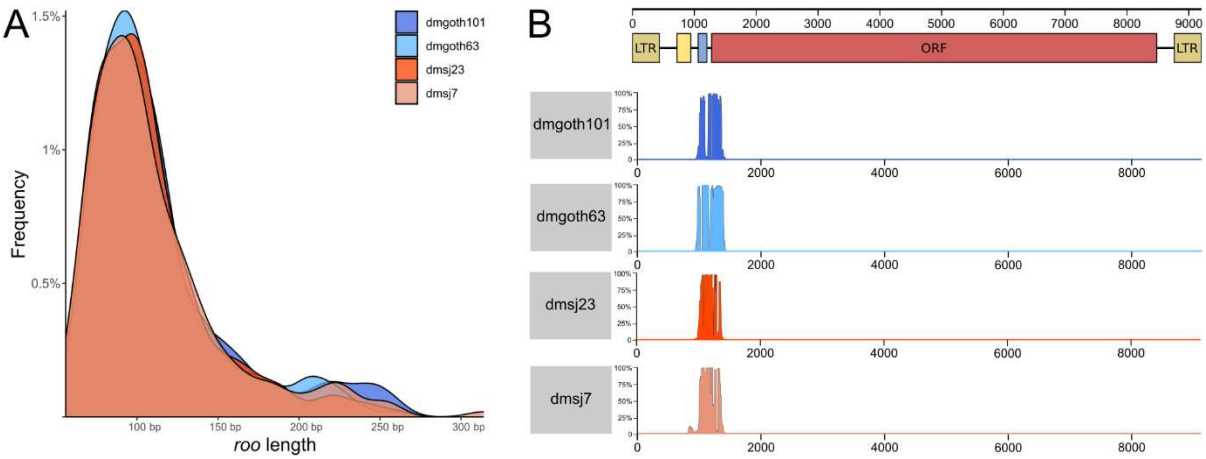

Figure 10: **A**) Length of exonized *roo* elements in the four wild type-strains. **B**) Alignment depth of exonized *roo* insertions with *roo* consensus. At the top, the scheme of the *roo* element: brown boxes: LTRs; yellow box: first tandem repeats at 5' UTR; blue box: second tandem repeat at 5' UTR; red box: Open reading frame (ORF). The coverage depth of the multiple alignments between exonized *roo* insertions and the consensus is separated by strain.

Polymorphic TE insertions in the wild-type strains generate 45 chimeric transcripts

Polymorphic TE insertions are common across *D. melanogaster* strains (67, 85). In order to quantify how many chimeric transcripts in these strains are derived from polymorphic TE insertions, we used the list of genes with TEs located 3 Kb upstream, inside (introns and exons), and 3 Kb downstream generated previously. These genes were selected as potential sources of chimeras in the *dm6* genome and then compared with the list of chimeric transcripts generated by ChimeraTE in the four wild-type strains. We found that 45 genes with chimeric transcripts in the wild-derived strains were generated by TE insertions that are absent in the reference genome (Sup. Table 15). There are 29 TE families generating polymorphic chimeric transcripts: *roo* (36.86%), *P-element* (7.89%), *412* (7.89%), *HMS-Beagle* (5.26%), *INE-1* (3.94%), among others. Except for *INE-1*, all these TE families are known to be active in *D. melanogaster* (82). *INE-1* chimeras are unexpected since it is an old TE family in *D. melanogaster* (86), however, *INE-1* polymorphism among *D. melanogaster* populations has previously been shown (87).

The study of TEs in wild-type strains offers new insights into their ability to provide genetic variability. Depending upon the position where TEs are inserted regarding genes, they can contribute to gene expression or protein sequence variation. We then compared whether polymorphic TE insertions generating chimeric transcripts are more likely inserted near or inside genes. We found that 6.83% correspond to TEs located upstream, 69.23% inside (intron/exons), and 23.93% downstream (Figure 11). In addition, we investigated whether the TE insertions absent from the reference genome could either be population-specific or strain-specific. For TEs located upstream, only the *accord2* insertion upstream to the *CG42694* gene has been found in both French strains, dmgoth101, and dmgoth63, suggesting that it might be a population-specific (Fig. 11A). All the other chimeric transcripts from TE-

initiated cases are strain-specific. For TEs inserted inside genes (Fig. 11B), there are five chimeric transcripts common to all strains (*CG7239-roo*; *CG46385-HMSBeagle*; *CG10077-roo*; *CG3164-McClintock* and *Camta-roo*). Then, five chimeric transcripts were found in dmgoth-specific strains *simj*; *scb*; *CG9518*; *B4,* and *Cyp6a2*. The latter is a gene from the P450 family, related to insecticide metabolism (88). Two chimeric transcripts were found only in dmsj23 and dmsj7 strains: *VhaM9.7b-G5A* and *Nelf-A-roo*. Finally, for TEs located downstream genes (Fig. 11C), we found that *Mps1-FB* is the only chimera found in French strains, whereas dmsj23 and dmsj7 have seven chimeric transcripts, including one lncRNA gene (*CR46064-Juan*). Taken together, these results reinforce the potential of TEs in generating genetic novelty between *D. melanogaster* strains.

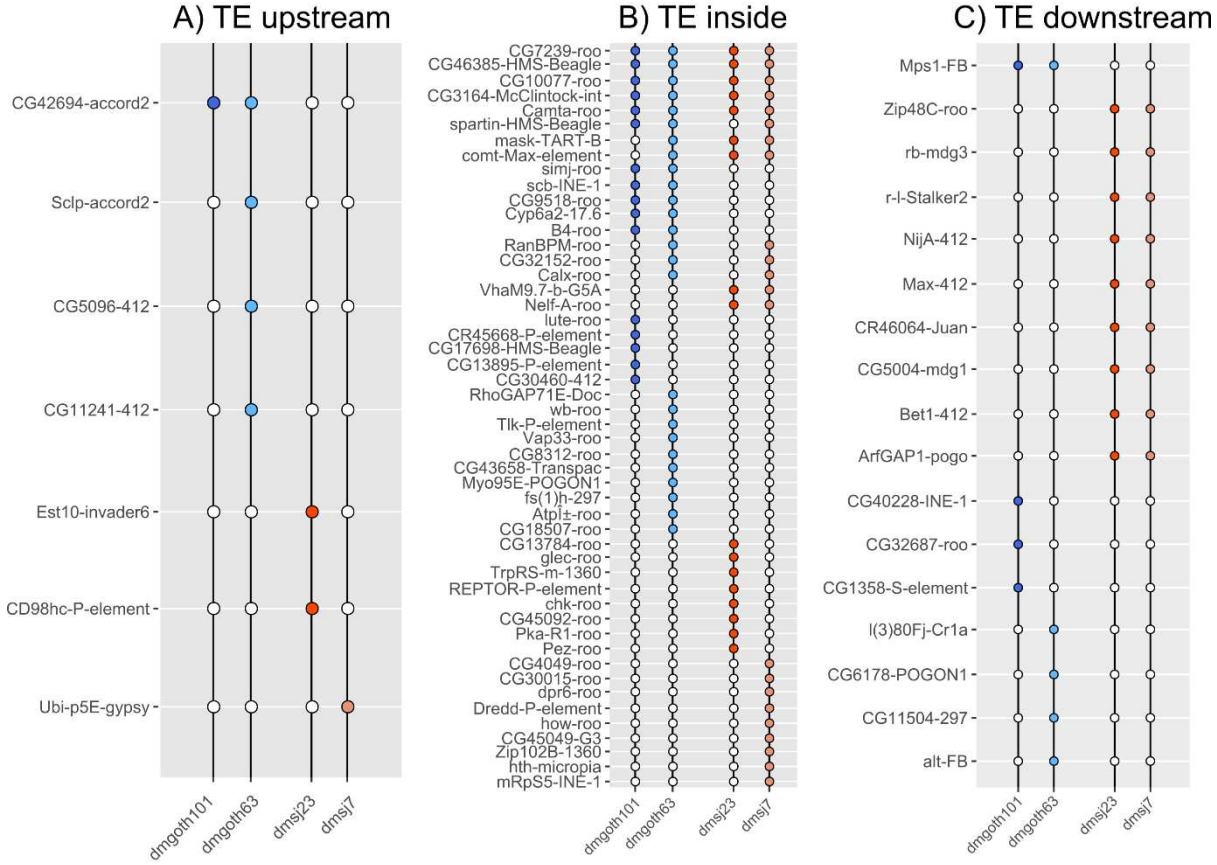

Figure 11: The 45 chimeric transcripts derived from TE insertions that are absent in the *dm6* reference genome. 6.83% of them correspond to TEs located upstream, 69.23% to TEs located inside genes (introns and exons), and 23.93% to TEs located downstream. **A**) TE upstream: Chimeric transcripts in which the TE is located up to 3kb upstream of the gene. Only the *CG42694-accord2* was found in both strains of the French strains, all the other are strain-specific. **B**) TE inside: Chimeric transcripts with TE insertions located inside the gene region (exons and introns). There are five chimeric transcripts found in all strains; five specific for French strains; and two specific for Brazilian strains. **C**) TE downstream: Chimeric transcripts in which the TE is located up to 3kb downstream of the gene. Only *Mps1-FB* was specific for French strains, whereas Brazilian strains have seven population-specific chimeras.

Chimeric transcripts do not change gene expression levels

TEs are able to modify gene expression when inserted inside and in their close vicinity. We hypothesized that genes with strain-specific chimeric transcripts may have differences in their

expression levels between French and Brazilian strains. However, we found that most genes have a similar expression level regardless of the presence of chimeric transcripts (Figure 12). The comparison between dmgoth63 and dmgoth101 reveals that only four genes have FPKM > 1 in one strain and FPKM < 1 in the other. We checked if the TE insertions generating such chimeras are unique from dmgoth63 or whether they are present in dmgoth101, and we found all of them in the dmgoth101 genome. All these cases were chimeric transcripts from old insertions located inside exons, except *l(2)41Ab* which harbors an exonized *Max* insertion located in the first intron. We performed the same analysis for *CG12581*, which has FPKM > 1 in dmgoth101 and FPKM < 1 in dmgoth63 and we found that it is also a common TE insertion, suggesting that when this gene is expressed, the chimeric isoform may also exist in dmgoth63. The results were very similar to Brazilian strains. From three genes that are expressed only in dmsj7 (*Cyp6a14*, *fuss*, and *Zasp52*), we found common TE insertions in comparison to dmsj23. Regarding genes expressed in dmsj23 and not in dmsj7, we found that *Spt20*, *CG17816*, *Smr*, *l(2)41Ab*, have common TEs, whereas *nej* and *CG8665* are exclusive of dmsj23. Therefore, by comparing the expression level of genes with chimeric transcripts in French and Brazilian strains, we found that only 4.92% of strain-specific chimeric transcripts are due to low expression levels in one of the strains.

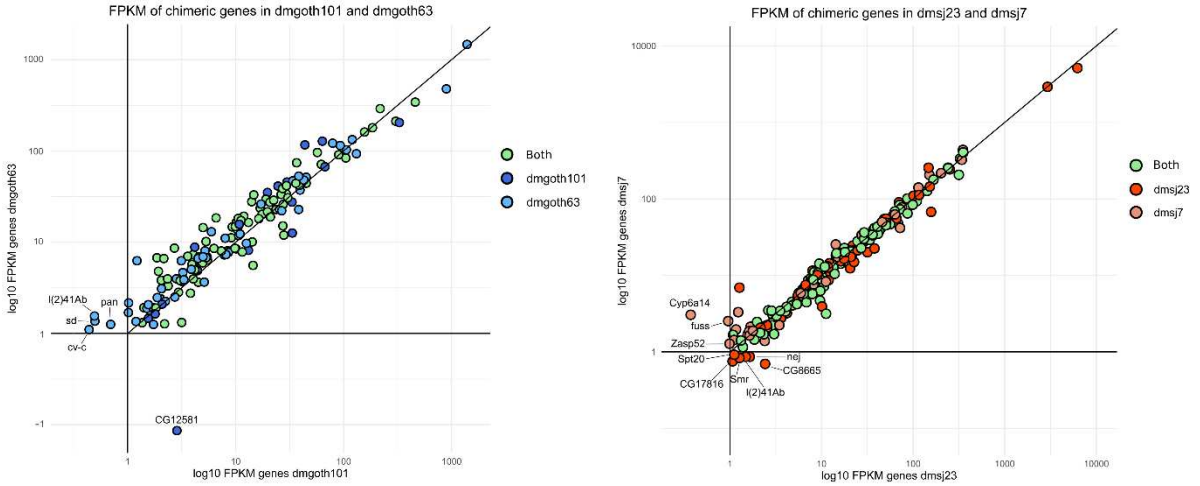

Figure 12: Expression level (FPKM) of genes that have chimeric transcripts in the French (dmgoth101, dmgoth63) and Brazilian strains (dmsj23, dmsj7). “Both” represents genes that have chimeric transcripts in both strains. Genes with the name indicated in the graph are regarding cases in which it is expressed (FPKM > 1) in only one strain.

CONCLUSION

In the last decades, RNA-seq has provided the opportunity to understand transcriptome plasticity, which can lead to phenotypic divergence, from related species or individuals from a population (89). Among several sources of modification in gene expression and novel isoform transcripts, TEs have been considered as fundamental suppliers of transcriptome plasticity, participating in gene networks and incorporating either regulatory sequences or protein domains into gene transcripts. The identification of chimeric transcripts is an important step to understanding transcriptome plasticity, since they may be triggered by ectopic conditions, such as cancer, oxidative stress, and heat shock (27, 90, 91), which may lead to both detrimental and advantageous outcomes (92). Therefore, uncovering the extent of

chimeric transcripts between individual/cell/strain transcriptomes is a crucial first step to investigate potential exaptation/domestication events, or gene disruption and loss of function.

Chimeric transcripts have been identified more recently by different methods exploiting RNA-seq data (32–34), but none of them provided the possibility to predict chimeras from TEs that are absent from reference genomes. Here, we developed ChimeraTE, a pipeline able to identify chimeric transcripts from TEs that are absent from the reference genome. The Mode 1 is the genome-guided approach and may be used either when the user is not interested in chimeras derived from TEs absent from the reference genome, or when the user has a high-quality genome assembly for each individual/cell/strain; whereas Mode 2 is the genome-blind approach, with the ability to predict chimeric transcripts without the assembled genome, but missing chimeras where the TE is smaller than the length of the reads.

We analyzed ovarian RNA-seq from four *D. melanogaster* wild-type strains, for which we have genome assemblies. Altogether, we found that ~3% of all genes are generating chimeric transcripts in ovaries, increasing the previously proposed abundance obtained with ESTs, and RNA-seq in midbrain tissue (31, 38). Furthermore, our results revealed that ~80% of all TE-exonized transcripts derive from *roo* elements, and most specifically, a small region between tandem repeats in the 5' UTR and the beginning of the *roo* ORF. These results suggest that these *roo* insertions have neutral or advantageous effects as they are maintained within these gene transcripts. However, we did not provide enough support to claim these *roo*-exonized transcripts as exaptation or domestication events, due to the lack of evidence regarding the functional role of these chimeras. Further studies must be performed to clarify this subject, mainly because chimeric transcripts detected by ChimeraTE can be degraded by surveillance pathways that degrade aberrant mRNA, such as non-sense-mediated mRNA decay (93), non-go decay (94), and non-stop decay (95).

Altogether, this new approach allows studying the impact of new mobilization events between populations or between treatment conditions, providing insights into biological questions from a broad community of researchers, ranging from cancer research, population transcriptomics, and adaptation studies. ChimeraTE implementation will be useful in the next discoveries regarding the evolutionary role of TEs and their impact on the host transcriptome.

AVAILABILITY

ChimeraTE is an open-source collaborative initiative available in the GitHub repository (https://github.com/OliveiraDS-hub/ChimeraTE).

ACCESSION NUMBERS

The RNA-seq data used in this study are available in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/), under PRJNA795668 accession.

ACKNOWLEDGEMENT

We thank Josefa González and Marta Coronado-Zamora for useful discussions and advices. This work was performed using the computing facilities of the CC LBBE/PRABI.

FUNDING

This work was supported by Agence Nationale de la Recherche [Exhyb ANR-14-CE19-0016-01 to C.V], Fondation pour la Recherche Médicale [DEP20131128536 to C.V.]; Idex Lyon fellowship to D.S.O., Campus France Eiffel [P769649C to D.S.O.], TIGER [H2020-MSCA-IF-2014- 658726 to R.R.]; National Council for Scientific and Technological Development [308020/2021-9 to C.M.A.C.]; and São Paulo Research Foundation [2020/06238-2 to C.M.A.C].

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTAL INFORMATION

Figure S1: Gene and TE annotation from *dm6* genome for chimeric transcripts identified by Mode 2 corresponding to genes that were not annotated in the wild-type genome assemblies.

Figure S2: The total number of TE-aligned reads between both ChimeraTE Modes, in all strains and their respective replicates.

Figure S3: Positive Person correlations between total chimeric reads and RNA-seq library size; and TE-aligned reads, in both Mode 1 and Mode 2.

Figure S4: The total hits of TE-exonized *roo* elements across their respective wild-type strain genomes.

Table S1: Gene and TE annotation in the four wild-type strains and dm6 genome

Table S2-5: Chimeric transcripts detected by Mode 1 in dmgoth101, dmgoth63, dmsj23 and dmsj7 strains.

Table S6: 17 chimeric transcripts found by Lipatov et al., 2005 and Kmn1-pogo chimera found by Mateo et al., 2014.

Table S7: Common chimeric transcripts found by Treibber & Waddell 2020 and our study.

Table S8-11: Chimeric transcripts detected by Mode 2 in dmgoth101, dmgoth63, dmsj23, and dmsj7 strains.

Table S12: Running time of ChimeraTE Mode 1 and Mode 2, without and with --assembly option, in the four D. melanogaster wild-type strains.

Table S13: Chimeric transcripts found by Mode 2, but not by Mode 1, due to lack of genomic annotation.

Table S14: Number of fusion genes from different chromosomes with chimeric reads support.

Table S15: TE insertions absent from the dm6 reference genome generating chimeric transcripts in the wild-type strains.

REFERENCES

1. Quesneville,H., Bergman,C.M., Andrieu,O., Autard,D., Nouaud,D., Ashburner,M. and Anxolabehere,D. (2005) Combined Evidence Annotation of Transposable Elements in Genome Sequences. *PLoS Comp Biol*, **1**, e22.

2. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

3. Schnable,P.S., Ware,D., Fulton,R.S., Stein,J.C., Wei,F., Pasternak,S., Liang,C., Zhang,J., Fulton,L., Graves,T.A., *et al.* (2009) The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science*, **326**, 1112–1115.

4. Sotero-Caio,C.G., Platt,R.N., Suh,A. and Ray,D.A. (2017) Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biology and Evolution*, **9**, 161–177.

5. Danilevskaya,O.N., Arkhipova,I.R., Pardue,M.L. and Traverse,K.L. (1997) Promoting in Tandem: The Promoter for Telomere Transposon HeT-A and Implications for the Evolution of Retroviral LTRs. *Cell*, **88**, 647–655.

6. Kapitonov,V.V. and Jurka,J. (2005) RAG1 Core and V(D)J Recombination Signal Sequences Were Derived from Transib Transposons. *PLoS Biol*, **3**, e181.

7. Kapitonov,V.V. and Koonin,E.V. (2015) Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. *Biol Direct*, **10**, 20.

8. Volff,J.-N. (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays*, **28**, 913–922.

9. Babaian,A., Romanish,M.T., Gagnier,L., Kuo,L.Y., Karimi,M.M., Steidl,C. and Mager,D.L. (2016) Onco-exaptation of an endogenous retroviral LTR drives IRF5 expression in Hodgkin lymphoma. *Oncogene*, **35**, 2542–2546.

10. Daborn,P.J., Yen,J.L., Bogwitz,M.R., Le Goff,G., Feil,E., Jeffers,S., Tijet,N., Perry,T., Heckel,D., Batterham,P., *et al.* (2002) A Single P450 Allele Associated with Insecticide Resistance in *Drosophila*. *Science*, **297**, 2253–2256.

11. Jordan,I.K., Rogozin,I.B., Glazko,G.V. and Koonin,E.V. (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends in Genetics*, **19**, 68–72.

12. Mateo,L., Ullastres,A. and González,J. (2014) A Transposable Element Insertion Confers Xenobiotic Resistance in Drosophila. *PLoS Genet*, **10**, e1004560.

13. Modzelewski,A.J., Shao,W., Chen,J., Lee,A., Qi,X., Noon,M., Tjokro,K., Sales,G., Biton,A., Anand,A., *et al.* (2021) A mouse-specific retrotransposon drives a conserved Cdk2ap1 isoform essential for development. *Cell*, **184**, 5541-5558.e22.

14. Brosius,J. and Gould,S.J. (1992) On 'genomenclature': a comprehensive (and respectful) taxonomy for pseudogenes and other 'junk DNA'. *Proceedings of the National Academy of Sciences*, **89**, 10706–10710.

15. Fueyo,R., Judd,J., Feschotte,C. and Wysocka,J. (2022) Roles of transposable elements in the regulation of mammalian transcription. *Nat Rev Mol Cell Biol*, 10.1038/s41580-022-00457-y.

16. Lamprecht,B., Walter,K., Kreher,S., Kumar,R., Hummel,M., Lenze,D., Köchert,K., Bouhlel,M.A., Richter,J., Soler,E., *et al.* (2010) Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nat Med*, **16**, 571–579.

17. McGinnis,W., Shermoen,A.W. and Beckendorf,S.K. (1983) A transposable element inserted just 5′ to a Drosophila glue protein gene alters gene expression and chromatin structure. *Cell*, **34**, 75–84.

18. Almeida,L.M., Amaral,M.E.J., Silva,I.T., Silva Jr,W.A., Riggs,P.K. and Carareto,C.M. (2008) Report of a chimeric origin of transposable elements in a bovine-coding gene. *Genet. Mol. Res.*, **7**, 107–116.

19. Sela,N., Mersch,B., Hotz-Wagenblatt,A. and Ast,G. (2010) Characteristics of Transposable Element Exonization within Human and Mouse. *PLoS ONE*, **5**, e10907.

20. Sorek,R. (2007) The birth of new exons: Mechanisms and evolutionary consequences. *RNA*, **13**, 1603–1608.

21. Bogwitz,M.R., Chung,H., Magoc,L., Rigby,S., Wong,W., O'Keefe,M., McKenzie,J.A., Batterham,P. and Daborn,P.J. (2005) Cyp12a4 confers lufenuron resistance in a natural population of Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, **102**, 12807–12812.

22. Farré,D., Engel,P. and Angulo,A. (2016) Novel Role of 3'UTR-Embedded Alu Elements as Facilitators of Processed Pseudogene Genesis and Host Gene Capture by Viral Genomes. *PLoS ONE*, **11**, e0169196.

23. Capy,P. (2021) Taming, Domestication and Exaptation: Trajectories of Transposable Elements in Genomes. *Cells*, **10**, 3590.

24. Magwire,M.M., Bayer,F., Webster,C.L., Cao,C. and Jiggins,F.M. (2011) Successive Increases in the Resistance of Drosophila to Viral Infection through a Transposon Insertion Followed by a Duplication. *PLoS Genet*, **7**, e1002337.

25. Cordaux,R., Udit,S., Batzer,M.A. and Feschotte,C. (2006) Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 8101–8106.

26. Ehrlich,M. (2009) DNA hypomethylation in cancer cells. *Epigenomics*, **1**, 239–259.

27. Babaian,A. and Mager,D.L. (2016) Endogenous retroviral promoter exaptation in human cancer. *Mobile DNA*, **7**, 24.

28. Lock,F.E., Rebollo,R., Miceli-Royer,K., Gagnier,L., Kuah,S., Babaian,A., Sistiaga-Poveda,M., Lai,C.B., Nemirovsky,O., Serrano,I., *et al.* (2014) Distinct isoform of FABP7 revealed by screening for retroelement-activated genes in diffuse large B-cell lymphoma. *Proceedings of the National Academy of Sciences*, **111**, E3534–E3543.

29. Faulkner,G.J., Kimura,Y., Daub,C.O., Wani,S., Plessy,C., Irvine,K.M., Schroder,K., Cloonan,N., Steptoe,A.L., Lassmann,T., *et al.* (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics*, **41**, 563–571.

30. Babarinde,I.A., Ma,G., Li,Y., Deng,B., Luo,Z., Liu,H., Abdul,M.M., Ward,C., Chen,M., Fu,X., *et al.* (2021) Transposable element sequence fragments incorporated into coding and noncoding transcripts modulate the transcriptome of human pluripotent stem cells. *Nucleic Acids Research*, **49**, 9132–9153.

31. Lipatov,M., Lenkov,K., Petrov,D.A. and Bergman,C.M. (2005) Paucity of chimeric gene-transposable element transcripts in the Drosophila melanogaster genome. *BMC Biol*, **3**, 24.

32. Treiber,C.D. and Waddell,S. (2020) Transposon expression in the *Drosophila* brain is driven by neighboring genes and diversifies the neural transcriptome. *Genome Res.*, **30**, 1559–1569.

33. Pinson,M.E., Pogorelcnik,R., Court,F., Arnaud,P. and Vaurs-Barrière,C. (2018) CLIFinder: Identification of LINE-1 chimeric transcripts in RNA-seq data. *Bioinformatics*, **34**, 688–690.

34. Babaian,A., Thompson,I.R., Lever,J., Gagnier,L., Karimi,M.M. and Mager,D.L. (2019) LIONS: Analysis suite for detecting and quantifying transposable element initiated transcription from RNA-seq. *Bioinformatics*, **35**, 3839–3841.

35. Kircher,M., Sawyer,S. and Meyer,M. (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, **40**, e3–e3.

36. Evrony,G.D., Lee,E., Park,P.J. and Walsh,C.A. (2016) Resolving rates of mutation in the brain using single-neuron genomics. *eLife*, **5**, e12966.

37. Quail,M.A., Kozarewa,I., Smith,F., Scally,A., Stephens,P.J., Durbin,R., Swerdlow,H. and Turner,D.J. (2008) A large genome center's improvements to the Illumina sequencing system. *Nat Methods*, **5**, 1005–1010.

38. Treiber,C.D. and Waddell,S. (2017) Resolving the prevalence of somatic transposition in Drosophila. *eLife*, **6**, e28297.

39. Martin Cerezo,M.L., Raval,R., de Haro Reyes,B., Kucka,M., Chan,F.Y. and Bryk,J. (2022) Identification and quantification of chimeric sequencing reads in a highly multiplexed RAD -seq protocol. *Molecular Ecology Resources*, 10.1111/1755-0998.13661.

40. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods*, **9**, 357–359.

41. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, **7**, 562–578.

42. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R., and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

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

44. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods*, **9**, 357–359.

45. Roberts,A. and Pachter,L. (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods*, **10**, 71–73.

46. Grabherr,M.G., Haas,B.J., Yassour,M., Levin,J.Z., Thompson,D.A., Amit,I., Adiconis,X., Fan,L., Raychowdhury,R., Zeng,Q., *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*, **29**, 644–652.

47. SMIT, Arian FA (2004) RepeatMasker Open 3.0.

48. Storer,J., Hubley,R., Rosen,J., Wheeler,T.J. and Smit,A.F. (2021) The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*, **12**, 2.

49. Altschul,S. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

50. Fablet,M., Salcez-Ortiz,J., Jacquet,A., Menezes,B.F., Dechaud,C., Veber,P., Noûs,C., Rebollo,R. and Vieira,C. (2022) A quantitative, genome-wide analysis in Drosophila reveals transposable elements' influence on gene expression is species-specific. *BioRxiv*, https://doi.org/10.1101/2022.01.20.477049.

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

52. Mohamed,M., Dang,N.T.-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**, 1776.

53. Shumate,A. and Salzberg,S.L. (2021) Liftoff: accurate mapping of gene annotations. *Bioinformatics*, **37**, 1639–1643.

54. Bailly-Bechet,M., Haudry,A. and Lerat,E. (2014) "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. *Mobile DNA*, **5**, 13.

55. dos Santos,G., Schroeder,A.J., Goodman,J.L., Strelets,V.B., Crosby,M.A., Thurmond,J., Emmert,D.B., Gelbart,W.M., and the FlyBase Consortium (2015) FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Research*, **43**, D690–D697.

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

57. Sherman,B.T., Hao,M., Qiu,J., Jiao,X., Baseler,M.W., Lane,H.C., Imamichi,T. and Chang,W. (2022) DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Research*, **50**, W216–W221.

58. Lu,S., Wang,J., Chitsaz,F., Derbyshire,M.K., Geer,R.C., Gonzales,N.R., Gwadz,M., Hurwitz,D.I., Marchler,G.H., Song,J.S., *et al.* (2020) CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Research*, **48**, D265–D268.

59. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

60. Kumar,S., Stecher,G., Li,M., Knyaz,C. and Tamura,K. (2018) MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*, **35**, 1547–1549.

61. Larsson,A. (2014) AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, **30**, 3276–3278.

62. Zea,D.J., Anfossi,D., Nielsen,M. and Marino-Buslje,C. (2016) MIToS.jl: mutual information tools for protein sequence analysis in the Julia language. *Bioinformatics*, 10.1093/bioinformatics/btw646.

63. Fablet,M., Salcez-Ortiz,J., Jacquet,A., Menezes,B.F., Dechaud,C., Veber,P., Noûs,C., Rebollo,R. and Vieira,C. (2022) A quantitative, genome-wide analysis in Drosophila reveals transposable elements' influence on gene expression is species-specific. *bioRxiv*.

64. Kapun,M., Barrón,M.G., Staubach,F., Obbard,D.J., Wiberg,R.A.W., Vieira,J., Goubert,C., Rota-Stabelli,O., Kankare,M., Bogaerts-Márquez,M., *et al.* (2020) Genomic Analysis of European Drosophila melanogaster Populations Reveals Longitudinal Structure, Continent-Wide Selection, and Previously Unknown DNA Viruses. *Molecular Biology and Evolution*, **37**, 2661–2678.

65. Rech,G.E., Bogaerts-Márquez,M., Barrón,M.G., Merenciano,M., Villanueva-Cañas,J.L., Horváth,V., Fiston-Lavier,A.-S., Luyten,I., Venkataram,S., Quesneville,H., *et al.* (2019) Stress response, behavior, and development are shaped by transposable element-induced mutations in Drosophila. *PLoS Genet*, **15**, e1007900.

66. Rech,G.E., Radío,S., Guirao-Rico,S., Aguilera,L., Horvath,V., Green,L., Lindstadt,H., Jamilloux,V., Quesneville,H. and González,J. (2022) Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in Drosophila. *Nat Commun*, **13**, 1948.

67. Vieira,C., Lepetit,D., Dumont,S. and Biemont,C. (1999) Wake up of transposable elements following Drosophila simulans worldwide colonization. *Molecular Biology and Evolution*, **16**, 1251–1255.

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

69. Lu,Z., Marand,A.P., Ricci,W.A., Ethridge,C.L., Zhang,X. and Schmitz,R.J. (2019) The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nat. Plants*, **5**, 1250–1259.

70. Bakoulis,S., Krautz,R., Alcaraz,N., Salvatore,M. and Andersson,R. (2022) Endogenous retroviruses co-opted as divergently transcribed regulatory elements shape the regulatory landscape of embryonic stem cells. *Nucleic Acids Research*, **50**, 2111–2127.

71. Cridland,J.M., Macdonald,S.J., Long,A.D. and Thornton,K.R. (2013) Abundance and Distribution of Transposable Elements in Two Drosophila QTL Mapping Resources. *Molecular Biology and Evolution*, **30**, 2311–2327.

72. Lerman,D.N. and Feder,M.E. (2005) Naturally Occurring Transposable Elements Disrupt hsp70 Promoter Function in Drosophila melanogaster. *Molecular Biology and Evolution*, **22**, 776–783.

73. Rahman,R., Chirn,G., Kanodia,A., Sytnikova,Y.A., Bergman,C.M. and Lau,N.C. (2015) Unique transposon landscapes are pervasive across Drosophila melanogaster genomes. *Nucleic Acids Research*, **43**, 10655–10672.

74. Vieira,C. and Biemont,C. (2004) Transposable element dynamics in two sibling species: Drosophila melanogaster and Drosophila simulans. *Genetica*, **120**, 115–123.

75. Merenciano,M., Ullastres,A., de Cara,M.A.R., Barrón,M.G. and González,J. (2016) Multiple Independent Retroelement Insertions in the Promoter of a Stress Response Gene Have Variable Molecular and Functional Effects in Drosophila. *PLoS Genet*, **12**, e1006249.

76. Rebollo,R., Horard,B., Begeot,F., Delattre,M., Gilson,E. and Vieira,C. (2012) A Snapshot of Histone Modifications within Transposable Elements in Drosophila Wild Type Strains. *PLoS ONE*, **7**, e44253.

77. Yasuhara,J.C. and Wakimoto,B.T. (2008) Molecular Landscape of Modified Histones in Drosophila Heterochromatic Genes and Euchromatin-Heterochromatin Transition Zones. *PLoS Genet*, **4**, e16.

78. Díaz-González,J., Vázquez,J.F., Albornoz,J. and Domínguez,A. (2011) Long-term evolution of the *roo* transposable element copy number in mutation accumulation lines of *Drosophila melanogaster*. *Genet. Res.*, **93**, 181–187.

79. Díaz-González,J., Domínguez,A. and Albornoz,J. (2010) Genomic distribution of retrotransposons 297, 1731, copia, mdg1 and roo in the Drosophila melanogaster species subgroup. *Genetica*, **138**, 579–586.

80. Nefedova,L.N., Kuzmin,I.V., Makhnovskii,P.A. and Kim,A.I. (2014) Domesticated retroviral GAG gene in Drosophila: New functions for an old gene. *Virology*, **450–451**, 196–204.

81. Malik,H.S. and Henikoff,S. (2005) Positive Selection of Iris, a Retroviral Envelope–Derived Host Gene in Drosophila melanogaster. *PLoS Genet*, **1**, e44.

82. Lerat,E., Rizzon,C. and Biémont,C. (2003) Sequence Divergence Within Transposable Element Families in the *Drosophila melanogaster* Genome. *Genome Res.*, **13**, 1889–1896.

83. Díaz-González,J. and Domínguez,A. (2020) Different structural variants of roo retrotransposon are active in Drosophila melanogaster. *Gene*, **741**, 144546.

84. Kaminker,J.S., Bergman,C.M., Kronmiller,B., Svirskas,R., Patel,S., Frise,E., Lewis,S.E., Rubin,G.M., Ashburner,M. and Celniker,S.E. (2002) The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biology*, **3**, 1–20.

85. Barrón,M.G., Fiston-Lavier,A.-S., Petrov,D.A. and González,J. (2014) Population Genomics of Transposable Elements in *Drosophila*. *Annu. Rev. Genet.*, **48**, 561–581.

86. Kapitonov,V.V. and Jurka,J. (2003) Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 6569–6574.

87. Kofler,R., Betancourt,A.J. and Schlötterer,C. (2012) Sequencing of Pooled DNA Samples (Pool-Seq) Uncovers Complex Dynamics of Transposable Element Insertions in Drosophila melanogaster. *PLoS Genet*, **8**, e1002487.

88. Dunkov,B.C., Guzov,V.M., Mocelin,G., Shotkoski,F., Brun,A., Amichot,M., Ffrench-Constant,R.H. and Feyereisen,R. (1997) The *Drosophila* Cytochrome P450 Gene *Cyp6a2*: Structure, Localization, Heterologous Expression, and Induction by Phenobarbital. *DNA and Cell Biology*, **16**, 1345–1356.

89. Marguerat,S. and Bähler,J. (2010) RNA-seq: from technology to biology. *Cell. Mol. Life Sci.*, **67**, 569–579.

90. Oliveira,D.S., Rosa,M.T., Vieira,C. and Loreto,E.L.S. (2021) Oxidative and radiation stress induces transposable element transcription in *Drosophila melanogaster*. *J of Evolutionary Biology*, **34**, 628–638.

91. Horváth,V., Merenciano,M. and González,J. (2017) Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response. *Trends in Genetics*, **33**, 832–841.

92. Nicolau,M., Picault,N. and Moissiard,G. (2021) The Evolutionary Volte-Face of Transposable Elements: From Harmful Jumping Genes to Major Drivers of Genetic Innovation. *Cells*, **10**, 2952.

93. Neu-Yilik,G., Gehring,N.H., Hentze,M.W. and Kulozik,A.E. (2004) Nonsense-mediated mRNA decay: from vacuum cleaner to Swiss army knife. *Genome Biology*.

94. Harigaya,Y. and Parker,R. (2010) No-go decay: a quality control mechanism for RNA in translation. *WIREs RNA*, **1**, 132–141.

95. Vasudevan,S., Peltz,S.W. and Wilusz,C.J. (2002) Non-stop decay—a new mRNA surveillance pathway. *BioEssays*, **24**, 785–788.