Identification and quantification of transposable element transcripts using Long-Read RNA-seq in *Drosophila* germline tissues

Rita Rebollo[1], Pierre Gerenton[2,3], Eric Cumunel[2,3], Arnaud Mary[2,3], François Sabot[4], Nelly Burlet[2], Benjamin Gillet[5], Sandrine Hughes[5], Daniel S. Oliveira[2,6], Clément Goubert[7], Marie Fablet[2,8], Cristina Vieira*[2,3] and Vincent Lacroix*[2,3]

[1]Univ Lyon, INRAE, INSA-Lyon, BF2I, UMR 203, 69621 Villeurbanne, France.
[2]Université Claude Bernard Lyon 1, Laboratoire de Biométrie et Biologie Evolutive, CNRS, UMR5558, Villeurbanne, Rhône-Alpes, 69100, France.
[3]ERABLE team, Inria, Lyon Rhone-Alpes, Villeurbanne, France
[4]DIADE unit, Univ Montpellier, Cirad, IRD, F-34394 Montpellier Cedex 5, France
[5]Institut de Génomique Fonctionnelle de Lyon (IGFL), CNRS UMR 5242 , Ecole Normale Supérieure de Lyon, Université Claude Bernard Lyon 1 , F-69007 Lyon, France.
[6]São Paulo State University (Unesp), Institute of Biosciences, Humanities and Exact Sciences, São José do Rio Preto, SP, Brazil.
[7]Human Genetics, McGill University, Montreal, QC, Canada
[8]Institut Universitaire de France (IUF), Paris, Île-de-France , F-75231, France.

*Corresponding authors: vincent.lacroix@univ-lyon1.fr and cristina.vieira@univ-lyon1.fr

Abstract

Transposable elements (TEs) are repeated DNA sequences potentially able to move throughout the genome. In addition to their inherent mutagenic effects, TEs can disrupt nearby genes by donating their intrinsic regulatory sequences, for instance, promoting the ectopic expression of a cellular gene. TE transcription is therefore not only necessary for TE transposition per se but can also be associated with TE-gene fusion transcripts, and in some cases, be the product of pervasive transcription. Hence, correctly determining the transcription state of a TE copy is essential to apprehend the impact of the TE in the host genome. Methods to identify and quantify TE transcription have mostly relied on short RNA-seq reads to estimate TE expression at the family level while using specific algorithms to discriminate copy-specific transcription. However, assigning short reads to their correct genomic location, and genomic feature is not trivial. Here we retrieved full-length cDNA (TeloPrime, Lexogen) of *Drosophila melanogaster* gonads and sequenced them using Oxford Nanopore Technologies. We show that long-read RNA-seq can be used to identify and quantify transcribed TEs at the copy level. In particular, TE insertions overlapping annotated genes are better estimated using long reads than short reads. Nevertheless, long TE transcripts (> 4.5 kb) are not well captured. Most expressed TE insertions correspond to copies that have lost their ability to transpose, and within a family, only a few copies are indeed expressed. Long-read sequencing also allowed the identification of spliced transcripts for around 105 TE copies. Overall, this first comparison of TEs between testes and ovaries uncovers differences in their transcriptional landscape, at the subclass and insertion level.

Keywords: long-read sequencing, ONT, transposable elements, regulation, RNA-seq, full-length cDNA

Introduction

Transposable elements (TEs) are widespread DNA sequences that have the ability to move around genomes in a process called transposition (Bourque et al., 2018). TEs can transpose either using an RNA intermediate, in a copy-and-paste mechanism, *i.e.* retrotransposons, or directly through a DNA molecule using different cut-and-paste strategies, *i.e.* DNA transposons. In both cases, the synthesis of a messenger RNA is a fundamental step allowing the production of the transposition machinery, and hence promoting TE replication in the host genome. TE transposition is *per se* a mutational process, and several host mechanisms are in place in order to avoid novel TE insertions, including chromatin remodelling factors, DNA methylation, and small RNAs (Slotkin & Martienssen, 2007). For instance, in *Drosophila melanogaster* ovaries, TEs are the target of piwi-interacting RNAs (piRNAs) that promote TE transcript cleavage, but also deposition of repressive chromatin marks within the TE insertion, blocking any further transcription (Fabry et al., 2021).

In order to appreciate the dynamics of TE regulation, an accurate measure of TE expression is required, including copy-specific information (Lanciano & Cristofari, 2020). While such analyses may be easily obtained in genomes composed of mostly ancient TE copies, discrimination of young TE families, such as LINE-1, AluY and SVA in humans, or the study of genomes composed of mostly active copies as seen in many insects, remains a complex feat. Indeed, TE copies belonging to the same TE family have, by definition, more than 80 % of sequence identity, hampering the study of TE regulation and consequently TE expression in a copy-specific manner (Lanciano & Cristofari, 2020). Most genome-wide analyses interested in TE expression, and even their regulation, focus on TE family-level analysis, where short reads are mapped either against TE consensus sequences or to the genome/TE copy sequences followed by grouping of read counts at the family level (TEcount from the TEtools package (Lerat et al., 2017), TEtranscripts (Jin et al., 2015)). In the past years, many methods have surfaced to take advantage of short-read sequencing datasets and circumvent the multi-mapping problem in order to develop copy-level analysis (for a review see (Lanciano & Cristofari, 2020)). These methods are based on different algorithms that are able to statistically reassign multi-mapped reads to unique locations, for instance with the expectation-maximization algorithm used in TEtranscripts (Jin et al., 2015), SQuIRE (Yang et al., 2019) and Telescope (Bendall et al., 2019).

In the past years, long-read sequencing has become an attractive alternative to study TE biology. Such reads are able to refine TE annotation (Jiang et al., 2019; Panda & Slotkin, 2020), pinpoint new TE insertions (Mohamed et al., 2020; Rech et al., 2022), determine TE DNA methylation rates at the copy level (Ewing et al., 2020), estimate TE expression (Berrens et al., 2022), and finally, detect TE-gene fusion transcripts (Panda & Slotkin, 2020; Dai et al., 2021; Babarinde et al., 2021). Furthermore, long-read RNA sequencing can not only determine which TE copies are expressed but also discriminate between isoforms of a single TE copy produced by alternative splicing. Indeed, TE alternative transcripts have been described in the very first studies of TEs, using techniques such as northern blot (Belancio et al., 2006), but concomitantly with accessible short-read genome-wide analysis, low interest has been given to TE transcript integrity. Nonetheless, TE isoforms have been shown to participate in TE regulation, as observed for the P element in *D. melanogaster*, where a specific germline isoform encodes a functional transposase protein, while in somatic tissues, another isoform acts as a P element repressor (Laski et al., 1986). The regulation of such tissue-specific splicing has recently been attributed to piRNA-directed heterochromatinization of P element copies (Teixeira et al., 2017). The retrotransposon *Gypsy* also produces two isoforms, including an envelope-encoding infectious germline isoform, also controlled by piRNA-guided repressive chromatin marks (Pélisson et al., 1994; Teixeira et al., 2017). Recently, Panda and Slotkin produced long-read RNA sequencing of *Arabidopsis thaliana* lines with defects in TE regulatory mechanisms (Panda & Slotkin, 2020), and were able to annotate TE transcripts, pinpoint TE splicing isoforms, and most importantly, demonstrate that properly spliced TE transcripts are protected from small RNA degradation.

D. melanogaster harbours around 12-20% of TE content, and recent studies have suggested that 24 TE superfamilies are potentially active (Adrion et al., 2017). Nevertheless, no indication of which copies are active has been documented. Here, we describe a bioinformatics procedure using long-read RNA sequencing, which enables the efficient identification of TE-expressed loci and variation in TE transcript structure and splicing. Furthermore, our procedure is powerful enough to uncover tissue-specific differences, as illustrated by comparing testes and ovaries data.

Methods

Reference genome and annotation

The dmgoth101 genome assembly was produced from Oxford Nanopore Technologies (ONT) long-read DNA sequencing and described in (Mohamed et al., 2020). Genome assembly has been deposited in the European Nucleotide Archive (ENA) under accession number PRJEB50024, assembly GCA_927717585.1. Gene annotation was performed as described in (Fablet et al., 2023). Briefly, gene annotation files were retrieved from Flybase (dmel-all-r6.46.gtf.gz) along with the matching genome sequence (fasta/dmel-all-chromosome-r6.46.fasta.gz). We then used LiftOff v1.6.1 (Shumate & Salzberg, 2021) with the command liftoff -g dmel-all-r6.46.gtf -f feature_types.txt -o dmgoth101.txt -u unmapped_dmgoth101.txt -dir annotations -flank 0.2 dmgoth101_assembl_chrom.fasta dmel-all-chromosome-r6.46.fasta to lift over gene annotations from the references to the GCA_927717585.1 genome assembly. One should note that feature_types.txt is a two line txt file containing 'gene' and 'exon'. In order to locate and count the reads aligned against TE insertions, we produced a GTF file with the position of each TE insertion. We have used RepeatMasker with DFAM dataset from *D. melanogaster* (-species Drosophila) TE copies (Dfam_3.1) and then used OneCodeToFindThemAll (Bailly-Bechet et al., 2014), downloaded on november 2020) to merge LTR and internal parts of a TE into one unique feature ./build_dictionary.pl --rm dmgoth101_assembl_chrom.fasta.out --unknown > dmgoth101_dico" and "./one_code_to_find_them_all.pl --rm dmgoth101_assembl_chrom.fasta.out --ltr dmgoth101_dico --unknown. Visualization of alignments of TE copies to their consensus sequences were performed using blastn (Altschul et al., 1990) with the consensuses sequences from the Bergman laboratory that can also be found in GitLab/te_long_read.

Drosophila *rearing*

D. melanogaster dmgoth101 strain was previously described by (Mohamed et al., 2020). Briefly, an isofemale line was derived from a wild-type female *D. melanogaster* from Gotheron, France, sampled in 2014, and sib-mated for 30 generations. Flies were maintained in 12-hour light cycles, and 24º C, in vials with nutritive medium, in small-mass cultures with approximately 50 pairs at each generation.

Long-read RNA-seq and analysis

RNA extraction and library construction

Forty-five pairs of ovaries and 62 pairs of testes were dissected in cold PBS 1X from 4 to 8-day-old adults. Total RNA was extracted using the QiagenRNeasy Plus Kit (Qiagen, reference 74104) after homogenization (using a pellet pestle motor) of the tissues. DNA contamination was controlled and removed by DNAse treatment for 30 minutes at 37°C (Ambion). Total RNA was visualized in agarose gel to check DNA contamination and RNA integrity before storing at -80°C. The two RNA extracts were quantified with RNA BR reagents on Qubit 4.0 (Thermo Fisher Scientific) and qualified with RNA ScreenTape on Tapestation 4150 instrument (Agilent Technologies), the results showing no limited quantity and a high quality of the RNA molecules (RIN >9.8). We then took advantage of the TeloPrime Full-Length cDNA Amplification kit V2 (Lexogen) in order to enrich ovary and testis total RNA in full-length cDNAs (Figure S1). One should note that the amplified cDNAs are smaller than ~3.5 kb. This protocol is highly selective for mRNAs that are both capped and polyadenylated and allows their amplification. TeloPrime recommends 2 µg total RNA per reaction and we performed two reactions for testis (total of 4 µg) and three reactions for ovaries (total of 6 µg). We determined the optimal PCR cycle number for each sample by quantitative PCR. The quantity and quality of the cDNA produced were checked again with Qubit (dsDNA BR) and Tapestation (D5000 DNA ScreenTape) to confirm the correct amplification of the cDNA and absence of degradation in cDNA fragment length profiles. It is important to note that we do not have replicates for the long-read dataset as the primary goal for this experiment was to evaluate the potential of this technique to identify the largest number of expressed TE copies and isoforms. Enriched full-length cDNAs generated from ovaries and testes were then processed into libraries using the SQK-LSK109 ligation kit (ONT) using 3 µg as starting material. The two libraries were sequenced separately in two flow cells R10 (FLO-MIN110) with a MinION device (Guppy version 2.3.6 and fast basecalling). We obtained 1,236,000 reads for ovaries and 2,925,554 for testes that passed the default quality filter (>Q7). Data are available online at the BioProject PRJNA956863.

Mapping

The analysis performed here can be replicated through https://gitlab.inria.fr/erable/te_long_read/, a GitLab containing all the scripts along with links and/or methods to retrieve the datasets used. Quality control was performed with NanoPlot v1.41.6 (De Coster et al., 2018). The median read length was 1.18 kb for ovaries and 1.44 kb for testes, the N50 read length was 1.7 kb for ovaries and 2.19 kb for testes, and the median quality was 7.7 for ovaries and 8.4 for testes (Table S1, Figure S2). Reads were mapped to the dmgoth101 genome using minimap2 (version 2.26) (Li, 2018) with the splice preset parameter (exact command line given in the GitLab). Most of reads (91.3% for ovaries, 98.8% for testes) could be mapped to the genome (Table S1). Out of those mapped reads, the majority (98.8% for ovaries and 95.1% for testes) mapped to a unique location (*i.e.* had no secondary alignment), and the vast majority (99.9% for ovaries and 97.7% for testes) mapped to a unique best location (*i.e.* in presence of secondary alignments, one alignment has a score strictly higher than the others). Indeed, if a read has several alignments with the same alignment score, then this means the read stems from exact repeats in the genome and they cannot be told apart, hence, one cannot know which copy is transcribed. However, if a read has several alignments with distinct alignment scores, then it means that the read stems from inexact repeats. The presence of this read in the dataset means that one of the copies is transcribed and we consider that it is the one with the highest alignment score. While it could be possible that the read actually stems from the copy with suboptimal alignment, this is highly unlikely because it would mean that there is a sequencing error at the position of the divergence between the two copies of the repeat. A sequencing error in any other position of the read would cause a decrease in the alignment score of both locations. An example of a read that maps to several locations, one with an alignment score larger than the others is given in Figure S3.

We also noticed that some reads were only partially mapped to the genome. In practice the query coverage distribution is bimodal (Figure S4), 80% of reads have a query coverage centered on 90%, while the remaining 20% have a query coverage centered on 50%. A thorough inspection of the unmapped regions of these partially mapped reads reveals that they stem from transcripts located elsewhere on the genome. Given that the transcripts covered by the read are themselves fully covered (both the primary locus and the secondary locus), we think that these chimeras are artifactual and were probably generated during ligation steps as previously described (White et al., 2017). Here, we chose to focus on the locus corresponding to the primary alignment and discard the secondary loci. In practice, this corresponds to the longest of the two transcripts. We also ran the same analyses after completely discarding those 20% chimeric reads, but the quantification of TEs is essentially the same (R=0.992, Figure S5). In the remainder of the paper, unless stated otherwise, for each read, we focused on the alignment with the best AS score. We discuss the few cases of ties when required.

Feature assignment

Once a read is assigned to a genomic location, it does not yet mean that it is assigned to a genomic feature. In order to decide which reads could correspond to a TE, we applied the following filters. First, we selected all reads where the mapping location overlaps the annotation of a TE, for at least one base. Then, we discarded all reads that covered less than 10 % of the annotated TE. It is important to note that no filter is based on the number of basepairs or proportion of the read that extends beyond the TE boundaries (Figure S6). Finally, in the case where a read mapped to a genomic location where there are several annotated features (a TE and a gene, or two TEs), we assign the read to the feature whose genomic interval has the smallest symmetric difference with the one of the read. The rationale for introducing this filter is best explained with examples. Figure S7 corresponds to a TE annotation overlapping a gene annotation. All reads map to both features, but the gene is fully covered while the TE is only partially covered. We conclude that the gene is expressed, not the TE. Figure S8 corresponds to a genomic location where there are five annotated TEs and a gene, which all overlap. The features that are best covered are Jockey and the gene. In general, several features may be partially covered by a read and a read may extend beyond each of these features. For each pair read-feature we compute the number of bases that are in the read and not the feature (nr) and the number of bases that are in the feature and not in the read (nf). The sum of these two terms nr + nf is the size of the symmetric difference between the two intervals. We assign the read to the feature with the smallest symmetric difference (Figure 1). This situation occurs frequently and assigning a read to a TE only because it covers it yields an overestimation of TE expression (Figure S9 is an example). The impact of each filter is given in Figure 1. After all filters are applied, there are 1 252 (1 202 uniquely mapping (Table S4) in addition to 50 multi-mapping (Table S6)) reads in ovaries and 8 138 (7 914 uniquely (Table

S3) and 224 multi-mapped (Table S5)) reads in testes that are assigned to TE copies. Our method enables to detect intergenic TEs, intronic TEs and exonic TEs. Counts are summarised in Table S1.

Breadth of coverage

To calculate the breadth of coverage of annotated transcripts, we mapped reads to the reference transcriptome and computed for each primary alignment the subject coverage and the query coverage. Scripts used are available on the git repository (sam2coverage_V3.py).

Gene ontology

To identify whether ovary and testis had genes associated with their tissue-specific functions, we firstly selected genes with at least one read aligned in each sample and then we submitted the two gene lists to DAVIDGO separately (Sherman et al., 2022). Due to the high number of biological terms, we selected only the ones with > 100 associated-genes.

Subsampling analysis

Subsampling of reads was performed using seqtk_sample (Galaxy version 1.3.2) at the European galaxy server (usegalaxy.eu) with default parameters (RNG seed 4) and the fastq datasets. Subsampled reads were then mapped, filtered and counted using the GitLab/te_long_read pipeline.

Splicing

We mapped reads to both the transcriptome and genomic copies of TEs, we selected the ones whose primary mapping was on a TE. We then filtered those exhibiting Ns in the CIGAR strings. Those are the reads aligning to TEs with gaps. We then extracted the dinucleotides flanking the gap on the reference sequence. Scripts used are available on the git repository (SplicingAnalysis.py, splicing_analysis.sh)

Short-read RNA-seq and analysis

RNA extraction and short-read sequencing were retrieved either from (Fablet et al., 2023), at the NCBI BioProject database PRJNA795668 (SRX13669659 and SRX13669658), or performed here and available at BioProject PRJNA981353 (SRX20759708, SRX20759707). Briefly, RNA was extracted from 70 testes and 30 ovaries from adults aged three to five days, using RNeasy Plus (Qiagen) kit following the manufacturer’s instructions. After DNAse treatment (Ambion), libraries were constructed from mRNA using the Illumina TruSeq RNA Sample Prep Kit following the manufacturer's recommendations. Quality control was performed using an Agilent Bioanalyzer. Libraries were sequenced on Illumina HiSeq 3000 with paired-end 150 nt reads. Short-read analysis was performed using TEtranscripts (Jin et al., 2015) at the family level, and SQUIRE (Yang et al., 2019) was used for mapping and counting TE copy-specific expression. A detailed protocol on SQUIRE usage in non-model species can be found here https://hackmd.io/@unleash/squireNonModel. Family-level differential expression analysis was performed with TE transcript (Jin et al., 2015). RNA-seq reads were first aligned to the reference genome (GCA_927717585.1) with STAR (Dobin et al., 2013): the genome index was generated with the options --sjdbOverhang 100 and --genomeSAindexNbases 12; next, alignments were performed for each read set with the parameters -sjdbOverhang 100 --winAnchorMultimapNmax 200 and --outFilterMultimapNmax 100 as indicated by the authors of TE transcript (Jin & Hammell, 2018). TE transcript was ran in two distinct modes, using either multi-mapper reads (--mode multi) or only using single mapper reads (--mode uniq) and the following parameters: --minread 1 -i 10 --padj 0.05 --sortByPos.

Results and discussion

Transposable element transcripts are successfully detected with long-read RNA-seq

In order to understand the TE copy transcriptional activity and transcript isoforms in gonads of *D. melanogaster*, we extracted total RNA from ovaries and testes of dmgoth101 adults, a French wild-derived strain previously described (Mohamed et al., 2020). Prior to long-read sequencing, we enriched the total RNA fraction into both capped and polyadenylated mRNAs in order to select mature mRNAs potentially associated with TE activity (see material and methods for the details on the TeloPrime approach). Sequencing yielded between ~1 to ~3 million reads per tissue, ranging from 104 to 12,584 bp (median read length ~1.4 Kb, Figure S1-2, Table S1). Reads were subsequently mapped to the strain-specific genome assembly (Mohamed et al., 2020) using the LR aligner Minimap2 (version 2.26) (Li, 2018). Most reads mapped to the genome (91.3% for ovaries and 98.8% for testes, Table S1), and the majority of them mapped to a unique location (*i.e.* had no secondary alignment, 98.8% for ovaries and 95.1% for testes), and the vast majority mapped to a unique best location (*i.e.* presence of secondary alignments, one alignment has a score strictly higher than the others, see Methods, 99.9% for ovaries and 97.7% for testes).

In order to validate the long-read RNA-seq approach, we first determined the breadth of coverage of all expressed transcripts and showed that the majority harbour at least one read covering more than 80% of their sequence (70.2% in ovaries and 71.8% in testes). Only a few reads correspond to partially covered transcripts, as most reads cover more than 80% of the transcript sequence (63.4% in ovaries, 77.4% in testes - Figure 1A), although very long transcripts (≥ 5 kb) are poorly covered (Figures S10-11). The transcriptomes obtained are enriched in typical germline ontology terms, such as "spermatogenesis" for testes, and "oogenesis" for ovaries (Figure S12). Finally, while the first version of the TeloPrime protocol could not be used for quantification (Sessegolo et al., 2019), the quantifications obtained here correlate well with available short-read sequencing (rho=0.78, R=0.44, Figure 1B). We also noticed that the correlation between the two technologies is weaker for very long transcripts.

Although most long reads map to a unique location on the genome, ensuring that a read stems from the TE copy is not straightforward. First, a read may overlap with a TE copy only for a few nucleotides, suggesting the read is not a consequence of TE transcription. To rule out these cases, all reads covering less than 10% of the TE copy were discarded. Second, in many instances, the read overlapped both a gene and one or several TEs. In these cases, the feature for which the coverage was best (see Methods, Figure 1C and Figures S7-8) was selected. Overall, after applying these filters, 1 252 reads in ovaries and 8 138 reads in testes were assigned to TE copies (Table S1, Figure 1D). Out of these, 37% overlap exons, 21% overlap genes but not exons and 41% do not overlap genes in ovaries. In testes, 20% are exonic, 14% intronic and 65% intergenic (Figure 1E). Additionally, a large fraction of reads overlapping TEs and genes are assigned to genes (52% in testes and 79% in ovaries) (Figure 1D, 1E).

To ensure this long-read dataset is able to recover transcripts encompassing all TE copy lengths present in the genome, we compared the length distribution of all TE insertions with the length of all mapped reads (Figure 1F). While genomic TE copies range from a few base pairs to ~15 Kb, 75% are smaller than 2 Kb. The average length of reads mapping to TEs encompasses the majority of TE copies, but does not cover TE transcripts longer than 4.5 Kb. Reads mapping to genes have a similar distribution (Figure 1F). The absence of very long reads indicates that either very long mRNAs are absent from the sample (as suggested by the cDNA profile, Figure S1), or the TeloPrime technique is not well tailored for capturing very long transcripts. In order to clarify this point, we compared the quantification obtained by Illumina and ONT TeloPrime for short (<3 Kb), long (3 Kb-5 Kb) and very long transcripts (>5 Kb), and obtained the following Spearman correlations of 0.83 (n=3 603 genes), 0.71 (n=378) and 0.62 (n=130), respectively (Figure 1B for ovaries, Figure S13 for testes). Furthermore, reads covering very long annotated transcripts (>5kb) tend to be partial (Figure S10, S11, S14). We conclude that, although very long transcripts are rare (<0.1% of reads), the Teloprime protocol could underestimate their presence.

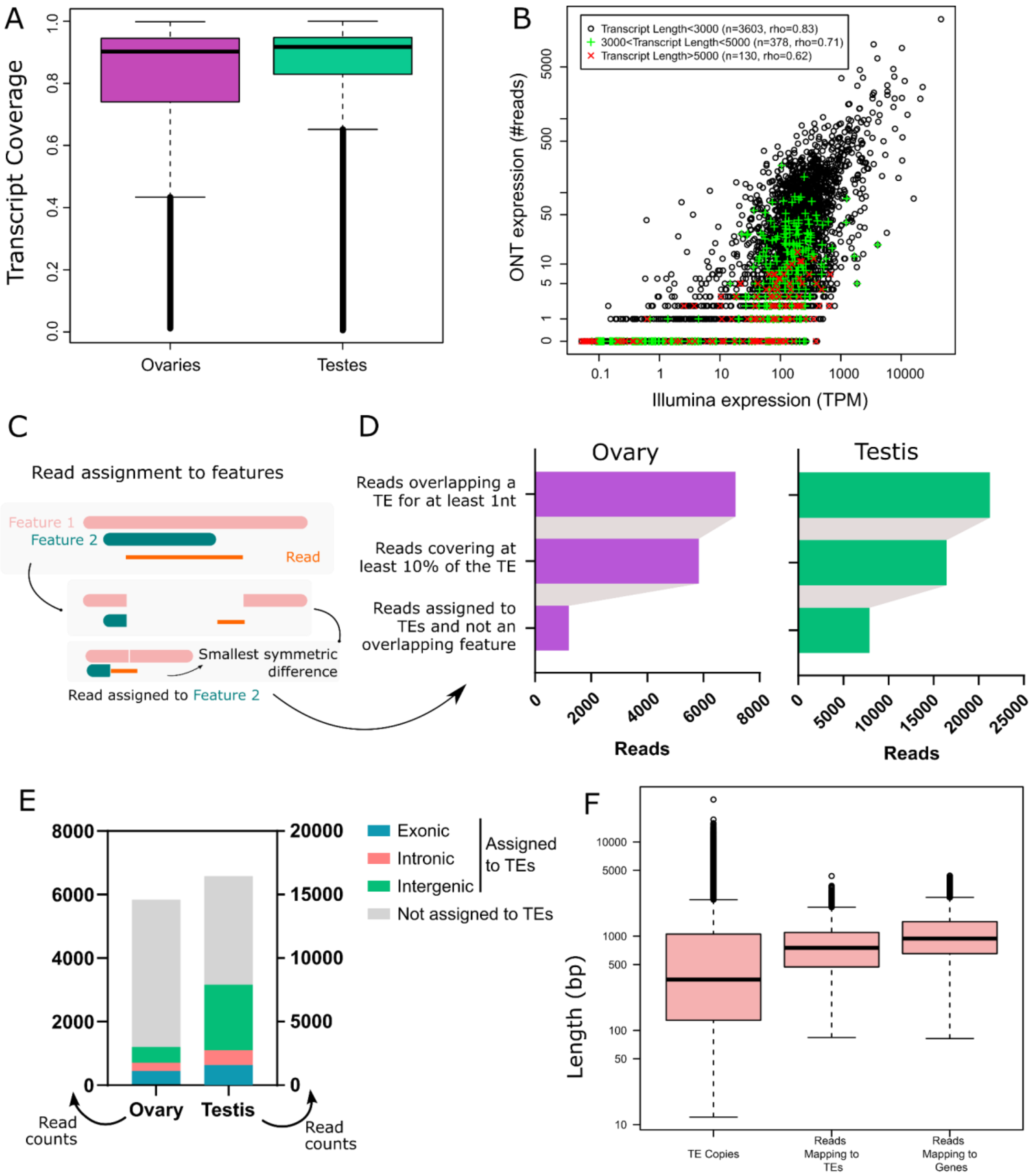

Figure 1: Long-read RNA-seq of *Drosophila melanogaster* ovaries and testes. **A.** Transcript coverage by long-read RNA-seq in ovaries and testes. **B.** Gene expression quantification using Illumina and ONT sequencing in ovaries. Each dot is a gene with a single annotated isoform. Transcripts longer than 5 kb tend to be undersampled using TeloPrime. **C.** Read assignment to features. In the case where a read aligns to a genomic location where two features are annotated, the read is assigned to the feature with the best coverage. Two dimensions are considered. The read should be well covered by the feature, and the feature should be well covered by the read. In practice, we calculate the symmetric difference for each read/feature and select the smallest. In this example, the read is assigned to Feature 2. **D.** Impact of filters on the number of reads assigned to TEs. **E.** Number of reads mapping to TEs separated into three categories (intronic, exonic or intragenic), and reads that have not been assigned to TE copies. **F.** TE copy and read length distribution. Reads mapping to TEs encompass most TE copy length but lack transcripts longer than 4.5 Kb, as also observed for reads mapping to genes.

TE mRNA landscape is sex-specific

Taking into account all the filtering steps, only 0.28% (8 138/2 925 554) and 0.10% (1 252/1 236 000) of long reads aligned to TE copies in testes and ovaries respectively (Table S1). Given the differences in sequencing depth between both tissues, we have computed the number of reads assigned to TEs based on different sets of subsampled reads, and show that TE reads are more abundant in testes than in ovaries (Figure 2A). We identified 130 TE families capable of producing capped-polyadenylated mRNAs (Table S2), of which 70 belong to Long-terminal repeat (LTR) elements (retrotransposons that possess LTR sequences surrounding a retroviral-like machinery). Despite the high number of shared transcribed TE families (96/130), the transcriptional landscape between ovaries and testes is quite different (Figure 2B for the complete dataset and Figure S15 for a subsampled dataset). While LTR elements dominate the transcriptional landscape in both tissues, LINE elements are the second most transcribed TE subclass in testes, while in ovaries, DNA families harbor more read counts (Figure 2B). The transcriptional landscape within TE subclasses between tissues is very similar for retrotransposons, with Gypsy and I-Jockey superfamilies being the most expressed LTR and LINE elements respectively. However, the DNA subclass transcriptional landscape is quite different between testes and ovaries: TcMar-Pogo is the most expressed DNA superfamily in ovaries, while RC elements are abundantly transcribed in testes.

Globally, TE families show higher long-read counts in males compared to females (Figure 2C), not only because male samples were more deeply sequenced (2.3 times more), but also because the proportion of reads that map to TEs is higher in males even when subsampling the same number of reads between tissues (Figure S16 for a subsampled dataset). *HETA* (I-Jockey) and *DNAREP1_D* (RC) are the top two families in male TE transcript counts, with 743 and 713 long-read transcripts respectively. In females, *pogo* (TcMar-pogo) is the TE family amounting the most transcripts with 213 long reads (while only 32 in males), followed by *DNAREP1_D* (RC) with 141, and *Copia* (Copia) with 99 long reads (but both families with higher transcript counts in males, Table S2). There are only five TE families that yielded long-reads in ovaries and not in testes, *BARI_DM* (TcMar-Tc1 - DNA), *Gypsy7* (Gypsy - LTR), *Gypsy11* (Gypsy - LTR), *Copia2_LTR_DM* (Copia - LTR) and *Helena_RT* (I-Jockey - LINE), but they all harbor only one or two long reads suggesting their expression is low. There are 29 families detected only in testes, four DNA elements (*Transib-N1_DM* (CMC-Transib), *NOF_FB* (MULE-NOF), and two TcMar-Tc1 (*Bari1* and *Minos*)), 13 LINE elements (10 I-Jockey, two R1, and one R1-LOA, see Table S2 for details), and 12 LTR families (two Copia, four Gypsy, two Pao and four unknown families), ranging from one to 74 long reads per TE family. Finally, eleven TE families show no long-read mapping in either tissue. Collectively, long-read sequencing is able to discriminate between ovaries and testes TE transcriptional landscapes.

Short-read RNA sequencing of ovaries and testes, followed by estimation of TE family expression with TEtranscripts (Jin et al., 2015) - TE expression estimation *per* TE family, see material and methods for more information) recapitulates the long-read RNA sequencing profiles (Figure 2D and Figure S17). Despite the fact that TE transcripts are overall poorly expressed, the estimation of their expression level is reproducible across technologies. The correlation is higher for testes (r=0.74, rho=0.86) than for ovaries (r=0.45, rho=0.67), where the coverage is weaker. Indeed, as previously stated, the total contribution of TE to the transcriptome is weaker for ovaries and the sequencing shallower.

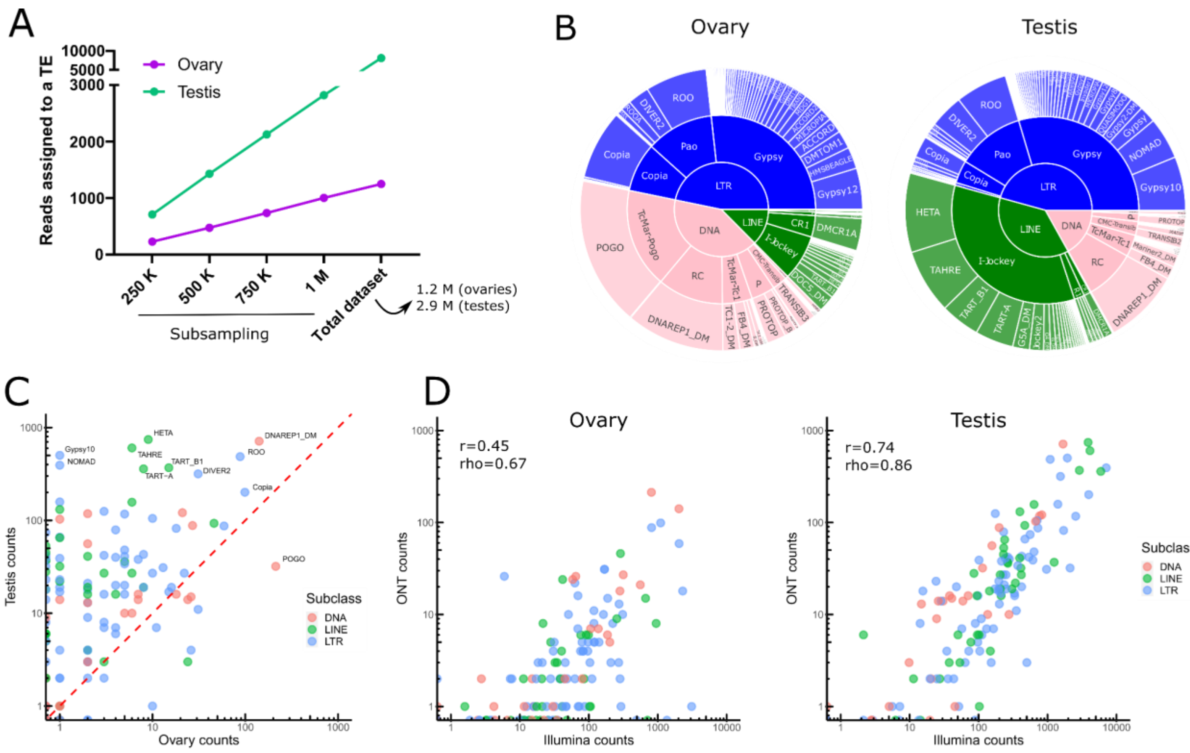

Figure 2: Transposable element transcriptional landscape in ovaries and testes of *Drosophila melanogaster*. **A.** Reads assigned to TEs are more abundant in testis. Subsampling of reads from 250 000 to 1 million reads, along with the complete dataset, show a higher number of reads assigned to TEs in testes than in ovaries. **B.** Global TE transcriptional landscape using ONT long-read sequencing. The outer ring, middle ring and inside circle represent TE family, superfamily and subclass respectively. The area in the circle is proportional to the rate of expression. **C.** Comparison of TE expression ratio between testes and ovaries ONT long-read datasets. **D.** Comparison between Illumina and ONT datasets for estimating the expression levels at the TE family level. Each dot is a TE family. TEtranscript is used for short reads. The correlation between quantifications by both technologies is higher for testes (rho=0.86) than for ovaries, (rho=0.67) where the coverage is weaker. In both cases, it is compatible with what is observed for genes.

Long-read sequencing successfully retrieves single-copy transcripts

The main objective of using long-read sequencing after the TeloPrime full-length cDNA enrichment protocol is to recover copy-specific mature TE transcripts and potential isoforms. There are 1 202 long reads mapping uniquely to a TE copy in ovaries, while 47 map to multiple copies within the same family and three reads are unable to be assigned to a specific TE family. In testis, 7 914 reads are assigned to specific TE copies, 206 to TE families and 18 are assigned at the superfamily or subclass level. The overall percentage of reads unable to be assigned to a particular copy is therefore quite small (4% and 2.7% for ovaries and testes respectively). The only family harboring only multimapped reads is *NOMAD* with one single long-read in ovaries that matches three different copies. In contrast, in testes, 383 long-reads are assigned uniquely to *NOMAD* copies and only 10 multi-mapped reads are detected.

In ovaries, out of 101 TE families detected (at least one read), there are 16 families that harbor only one multi-mapped read, and three families with 3 to 21 multimapped reads, *copia, pogo* and *micropia* (Figure 3A). While only 3% of *pogo* reads are multimapped in ovaries (7 out of 213 reads), *micropia* and *copia* harbor a higher percentage of multimapped reads, 17% of 18 reads for *micropia* and 21% of 99 reads for *copia.* In testis, 125 TE families are expressed (at least one read), and 39 of them have multi-mapped reads (Figure 3A). As observed in the ovary dataset, the number of multimapped reads is low for most families, with only seven families harboring more than 10 multimapped reads. While *copia* harbors the most multimapped reads in testis (47 out of 201 long-reads), *blastopia* shows a higher multi-mapped ratio with 54% of reads multimapped (26 out of 48 long-reads). *Transpac* also shows an important number of multi-mapped reads with 22 out of 66 reads mapped. In total, 45 TE families have both uniquely and multi-mapping reads in ovaries and testis (Figure 3A).

We uncovered 404 and 1 078 TE copies harbouring at least one long-read unambiguously in ovaries and testes respectively. When taking into account multi-mapped reads, an additional 53 and 94 TE copies are potentially expressed in each tissue (Table S2). However, it is important to note that the number of assigned multi-mapped reads to each copy is quite small. For instance, in ovaries, 47 of these potentially expressed copies only harbor one multi-mapped read, five copies harbor two, and a single *pogo* copy harbors three multi-mapping reads. As a comparison, the two most expressed copies in ovaries are two *Pogo* copies, POGO$3L_RaGOO$9733928$9735150 and POGO$X_RaGOO$21863530$21864880, with 80 and 77 uniquely mapping reads, and no multi-mapping read (Table S4). In testis, out of 94 copies, 71 have only a single multi-mapped read and only eight copies show more than five multi-mapped reads. Among these copies, there are three *Blastopia* copies with 23, 22 and 19 multi-mapped reads, two *Transib2* copies (16 reads each), one *M4DM* (15 reads), one *Burdock* (12 reads) and finally one *Copia* (six reads) (Table S3). As a comparison, the top expressed copy in testis is a *Gypsy10* (Gypsy10$3R_RaGOO$761442$762629) with 493 uniquely mapping reads and no multimapping ones. If searching for the most expressed copy among the same TE families as described having only abundant multi-mapped reads in testis, *Blastopia* harbors only 11 uniquely mapping reads, *Transib2* has 41 uniquely mapped reads and no multimapped ones, M4DM copy has 35, and *Burdock* shows 15 uniquely mapping reads and five multimapped ones (Figure 3B). Therefore, despite a few exceptions (*blastopia*, *transib2, M4DM and Burdock)*, long-read sequencing can indeed identify single-copy TE transcripts.

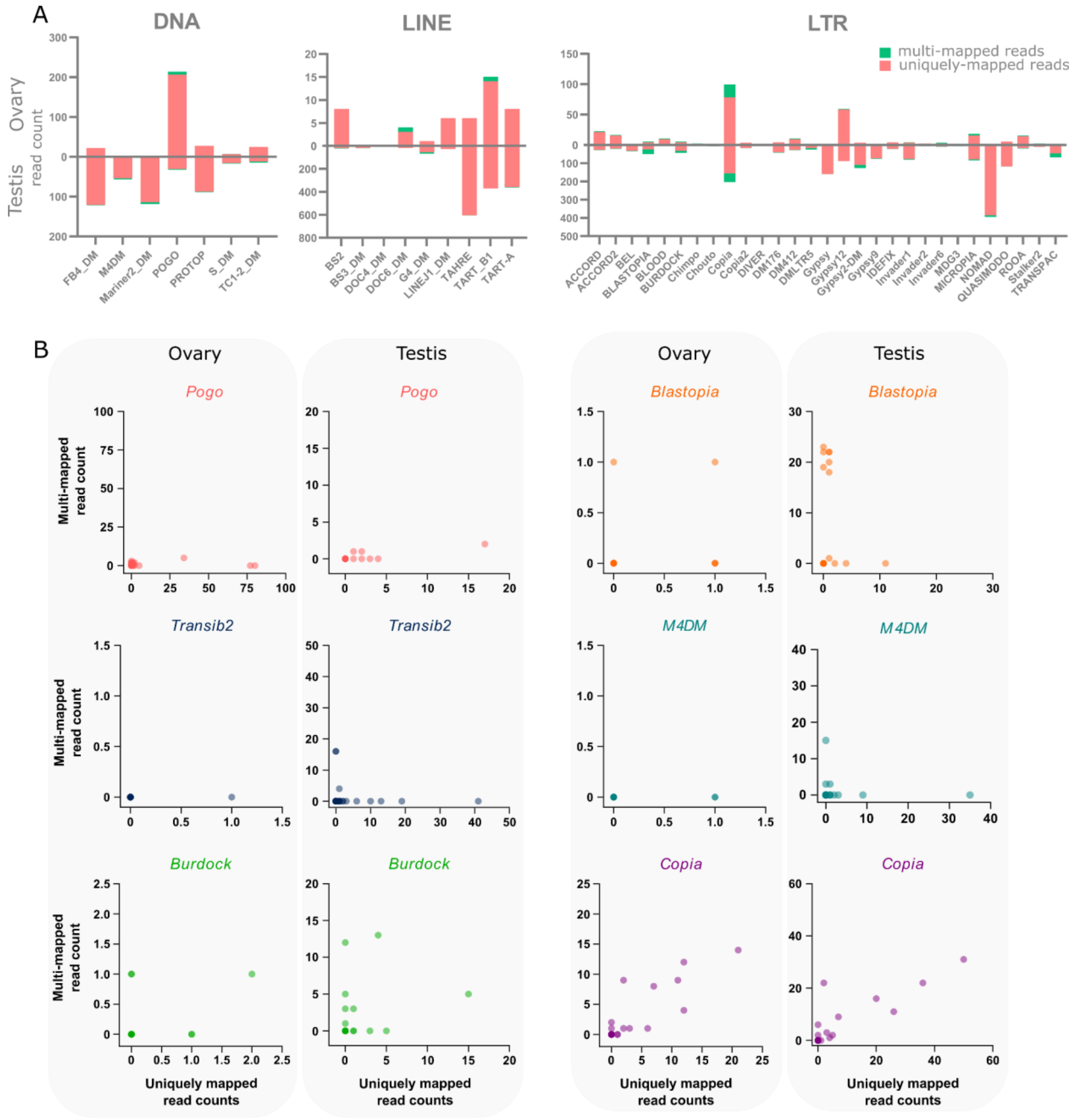

Figure 3: Multi-mapping and uniquely mapping ONT reads. **A.** Distribution of uniquely and multimapped reads across TE families in ovaries and testes (only TE families harboring at least one multimapped read are shown). **B.** Association between multi-mapped and uniquely mapped reads at the copy-level for the TE families showing copies harboring abundant multi-mapped reads.

Within a TE family, the contribution of each TE copy to the family transcriptional activity is variable. In general, only a few insertions produce transcripts, even if taking into account multi-mapped reads (Figure 4A for uniquely mapping reads and Figure S18 for all reads). However, *Transpac* (LTR, Gypsy) copies are nearly all expressed in testes (10 expressed copies and two potentially expressed copies out of 15), while in ovaries, *pogo* (DNA, TC-mar-Pogo) harbors 12 copies producing transcripts, and five potentially expressed copies out of 26. *DNArep1* (RC) is the most abundant TE family in the *D. melanogaster* genome and is also the family harbouring the most transcribed copies in both ovaries and testes (72 and 170 respectively out of 2 555 copies). In ovaries, out of the 404 insertions with at least one mapped read, 23 had more than 10 mapped reads. In testes, out of the 1 078 insertions with at least one uniquely mapped read, 157 had in fact more than 10 mapped reads.

While many TE copies within a family are indeed able to produce transcripts, there are significant differences in copy expression rate (Figure 4B for the 10 most expressed TE families in ovaries and testes, and Figure S19 for a subsampled dataset). For instance, *Gypsy10* (LTR - Gypsy*)* harbors a highly active copy with 493 uniquely mapping reads out of 502 total counts in testes, while only one read is detected in ovaries. As a contrast, *DNAREP1* (DNA, RC) and *Roo* (LTR, Pao) have several copies that contribute to the TE family expression (Figure 4B). Finally, some families show copies transcriptionally active in both ovaries and testes, as for instance *Copia* (LTR, Copia).

In ovaries, where *pogo* has the highest number of long reads (213), an insertion of 1 222 bp in the 3L chromosome (POGO$3L_RaGOO$9733928$9735150) accumulates nearly 37% of the family total read count (Figure 4B and C). This specific *pogo* insertion is located in the intron of the CG10809 gene. The same pattern is observed for the second most expressed *pogo* insertion (POGO$X_RaGOO$21863530$21864880), also located in the intron of a gene (CG12061), expressed in testes and not in ovaries (Figure S20). CG12061 is a potential calcium exchange transmembrane protein and has been previously shown to be highly expressed in the male germline (Li et al., 2022). Indeed, using long-read sequencing, CG12061 is highly expressed in testes compared to ovaries, and curiously, the intronic *pogo* insertion is only expressed when the gene is silent (in ovaries). The other expressed insertions of *pogo* are located in intergenic regions (Figure S21).

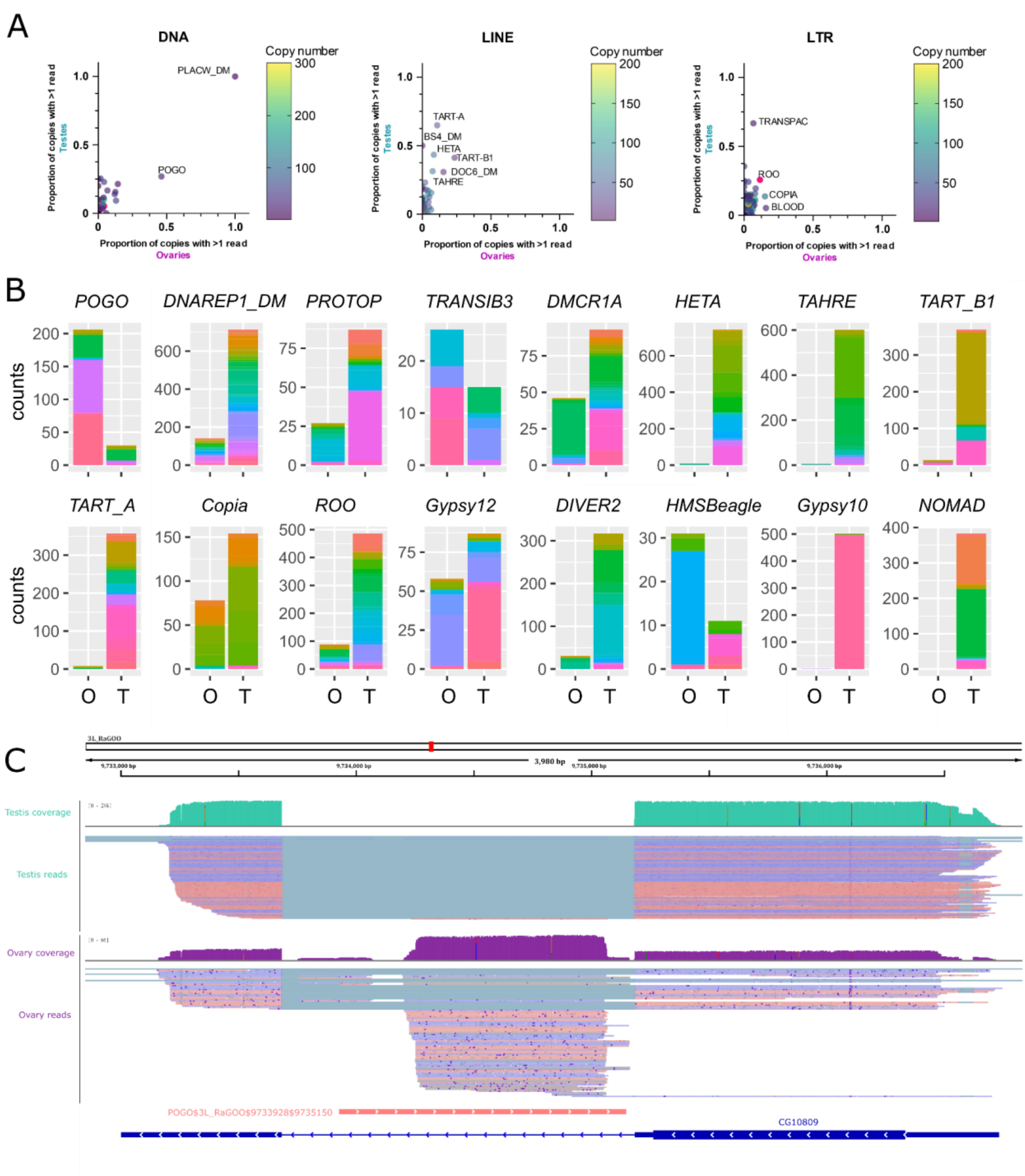

Figure 4: Transcription of transposable element copies. **A.** Frequency of transcribed copies (read > 1) within TE subfamilies in ovaries and testes, along with genomic copy number (color bar, 1 to 200 (LINE/LTR) or 300 (DNA) copies). All TE families harbouring more than 200 (LINE/LTR) or 300 (DNA) copies are depicted in pink. For DNA elements, *DNAREP1_DM* harbours 2 555 copies, *Protop* 305 and *ProtopA* 347. For LINE families, *DMCR1A* has 583 copies and *FW2_DM* 216. The LTR families, *idefix* (227) and *roo* (218) are also depicted in pink. Most TE subfamilies have only a couple of copies producing transcripts, while the majority of HETA copies are expressed in testes for instance (middle panel). **B.** Distribution of read counts per copy for the 10 most expressed copies in ovaries and testes (16 TE families total), showing the overall expression of specific copies within a TE family (Table S3 and S4). Copies are represented by different colors within the stacked bar graph. O: ovaries, T: testes **C.** IGV screenshot of a *pogo* copy (POGO$3L_RaGOO$9733928$9735150, in pink). In green, testis coverage and below mapped reads, in purple the same information for ovaries. Dmgoth101

repeat and gene tracks are also shown and more information on the annotation can be seen in the material and methods section.

Finally, using short-read sequencing and a tool developed to estimate single-copy expression (Squire (Yang et al., 2019)), we compared the overall TE copy transcriptional landscape between short and long reads (Figure S20). There was a poor correlation with the ONT estimations (rho=0.23, r=0.18 for ovaries and rho=0.37, r=0.32 for testes). At the family level, the quantifications obtained by Squire were comparable to the ones obtained with long-reads (rho=0.66, r=0.49 for ovaries and rho=0.77, r=0.34 for testes, Figure S22). Examination of instances where the two techniques differed most, shows that Squire tends to overestimate the expression of TE insertions completely included in genes (Figure S9). Indeed, while long-reads can easily be assigned to the correct feature because they map from the start to the end of the feature, many of the short-reads originating from the gene also map within the boundaries of the TE. Methods based on short reads could clearly be improved, based on the study on such instances where there is a discordance.

Transcripts from full-length transposable element copies are rarely detected

Many insertions produce transcripts that are shorter than the annotated TE and are likely unable to participate in TE transpositional activity. Furthermore, even in the case where the transcript fully covers the insertion, the copy itself might have accumulated mutations, insertions and deletions making it unable to transpose. To assess this, we computed the query coverage of the reads with regards to the insertion they correspond to. We find that one-third of the insertions have at least 80% query coverage (Figure 5A). However, out of these insertions, only a few of them are close in length to a functional full-length sequence. In order to search for potentially functional, expressed copies, we filtered for copies with at least five long-reads detected, and covering at least 80% of their consensus sequences. In ovaries these filters correspond to seven insertions: one *pogo*, five *Copia* insertions and one *MAX*, and in testes, there are nine potentially functional insertions, five *Copia* insertions, three *Mariner2*, and one *DM1731*. While all *copia* insertions expressed with at least five reads are full-length, other TE families show mostly internally deleted expressed copies (Figure 5B). Indeed, a closer analysis of *pogo*, the most expressed TE subfamily in ovaries, shows only one full-length copy expressed (POGO$2L_RaGOO$2955877$2958004), but at low levels (five reads in ovaries, and two in testes). Instead, the other three expressed *pogo* copies with at least five reads in ovaries (80, 77 and 34), are internally deleted (Figure 5B). Hence, ONT long-reads detects only a small number of expressed full-length copies.

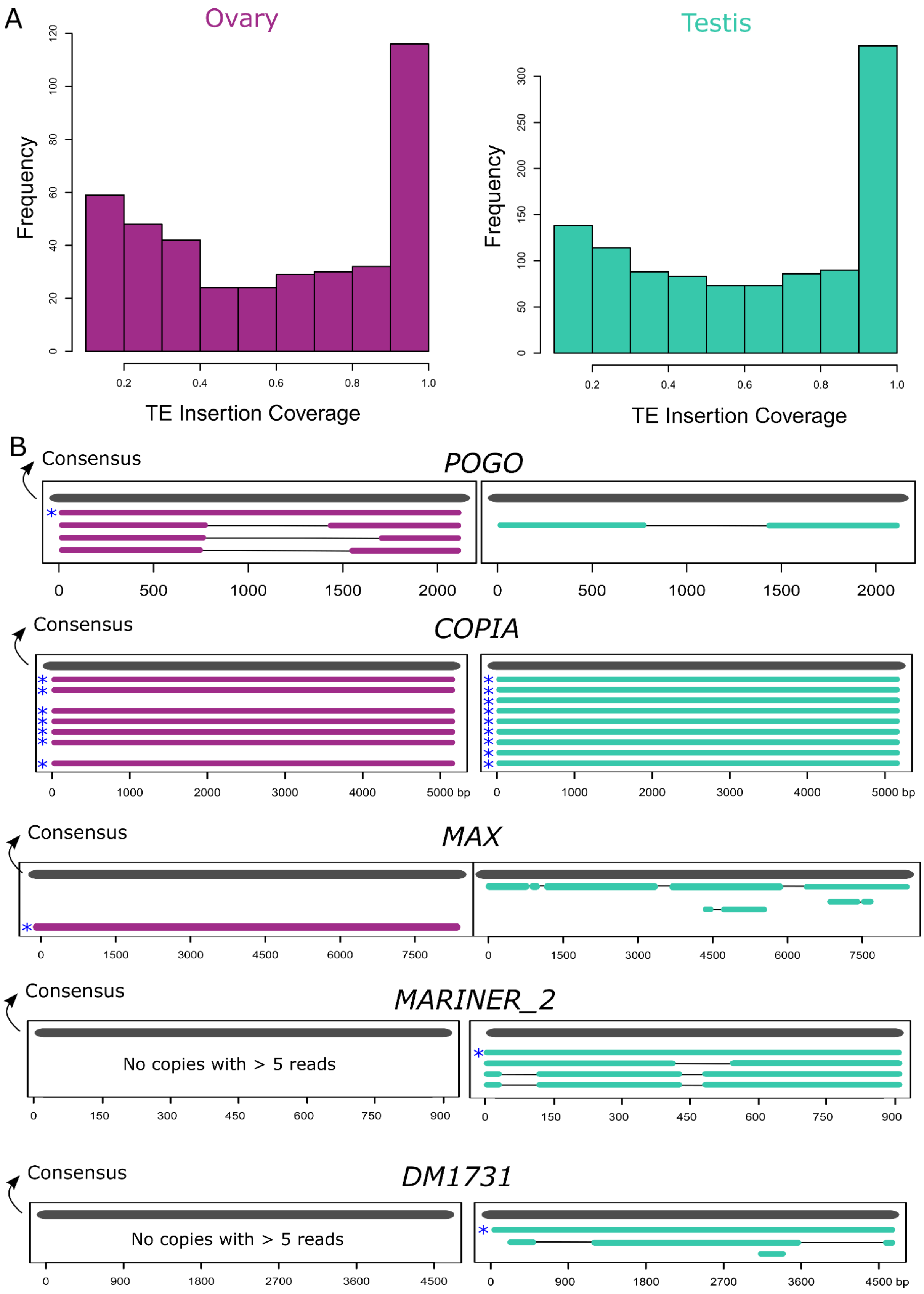

Figure 5. TE transcripts stem mostly from deleted or truncated copies. **A.** Coverage of ONT reads on TE insertions. One-third of copies are covered for at least 80% of their length. **B.** Alignment of copies belonging to the five TE families where at least one full-length expressed copy (80% of consensus) was observed with more

than five long-reads. All copies represented have at least five long-reads. Consensus sequences are represented in grey and copies are either purple for ovaries or green for testes. Asterisks depict the full-length copies.

Long-read sequencing unveils novel spliced TE isoforms

A closer analysis of the reads stemming from the detected full-length copies shows that many of them do not cover the copies completely (Figure 6). For instance, five *Copia* copies are at least ~80% of the consensus sequences and have at least five long-reads detected (Figure 5B), however, although the reads map from the 5' end to the 3' end of the copy, they map with a gap (Figure 6 and Figure S23-S25). *Pogo, Max* and *DM1731* also show such gapped alignments. Inspection of these gaps reveals that they are flanked by GT-AG consensus, suggesting that those transcripts are spliced. Only *Mariner2* shows reads that correspond to the full-length copy, but one should note that the consensus sequence of *Mariner2* is smaller than 1 Kb, while the other elements are much longer. As stated before, very few cDNA molecules longer than ~4 Kb have been sequenced (Figure S1), suggesting either that such longer transcripts are rare, and/or that the method used here for cDNA amplification induces a bias towards smaller sequenced fragments. Collectively, long-read sequencing shows that despite the presence of potentially functional, full-length copies in the *D. melanogaster* genome, only a few of these are detected as expressed in testes and ovaries, and the reads that are indeed recovered seem to be spliced.

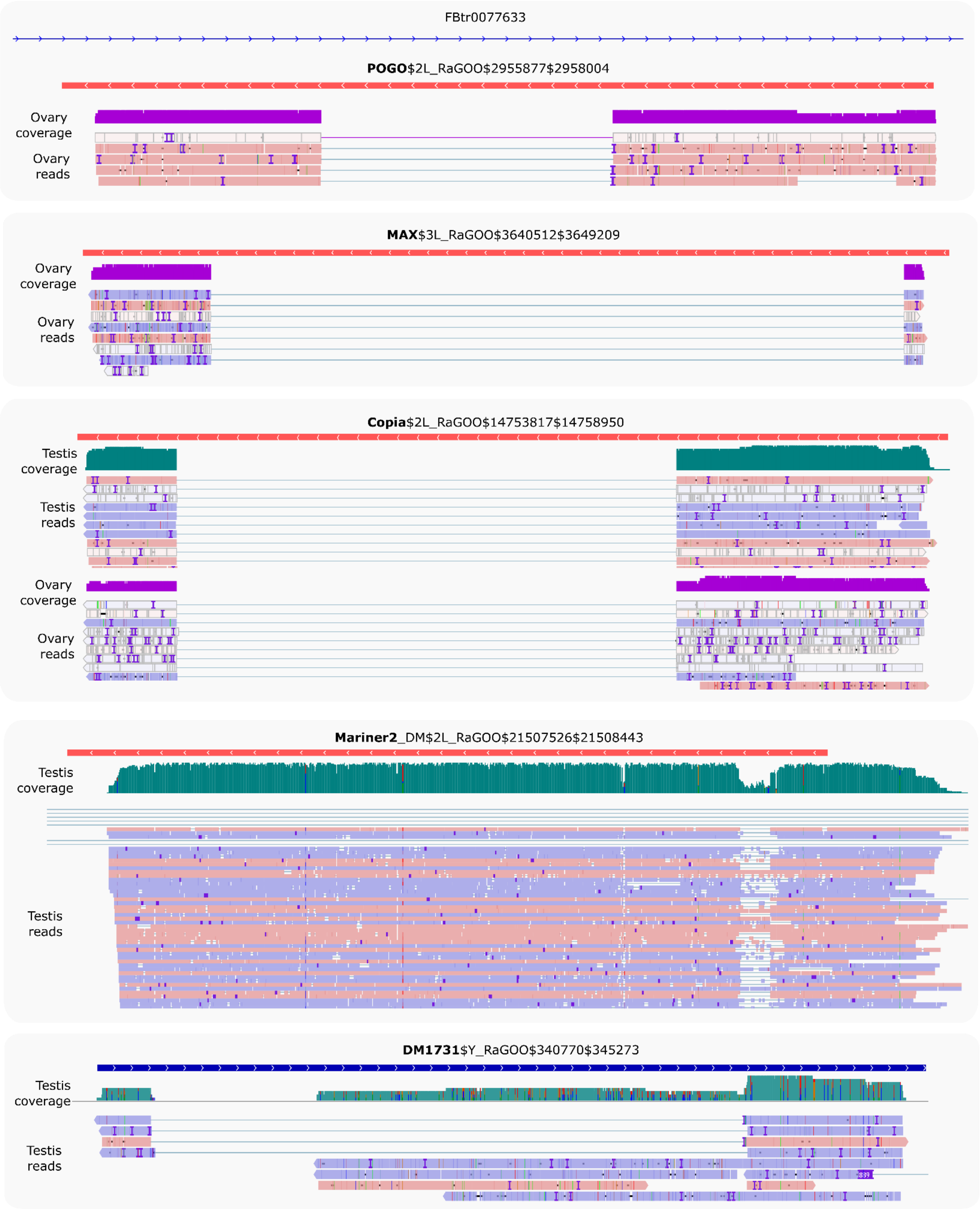

Figure 6. Full-length copies produce spliced transcripts. IGV screenshot of uniquely mapping reads against putative full-length copies (copies > 80% of the consensus sequence length) harboring at least five reads (see Figure 5B). Only *copia* has multi-mapped reads that can be appreciated in Figure S23. Dmgoth101 repeat and gene tracks are also shown and more information on the annotation can be seen in the material and methods section. Ovary and/or testis coverage and reads are shown below the TE copies.

While most TEs do not harbor introns, there are a couple of exceptions previously described in *D. melanogaster.* Indeed, P elements are known to be regulated tissue-specifically by alternative splicing mechanisms, involving piRNA targeting (Laski et al., 1986; Teixeira et al., 2017). *Gypsy* copies are able to produce ENV proteins through mRNA alternative splicing (Pélisson et al., 1994; Teixeira et al., 2017). As with P elements, *gypsy* splicing is also thought to be regulated by piRNAs. Finally, *Copia* elements produce two isoforms, a 5 Kb and a 2.1 Kb (which is a spliced product of the 5 Kb mRNA) (Miller et al., 1989; Yoshioka et al., 1990). The 2.1 Kb encodes the GAG protein and is produced at higher levels than the other proteins (Brierley & Flavell, 1990). While the shorter transcript can be processed by *Copia* reverse transcriptase, the 5 Kb full-length isoform is clearly preferred (Yoshioka et al., 1990). Most of these discoveries were obtained through RT-PCR sequencing of amplicons, or recently, through short-read mapping. Nevertheless, systematic analysis of TE alternative splicing in *D. melanogaster* is lacking, due to the difficulty of detecting such isoforms from short-read data. Here we used long-read sequencing to mine for such splicing isoforms. We searched for reads harboring a gap compared to the reference sequence (presence of N's in the CIGAR string). In order to ensure that those gaps corresponded to introns, we searched for flanking GT-AG splice sites (see methods, and Figures S26-S29). In ovaries, out of 22 insertions supported by at least 5 reads, 15 exhibited at least one gapped read (Figure 7). For all tested insertions, the majority of gapped reads exhibited a GT-AG consensus, except for one insertion (it was CT-TA). In testes, out of 163 insertions supported by at least 5 reads, 100 exhibited at least one gapped read, 91 with a GT-AG consensus (Figure 7). Out of the 9 others, 6 exhibited only one of two gapped reads, the 3 remaining ones with a CT-AT, GA-CG and AT-AG consensus. Those could correspond to non-canonical splicing. They could also correspond to a heterozygous deletion or to the expression of a deleted copy located in a non-assembled part of the genome. Overall, we find that the vast majority of gaps are flanked by GT-AG consensus, and we conclude that they correspond to spliced introns. These introns are however not systematically spliced, because in many cases the proportion of spliced reads is between 0 (never spliced) and 1 (always spliced).

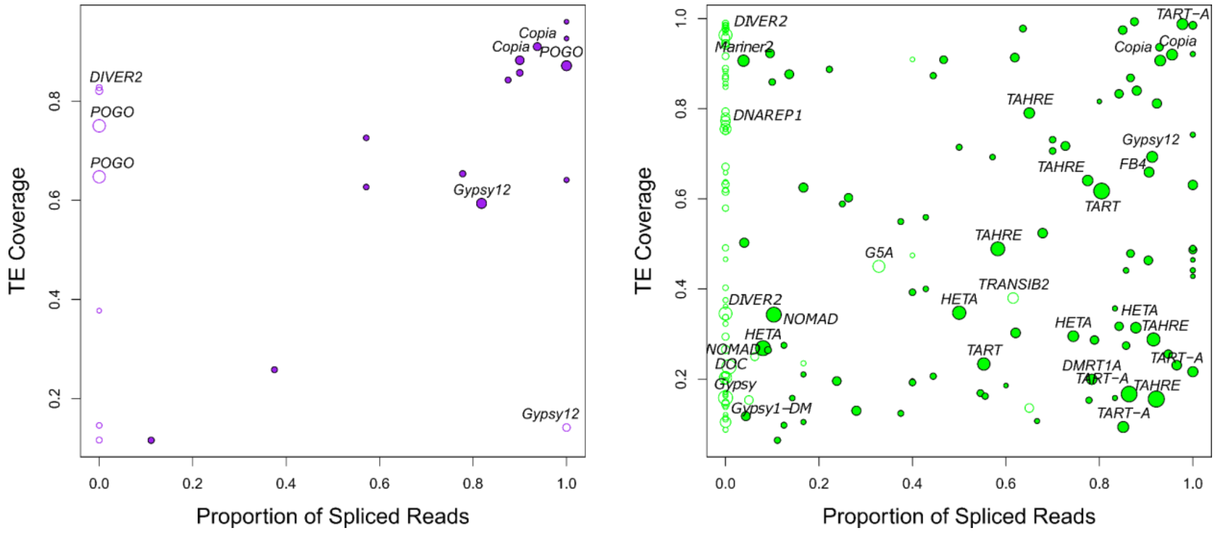

Figure 7. TE spliced transcripts are ubiquitous. Left depicts ovaries, right depicts testes. Each circle depicts a TE insertion supported by at least 5 reads and their size is proportional to the expression level of the insertion. The TE family name is written for highly expressed insertions (>30 reads in testis, >10 reads in ovaries). The X-axis represents the proportion of reads that align with a gap (presence of N's in the CIGAR

string), while the Y-axis represents the proportion of the insertion covered by reads. Unfilled circles correspond either to TE insertions with no gaps, or to TE insertions with gaps that do not exhibit GT-AG sites.

While the proportion of spliced transcripts stemming from a TE copy can vary, there are a couple of copies that only produce spliced transcripts, as POGO$2R_RaGOO$7201268$7202754 for instance. *Pogo* is the most expressed TE family in ovaries, with 12 out of 26 copies producing capped poly-A transcripts corresponding to 213 long reads, while only 7 expressed copies with a total of 32 long-reads are observed in testes, despite the higher coverage. While we previously noted that only one full-length copy is transcribed in ovaries (and in testes albeit with a lower number of reads), there are many truncated or deleted copies that are transcribed (Figure 5B). POGO$2R_RaGOO$7201268$7202754 is one of the internally deleted copies, and it produces a spliced transcript present in both testis and ovaries (Figure S21). The splicing of this short intron (55 nt) has been previously reported (Tudor et al., 1992) and enables the splicing of the two ORFs of *pogo* into a single continuous ORF. This particular copy (POGO$2R_RaGOO$7201268$7202754) is however non-functional since it contains a large genomic deletion located in the ORF near the intron. POGO$X_RaGOO$21863530$21864880 (Figure S20) also contains a large genomic deletion, encompassing the intron, explaining why there are no spliced transcripts for this copy.

Despite the presence of full-length *Copia* insertions in the genome, only spliced transcripts were uncovered in the long-read sequencing (Figure 6). In contrast, with Illumina short reads, we see both spliced and unspliced transcripts (Figure 8). A similar pattern occurs with MAX$3L_RaGOO$3640512$3649209 (Figure S30), but not with POGO$2L_RaGOO$2955877$2958004 or DM1731$Y_RaGOO$340770$345273 (Figure S31-32). The full-length Copia transcripts are 5 Kb, and are less abundant than the spliced transcripts (~10 times less). The lack of such a full-length transcript in the long-read sequencing data might be explained by the lower expression level and the length of the transcript. One can not discard the possibility that deeper long-read coverage might uncover full-length, unspliced, *Copia* transcripts. It is important to stress that by using only short-reads it is nearly impossible to determine which *Copia* sequence is being expressed as the vast majority of short reads map to multiple locations with the exact same alignment score. With short-reads, at least one full-length *Copia* insertion is expressed but its specific location remains unknown. Furthermore, if we restrict the analysis to primary alignments (i.e. a randomly chosen alignment in the case of multiple mapping), then the coverage of the intronic sequence decreases and it is no longer clear if the insertion produces both spliced and unspliced transcripts (Figure S24). Overall, for *Copia*, the long-reads enable the identification of which insertion is being transcribed, and the short-reads enable the detection of the presence of the two splice variants. Some multi-mapping long reads could support the presence of the unspliced transcript because they partially map to Copia intron, but we cannot know from which insertion they were transcribed (Figure S25). Finally, spliced transcripts are unable to produce the complete transposition machinery as they lack the reverse transcriptase enzyme and are only able to produce the gag protein.

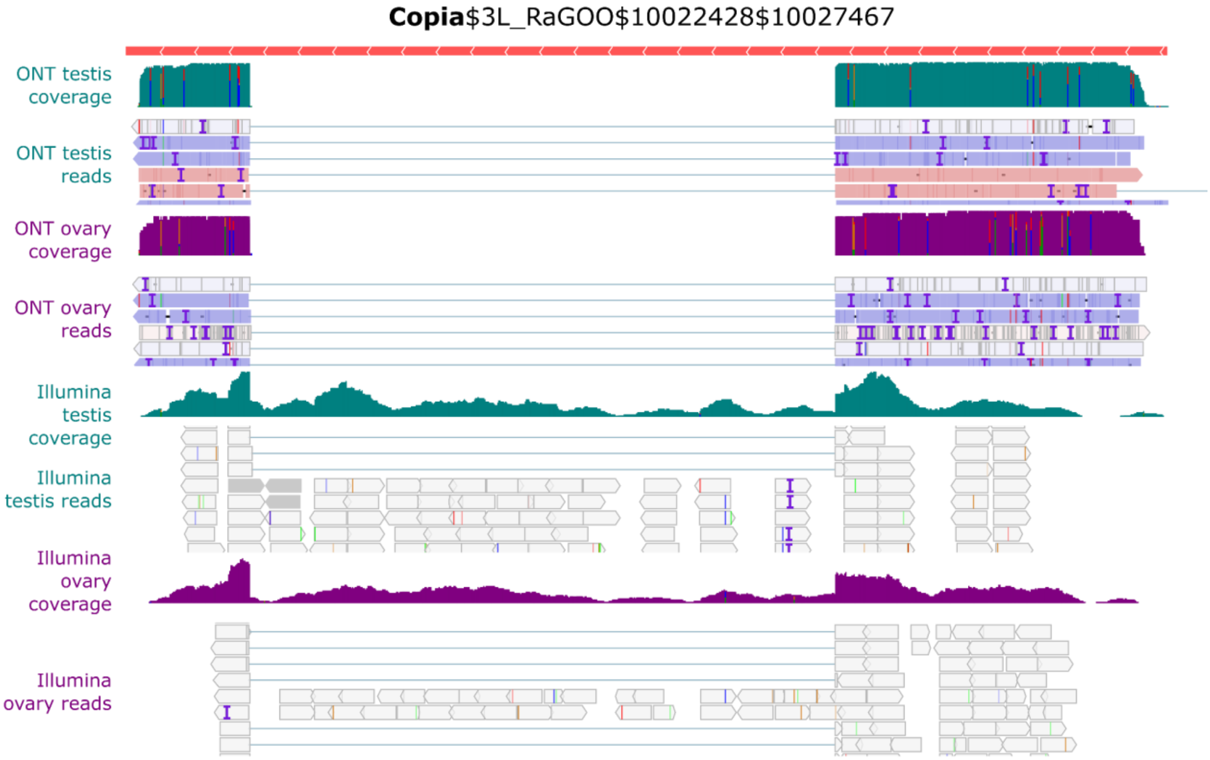

Figure 8. Example of *Copia* splicing. IGV screenshot shows spliced transcripts using long and short-read datasets. In green, testis coverage and an excerpt of mapped reads, in purple the same information for ovaries. In the excerpt of mapped reads, white rectangles correspond to multi-mapping reads. Dmgoth101 repeat and gene tracks are also shown and more information on the annotation can be seen in the material and methods section.

Conclusion

Long-read sequencing remains a major progress in the study of repeat transcription. Here we demonstrated the feasibility of assigning long reads to specific copies. In addition, quantification of TE expression with long-read sequencing is similar to short-read analysis, suggesting not only one could recover copy-specific information but also perform quantitative and differential expression analysis.

The genome of *D. melanogaster* contains many functional full-length copies but only a couple of such copies produce full-length transcripts in gonads. Given TEs are tightly controlled in the germline, one can wonder how many full-length copies might be expressed in somatic tissues. It is also important to stress that, to our knowledge, this is the first comparison of the expression of TEs between testes and ovaries, and we uncover a different TE transcriptional landscape regarding TE subclasses, using both short-reads and long-reads. Furthermore, in many instances, we see that TEs are spliced, independently of their structure or class. While some of these introns had been reported in the literature 30 years ago, the relevance and prevalence of these spliced transcripts have not always been investigated. Long-read sequencing could facilitate the exhaustive inventory of all spliceforms, in particular for recent TEs, where short reads are harder to use due to multiple mapping. A difficulty that remains when assessing if the intron of a particular TE insertion has really been spliced is the possibility that there exists a retroposed copy of a spliced version of this TE elsewhere in a non-assembled part of the genome. Here, taking advantage of the availability of raw genomic Nanopore reads for the same dataset (ERR4351625), we could verify that this was not the case for Copia, the youngest expressed element in our dataset. In practice, we mapped the genomic reads to both Copia and a spliced version of Copia and found no genomic read mapping to the spliced version.

Finally, it is important to note that we did not recover TE transcripts longer than ~4.5 Kb. While the detection of rare transcripts might indeed pose a problem to most sequencing chemistries, it would be important to verify if long transcripts necessitate different RNA extraction methods for ONT sequencing. For instance, the distribution of cDNA used here for ONT library construction reflects the distribution of reads, with a low number of cDNAs longer than 3.5 Kb (Figure S1).

Acknowledgements

This work was performed using the computing facilities of the LBBE/PRABI. We thank Josefa Gonzalez laboratory for discussing our preprint in their journal club and suggesting modifications to the manuscript.

Data, scripts, code, and supplementary information availability

Data are available online at the BioProject PRJNA956863 (ONT long-reads), PRJNA981353 (SRX20759708, SRX20759707, testes short-reads), PRJNA795668 (SRX13669659 and SRX13669658, ovaries short-reads). Scripts are available at https://gitlab.inria.fr/erable/te_long_read. Processed data (.bam files) are available at https://zenodo.org/records/10277511.

Conflict of interest disclosure

The authors declare that they have no financial conflicts of interest in relation to the content of the article.

Funding

This work was supported by French National research agency (ANR project ANR-16-CE23-0001 'ASTER and ANR-20-CE02-0015 LongevitY), along with UDL-FAPESP2020.

References

Adrion JR, Song MJ, Schrider DR, Hahn MW, Schaack S (2017) Genome-Wide Estimates of Transposable Element Insertion and Deletion Rates in *Drosophila Melanogaster*. *Genome Biology and Evolution*, **9**, 1329–1340. https://doi.org/10.1093/gbe/evx050

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Babarinde IA, Ma G, Li Y, Deng B, Luo Z, Liu H, Abdul MM, Ward C, Chen M, Fu X, Shi L, Duttlinger M, He J, Sun L, Li W, Zhuang Q, Tong G, Frampton J, Cazier J-B, Chen J, Jauch R, Esteban MA, Hutchins AP (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. https://doi.org/10.1093/nar/gkab710

Bailly-Bechet M, Haudry A, Lerat E (2014) “One code to find them all”: a perl tool to conveniently parse RepeatMasker output files. *Mobile DNA*, **5**, 13. https://doi.org/10.1186/1759-8753-5-13

Belancio VP, Hedges DJ, Deininger P (2006) LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic Acids Research*, **34**, 1512–1521. https://doi.org/10.1093/nar/gkl027

Bendall ML, Mulder M de, Iñiguez LP, Lecanda-Sánchez A, Pérez-Losada M, Ostrowski MA, Jones RB, Mulder LCF, Reyes-Terán G, Crandall KA, Ormsby CE, Nixon DF (2019) Telescope: Characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLOS Computational Biology*, **15**, e1006453. https://doi.org/10.1371/journal.pcbi.1006453

Berrens RV, Yang A, Laumer CE, Lun ATL, Bieberich F, Law C-T, Lan G, Imaz M, Bowness JS, Brockdorff N, Gaffney DJ, Marioni JC (2022) Locus-specific expression of transposable elements in single cells with CELLO-seq. *Nature Biotechnology*, **40**, 546–554. https://doi.org/10.1038/s41587-021-01093-1

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, Mager DL, Feschotte C (2018) Ten things you should know about transposable elements. *Genome Biology*, **19**, 199. https://doi.org/10.1186/s13059-018-1577-z

Brierley C, Flavell AJ (1990) The retrotransposon copia controls the relative levels of its gene products post-transcriptionally by differential expression from its two major mRNAs. *Nucleic Acids Research*, **18**, 2947–2951. https://doi.org/10.1093/nar/18.10.2947

Dai Z, Ren J, Tong X, Hu H, Lu K, Dai F, Han M-J (2021) The Landscapes of Full-Length Transcripts and Splice Isoforms as Well as Transposons Exonization in the Lepidopteran Model System, Bombyx mori. *Frontiers in Genetics*, **12**, 704162. https://doi.org/10.3389/fgene.2021.704162

De Coster W, D’Hert S, Schultz DT, Cruts M, Van Broeckhoven C (2018) NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, **34**, 2666–2669. https://doi.org/10.1093/bioinformatics/bty149

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

Ewing AD, Smits N, Sanchez-Luque FJ, Faivre J, Brennan PM, Richardson SR, Cheetham SW, Faulkner GJ (2020) Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. *Molecular Cell*, **80**, 915-928.e5. https://doi.org/10.1016/j.molcel.2020.10.024

Fablet M, Salces-Ortiz J, Jacquet A, Menezes BF, Dechaud C, Veber P, Rebollo R, Vieira C (2023) A Quantitative, Genome-Wide Analysis in Drosophila Reveals Transposable Elements’ Influence on Gene Expression Is Species-Specific. *Genome Biology and Evolution*, **15**, evad160. https://doi.org/10.1093/gbe/evad160

Fabry MH, Falconio FA, Joud F, Lythgoe EK, Czech B, Hannon GJ (2021) Maternally inherited piRNAs direct transient heterochromatin formation at active transposons during early Drosophila embryogenesis. *eLife*, **10**, e68573. https://doi.org/10.7554/eLife.68573

Jiang F, Zhang J, Liu Q, Liu X, Wang H, He J, Kang L (2019) Long-read direct RNA sequencing by 5'-Cap capturing reveals the impact of Piwi on the widespread exonization of transposable elements in locusts. *RNA biology*, **16**, 950–959. https://doi.org/10.1080/15476286.2019.1602437

Jin Y, Hammell M (2018) Analysis of RNA-Seq Data Using TEtranscripts. In: *Transcriptome Data Analysis: Methods and Protocols* Methods in Molecular Biology. (eds Wang Y, Sun M), pp. 153–167. Springer, New York, NY. https://doi.org/10.1007/978-1-4939-7710-9_11

Jin Y, Tam OH, Paniagua E, Hammell M (2015) TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics (Oxford, England)*, **31**, 3593–3599. https://doi.org/10.1093/bioinformatics/btv422

Lanciano S, Cristofari G (2020) Measuring and interpreting transposable element expression. *Nature Reviews. Genetics*, **21**, 721–736. https://doi.org/10.1038/s41576-020-0251-y

Laski FA, Rio DC, Rubin GM (1986) Tissue specificity of Drosophila P element transposition is regulated at the level of mRNA splicing. *Cell*, **44**, 7–19. https://doi.org/10.1016/0092-8674(86)90480-0

Lerat E, Fablet M, Modolo L, Lopez-Maestre H, Vieira C (2017) TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids Research*, **45**, e17. https://doi.org/10.1093/nar/gkw953

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li H, Janssens J, De Waegeneer M, Kolluru SS, Davie K, Gardeux V, Saelens W, David FPA, Brbić M, Spanier K, Leskovec J, McLaughlin CN, Xie Q, Jones RC, Brueckner K, Shim J, Tattikota SG, Schnorrer F, Rust K, Nystul TG, Carvalho-Santos Z, Ribeiro C, Pal S, Mahadevaraju S, Przytycka TM, Allen AM, Goodwin SF, Berry CW, Fuller MT, White-Cooper H, Matunis EL, DiNardo S, Galenza A, O'Brien LE, Dow JAT, FCA Consortium, Jasper H, Oliver B, Perrimon N, Deplancke B, Quake SR, Luo L, Aerts S (2022) Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. *Science*, **375**, eabk2432. https://doi.org/10.1126/science.abk2432

Miller K, Rosenbaum J, Zbrzezna V, Pogo AO (1989) The nucleotide sequence of Drosophila melanogaster copia-specific 2.1-kb mRNA. *Nucleic Acids Research*, **17**, 2134. https://doi.org/10.1093/nar/17.5.2134

Mohamed M, Dang NT-M, Ogyama Y, Burlet N, Mugat B, Boulesteix M, Mérel V, Veber P, Salces-Ortiz J, Severac D, Pélisson A, Vieira C, Sabot F, Fablet M, Chambeyron S (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. https://doi.org/10.3390/cells9081776

Panda K, Slotkin RK (2020) Long-Read cDNA Sequencing Enables a "Gene-Like" Transcript Annotation of Transposable Elements. *The Plant Cell*, **32**, 2687–2698.

https://doi.org/10.1105/tpc.20.00115

Pélisson A, Song SU, Prud'homme N, Smith PA, Bucheton A, Corces VG (1994) Gypsy transposition correlates with the production of a retroviral envelope-like protein under the tissue-specific control of the Drosophila flamenco gene. *The EMBO journal*, **13**, 4401–4411. https://doi.org/10.1002/j.1460-2075.1994.tb06760.x

Rech GE, Radío S, Guirao-Rico S, Aguilera L, Horvath V, Green L, Lindstadt H, Jamilloux V, Quesneville H, González J (2022) Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in *Drosophila*. *Nature Communications*, **13**, 1948. https://doi.org/10.1038/s41467-022-29518-8

Sessegolo C, Cruaud C, Da Silva C, Cologne A, Dubarry M, Derrien T, Lacroix V, Aury J-M (2019) Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. *Scientific Reports*, **9**, 14908. https://doi.org/10.1038/s41598-019-51470-9

Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, Imamichi T, 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. https://doi.org/10.1093/nar/gkac194

Shumate A, Salzberg SL (2021) Liftoff: accurate mapping of gene annotations. *Bioinformatics (Oxford, England)*, **37**, 1639–1643. https://doi.org/10.1093/bioinformatics/btaa1016

Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, **8**, 272–285. https://doi.org/10.1038/nrg2072

Teixeira FK, Okuniewska M, Malone CD, Coux R-X, Rio DC, Lehmann R (2017) piRNA-mediated regulation of transposon alternative splicing in the soma and germ line. *Nature*, **552**, 268–272. https://doi.org/10.1038/nature25018

Tudor M, Lobocka M, Goodell M, Pettitt J, O'Hare K (1992) The pogo transposable element family of *Drosophila melanogaster*. *Molecular & general genetics: MGG*, **232**, 126–134. https://doi.org/10.1007/BF00299145

White R, Pellefigues C, Ronchese F, Lamiable O, Eccles D (2017) Investigation of chimeric reads using the MinION. *F1000Research*, **6**, 631. https://doi.org/10.12688/f1000research.11547.2

Yang WR, Ardeljan D, Pacyna CN, Payer LM, Burns KH (2019) SQuIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic Acids Research*, **47**, e27. https://doi.org/10.1093/nar/gky1301

Yoshioka K, Honma H, Zushi M, Kondo S, Togashi S, Miyake T, Shiba T (1990) Virus-like particle formation of *Drosophila* copia through autocatalytic processing. *The EMBO journal*, **9**, 535–541. https://doi.org/10.1002/j.1460-2075.1990.tb08140.x