Peer Community In Genomics

RESEARCH ARTICLE
Open Access
Open Data
Open Code
Open Peer-Review

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

Rita Rebollo[1], 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[2,11], Cristina Vieira*[2,3] and Vincent Lacroix*[2,3]

[1]Univ Lyon, INRAE, INSA-Lyon, BF2I, UMR 203, 69621 Villeurbanne, France.
[2]Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, CNRS, UMR5558, Villeurbanne, Rhone-Alpes, 69100, France
[3] ERABLE team, Inria, Grenoble Rhône-Alpes, Villeurbanne, France
[4] DIADE unit, Univ Montpellier, Cirad, IRD, F-34394 Montpellier Cedex 5, France
[5] Institut de Génomique Fonctionnelle de Lyon (IGFL), UMR5242 CNRS, 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.
[8] Human Genetics, McGill University, Montreal, QC, Canada
[9]Canadian Centre for Computational Genomics, McGill University, Montreal, QC, Canada
[10]McGill Genome Centre, Montreal, QC, Canada.
[11] Institut Universitaire de France (IUF), Paris, Île-de-France , F-75231, France.

*Corresponding author

Cite as
XXXXthe "cite_as" we sent you. If too long to fit, use "et al."

Correspondence
vincent.lacroix@univ-lyon1.fr
cristina.vieira@univ-lyon1.fr

Recommender
XXXXthe First name and Family name of the recommender

Reviewers
XXXXthe First names and Family names of the reviewers we sent you

This version of the article has been peer-reviewed and recommended by
Peer Community in Genomics
XXXXthe_doi_we_sent_you

Abstract
Transposable elements (TEs) are repeated DNA sequences potentially able to move throughout the genome. In addition to their inherent mutagenic effects, TEs are also able to disrupt nearby genes by donating their intrinsic regulatory sequences, as 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 RNAseq reads to estimate TE expression at the family level, while using specific algorithms to try 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 RNAseq can be used to identify and quantify 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 (> 5kb) are not well captured. Most expressed TE insertions correspond to copies that have lost their ability to transpose. Some TEs exhibit interesting patterns, such as intronic POGO sequences, which are expressed in ovaries and not testes, in opposition to the host gene. In many instances, we see that TEs can be spliced, independent of their structure or class. Overall, this first comparison of TEs between testis and ovaries uncovers differences in their transcriptional landscape, at the subclass and insertion level.
Keywords: long-read sequencing, transposable elements, regulation, RNAseq, full-length cDNA, ONT

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 env-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 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. Gene annotation was performed as described in (Fablet et al., 2022). Briefly, we used LiftOff (Shumate & Salzberg, 2021) on FlyBase *D. melanogaster* annotation (dmel-all-r6.46.gtf.gz), with the option -flank 0.2. 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* TE copies and then used OneCodeToFindThemAll to merge LTR and internal parts of TE in one unique feature. Alignments of copies to their consensus were performed in SnapGene.

Drosophila *rearing and total RNA extraction*

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. 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.

Long read RNA-seq and analysis

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. This protocol is highly selective for mRNAs that are both capped and polyadenylated and allows their amplification. 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. Our 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 (Oxford Nanopore Technologies, 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 testis that passed the default quality filter (>Q7). Quality control was performed with NanoPlot v1.39.0 (De Coster et al., 2018). The median read length was 1.4 Kb, and the median quality was 8.4 (*i.e.* one error every 8 bases, Figure S1). Data are available online at the BioProject PRJNA956863. Reads were mapped to the dmgoth101 genome using minimap2 (version 2.17-r974-dirty) (Li, 2018) using the splice preset parameter: "minimap2 -ax splice FC30.fastq dmgoth101_assembl_chrom.fasta -o FC30.bam". Ninety percent of reads were mapped to the genome (Table S1). The vast majority (99%) mapped to a unique location (i.e. had no secondary alignment reported by minimap2). Out of the reads with several locations reported, 90% had a unique best location, i.e. the alignment score (AS flag in the SAM format) was higher for one of the locations. 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, we 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 S2.

We also noticed that some reads were only partially mapped to the genome. In practice the query coverage distribution is bimodal (Figure S3), 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 S4). 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.

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. 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 assigned the read to the feature with the best coverage. The rationale for introducing this filter is best explained with examples. Figure S5 corresponds to 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 S6 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.

Most commonly, several features may be partially covered by a read and a read may extend beyond each of these features. Both the subject and the query coverage need to be modelled. To address this, we consider genomic intervals as sets of positions. For each pair read-feature, we compute the symmetric difference between the two intervals, and we assign the read to the feature with the smallest symmetric difference (Figure S7). This situation occurs frequently and ignoring it yields an overestimation of TE expression. The impact of each filter is given in Figure S8. After all filters are applied, there are 1 322 reads in ovaries and 8 219 reads in testes that are assigned to TE copies (Table S2).

Short-read RNA-seq and analysis

RNA extraction and short-read sequencing were retrieved from (Fablet et al., 2022), at the NCBI BioProject database PRJNA795668. Briefly, RNA was extracted from 70 testis 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. Libraries were sequenced on Illumina HiSeq 3000 with paired-end 150 nt reads. Short-read analysis was performed as in (Fablet et al., 2022), 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.

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, Figure 1A). Sequencing yielded between ~1 to ~3 million reads per tissue, ranging from 104 to 12 584 bp (median read length ~1.4 Kb, Table S1). Reads were subsequently mapped to the strain-specific genome assembly (Mohamed et al., 2020) using the LR aligner Minimap2 (version 2.17-r974-dirty) (Li, 2018). Ninety percent of reads could be mapped to the genome (Table S1). The vast majority (99%) mapped to a unique location (*i.e.* had no secondary alignment), and 99.9% mapped to a unique best location (see Methods).

In order to validate the long-read RNAseq approach, we first determined the read coverage of all expressed genes, and show that for both ovaries and testes, the vast majority (99.2%) harbour at least one read covering more than 90% of their sequence, regardless of their longest transcript length. Besides, there are very few reads which correspond to partial transcripts, as 98.4% of reads correspond to well-covered transcripts (>90% coverage) (Figure 1B). This clearly shows that the TeloPrime protocol was successful in capturing full-length transcripts. Besides, the transcriptomes obtained are enriched in typical germline ontology terms, such as “sperm individualization and competition” for testis, and “oogenesis” for ovaries (Figure 1C). 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 for genes harbouring a single isoform (rho=0.78, R=0.74, Figure 1D). These correlation coefficients are in agreement with those obtained when comparing direct RNAseq *vs* Illumina Truseq sequencing (Sessegolo et al., 2019).

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 annotated TE 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 and Figure S5-6) was selected. Overall, after applying these filters, 1 322 reads in ovaries and 8 219 reads in testes were assigned to TE copies (Table S2). Out of these, 37-48% overlapped only one TE, 8-18% overlapped with multiple TEs, 16-21% were located in the intron of a gene, and 18-34% overlapped with a gene exon (but were assigned to the TE, not the gene) (Figure 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 copies with the length of all mapped reads (Figure 1F). While TE copies range from a few base pairs to ~15 Kb, 75% of annotated copies 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 2 Kb. On the other hand, reads mapping to genes may reach 5 Kb. The absence of very long reads indicates that either very long mRNAs are absent from the sample, 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 Nanopore Teloprime for short (<3 Kb), long (3 Kb-5 Kb) and very long transcripts (>5 Kb), and obtained the following Spearman correlations of 0.79 (n=3 605), 0.54 (n=378) and 0.38 (n=130), respectively (Figure 1D). We concluded that indeed long and very long transcripts are a minority (9% and 3% of all transcripts, respectively) but the TeloPrime protocol underestimates their abundance.

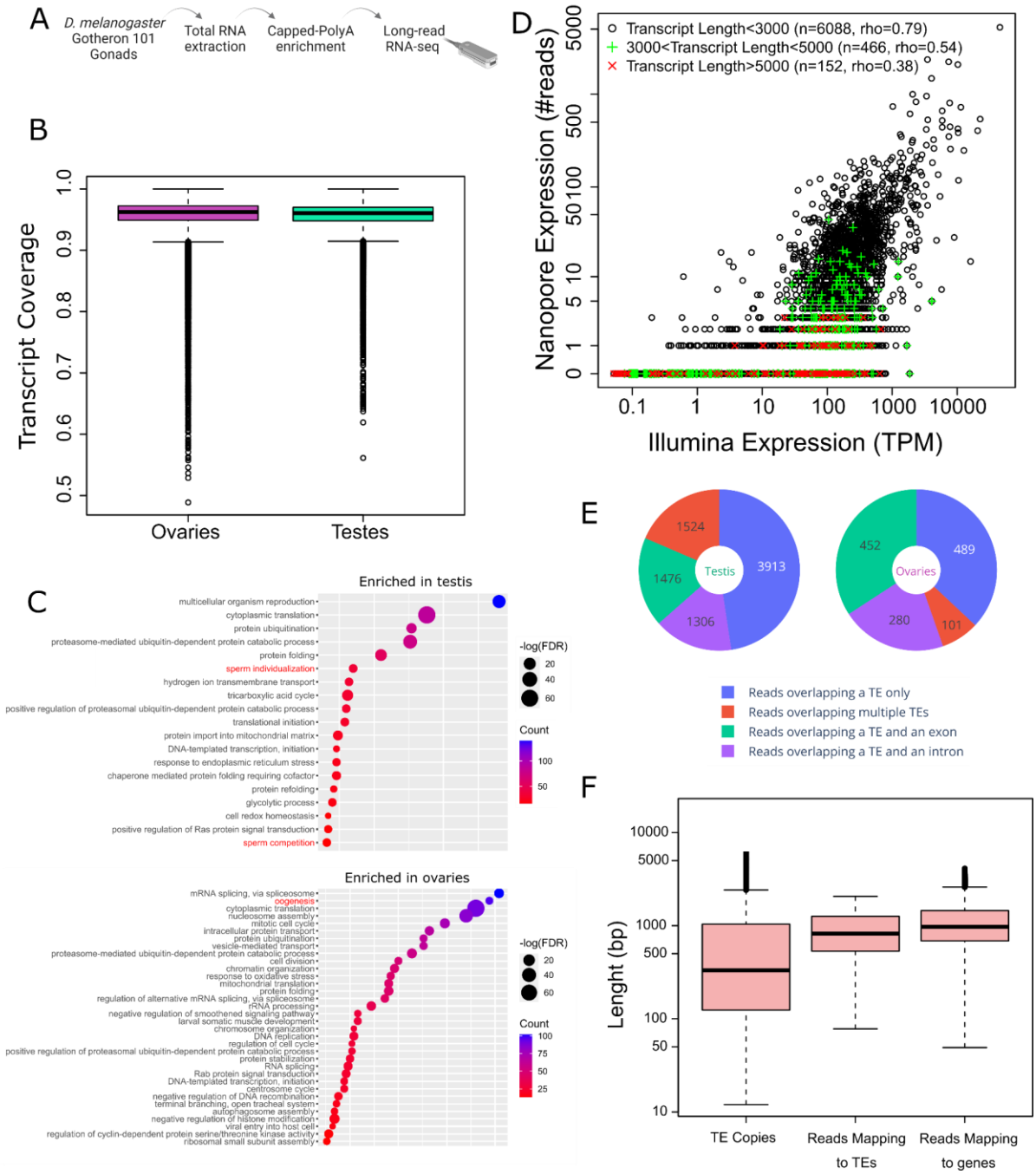

Figure 1: Long-read RNAseq of *Drosophila melanogaster* ovaries and testes. A. Experimental layout of TeloPrime approach. B. Transcript coverage by long-read RNAseq in ovary and testes. The majority of transcripts recovered are full-length. C. Gene Ontology enrichment analysis of long-read transcriptomes of testes and ovaries. In red are enriched terms related to male and female gonadal development. D. Gene expression quantification using Illumina and ONT sequencing. Each dot is a gene with a single annotated isoform. Transcripts longer than 5kb tend to be undersampled using Teloprime. E. Number of reads mapping to TEs separated into four categories whether they overlap a TE only, multiple TEs, a TE and an exon, a TE and an intron. F. TE copy and read length distribution. Reads mapping to TEs encompass most TE copy length but lack transcripts longer than 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% (8 219/2 925 554) and 0.11% (1 322/1 236 000) of long reads aligned to TE copies in testes and ovaries respectively, in agreement with the previous observations using short-read sequencing (Fablet et al., 2022). We identified 130 TE families capable of producing capped-polyadenylated mRNAs (Table S2), of which 69 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 (99/130), the transcriptional landscape between ovaries and testes is quite different (Figure 2A). While LTRs 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 2A). The transcriptional landscape within TE subclasses between males and females 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 males and females: TcMar-Pogo is the most expressed DNA superfamily in females, while RC elements are abundantly transcribed in males.

Globally, TE families show higher long-read counts in males compared to females (Figure 2B), both because male samples were more deeply sequenced (2.3 times more), and because the proportion of reads that map to TEs is higher in males (0.28% *vs* 0.1%). HETA (I-Jockey) and DNAREP1_D (RC) are the top two families in male TE transcript counts, with 734 and 722 long-read transcripts respectively. In females, Pogo (TcMar-Pogo) is the TE family amounting the most transcripts with 224 long-reads (while only 34 in males), followed by DNAREP1_D (RC) with 142 and Copia (Copia) with 136 long reads (but both families with higher transcript counts in males). There are only three TE families that are specific to ovaries, BARI_Dm (TcMar-Tc1 - DNA), Gypsy7 (Gypsy - LTR) and Helena (I-Jockey - LINE), but they all show only one single long-read suggesting their expression is low. There are 28 families specifically transcribed in testes compared to ovaries, five DNA elements (Transib-N1_DM (CMC-Transib), NOF_FB (MULE-NOF), Looper1_Dm (PiggyBac), and two TcMar-Tc1 (Bari1 and Minos)), 12 LINE elements (9 I-Jockey, two R1, and one R1-LOA, see Table S2 for details), and 11 LTR families (one Copia, four Gypsy, two Pao and four unknown families), ranging from one to 76 long-reads per TE family. Finally, 9 TE families show no long-read mapping in either tissue. Collectively, long-read sequencing is able to discriminate between female and male TE transcriptional landscapes and suggests retrotransposons might be strongly and specifically expressed in males compared to females.

Short-read RNA sequencing of ovaries and testes, followed by estimation of TE family expression with TEcount ((Lerat et al., 2017) - TE expression estimation *per* TE family, see material and methods for more information) recapitulates the long-read RNA sequencing profiles (Figure 2C). 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=0.48, rho=0.79) than for ovaries (r=0.36, rho=0.59), where the coverage is weaker. Indeed, the total contribution of TE to the transcriptome is weaker for ovaries (0.11% Vs 0.28% for males) and the sequencing was more shallow.

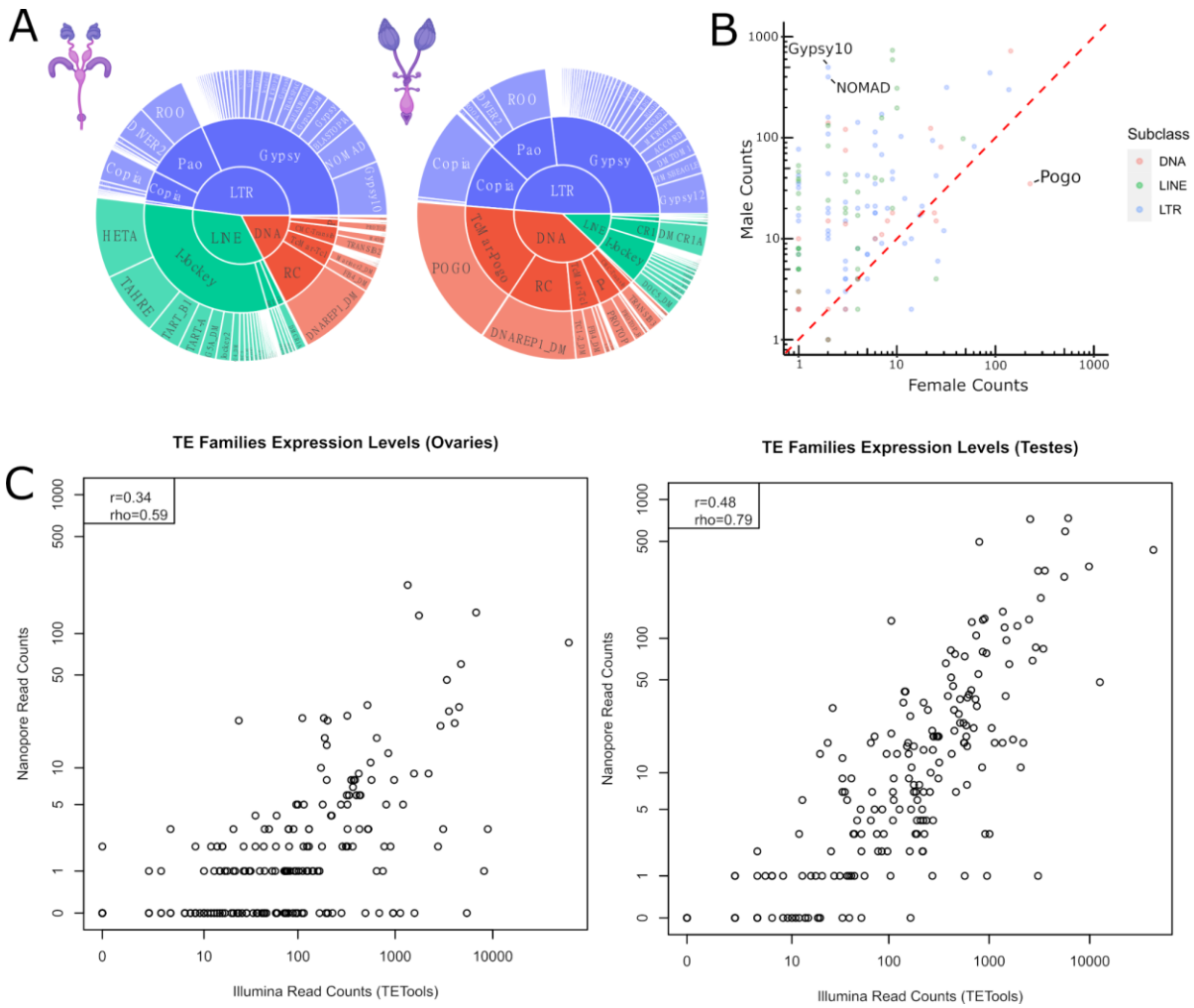

Figure 2: Transposable element transcriptional landscape in ovaries and testes of *Drosophila melanogaster*. A. Global TE transcriptional landscape using ONT long-read sequencing. The outer ring, middle ring and inside circle represent respectively family, superfamily and subclass of TE. The area in the circle is proportional to the rate of expression. The left chart depicts ovary data sets while the right chart depicts testis. B. Comparison of TE expression ratio between males and females ONT long-read datasets. C. Comparison between Illumina and ONT datasets for estimating the expression levels at the TE family level. Each dot is a TE family. TEcount is used for short reads. The correlation between quantifications by both technologies is higher for testes (rho=0.79) than for ovaries, (rho=0.59) 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. We uncovered 1 115 and 448 TE copies harbouring at least one long-read transcript unambiguously in testis and ovaries respectively. 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 (Figure 3A). However, HETA (I-Jockey) copies are nearly all expressed in males (14/15 annotated copies), while in females, POGO harbors 17 copies producing transcripts out of 25. 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 females and males (73 and 170 respectively). Finally, in testes, out of the 1 115 insertions with at least one mapped read, 165 had in fact more than 10 mapped reads. In ovaries, out of the 448 insertions with at least one mapped read, 26 had more than 10 mapped reads.

In addition, while many TE copies within a family are indeed able to produce transcripts, there are significant differences in copy expression rate. For instance, in ovaries, where POGO has the highest number of long-reads (224), an insertion of 1 352 bp in the X chromosome (X_RaGOO:21,861,530-21,866,880) accumulates nearly 30% of the family total read count (Figure 3B and C). This specific POGO insertion is located in an intron of a gene (CG12061), at 370 bp from the transcription start site, and anti-sense to the gene. 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) (Figure 3D). The same pattern is observed for the second most expressed POGO insertion, also located in the intron of a gene, expressed in testes and not in ovaries (Figure S9). The other expressed insertions of POGO are located in intergenic regions (Figure S10).

Finally, using short-read sequencing and a tool developed to estimate single-copy expression (Squire (Yang et al., 2019)), we compared the overall transcriptional landscape between short and long-reads, as we previously performed for family and superfamily levels, but here at the insertion level (Figure S11). There was a poor correlation with the Nanopore estimations (rho=0.22, r=0.17 for ovaries and rho=0.38, r=0.33 for testes). Examination of instances where the two techniques differed most, shows that Squire tends to overestimate the expression of TE insertions overlapping genes (Figure S11). At the family level, the quantifications obtained by Squire were comparable to the ones obtained with TEcount.

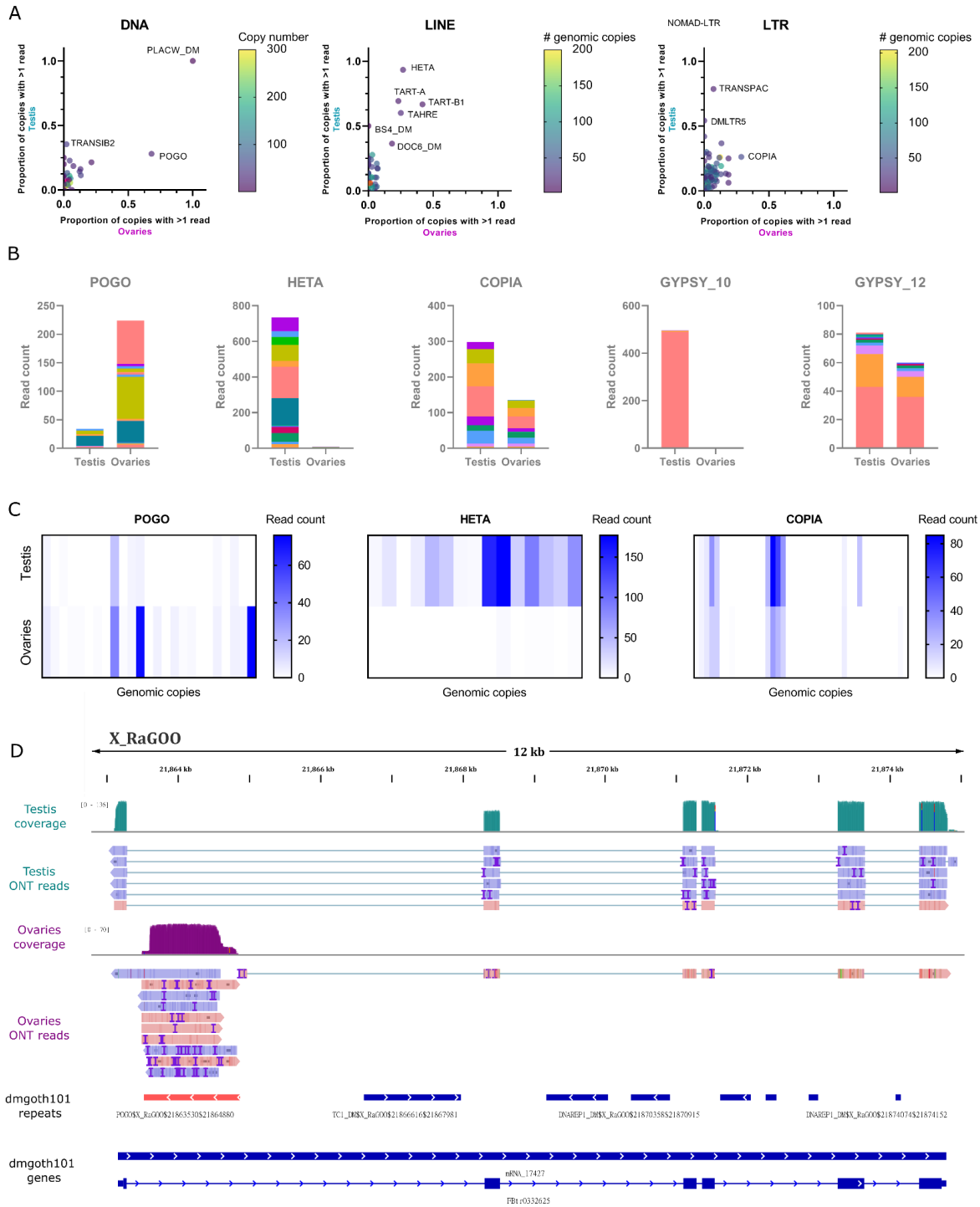

Figure 3: Transcription of transposable element copies. A. Frequency of transcribed copies (read > 1) within TE subfamilies in testis and ovaries, along with genomic copy number (color bar, 1 to 200 (LINE/LTR) or 300 (DNA) copies). All TE families harbouring more than 200 (LINE) or 300 (DNA) copies are depicted in pink. For DNA elements, DNAREP1_DM harbours 2 472 copies, and the LINE family DMCR1A has 534 copies. Most TE subfamilies have only a couple of copies producing transcripts, while the majority of HETA copies are expressed in testis for instance (middle panel). B. Distribution of read count per copy for POGO, HETA, COPIA, GYPSY_10 and GYPSY_12 showing the overall expression of specific copies within aTE family (Table S3 and S4). Copies are represented by different colors within a bar graph. C. Heatmap of copy read count per TE subfamily in testis and ovaries. POGO, HETA and COPIA are depicted. D. Simplified IGV screenshot of a Pogo copy (X_RaGOO:21,861,530-21,866,88, in pink). In green, testis coverage and an excerpt of 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.

Full-length transposable element copies are rarely transcribed

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 insertion itself might have accumulated mutations, insertions and deletions which makes it unable to transpose. To assess this, we computed the query coverage of the reads w.r.t. the insertion they correspond to. We find that one-third of the insertions have at least 80% query coverage (Figure 4A). However, out of these insertions, only a few of them are close in length to a functional full-length sequence. Indeed, a close analysis of POGO, one of the most expressed TE subfamilies in ovaries, shows only two potential complete copies expressed, and both at low levels (6 to 4 reads in ovaries and 2 to 3 in testis, Figure 4B). Instead, the most expressed POGO copies in ovaries (73 and 76 long reads), and in testis (18 long reads) are internally deleted. In order to search for potentially functional, expressed copies, we filtered copies that produced at least five long reads in ovaries, and ten in testis (due to coverage differences), and covered at least 80% of their consensus sequences. In ovaries these filters correspond to seven insertions: one POGO, five Copia insertions and one MAX. In testis, there are nine potentially functional insertions: five Copia insertions, three Mariner2, and one HETA (Figure 4B). 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 copies are indeed transcribed in testis and ovaries.

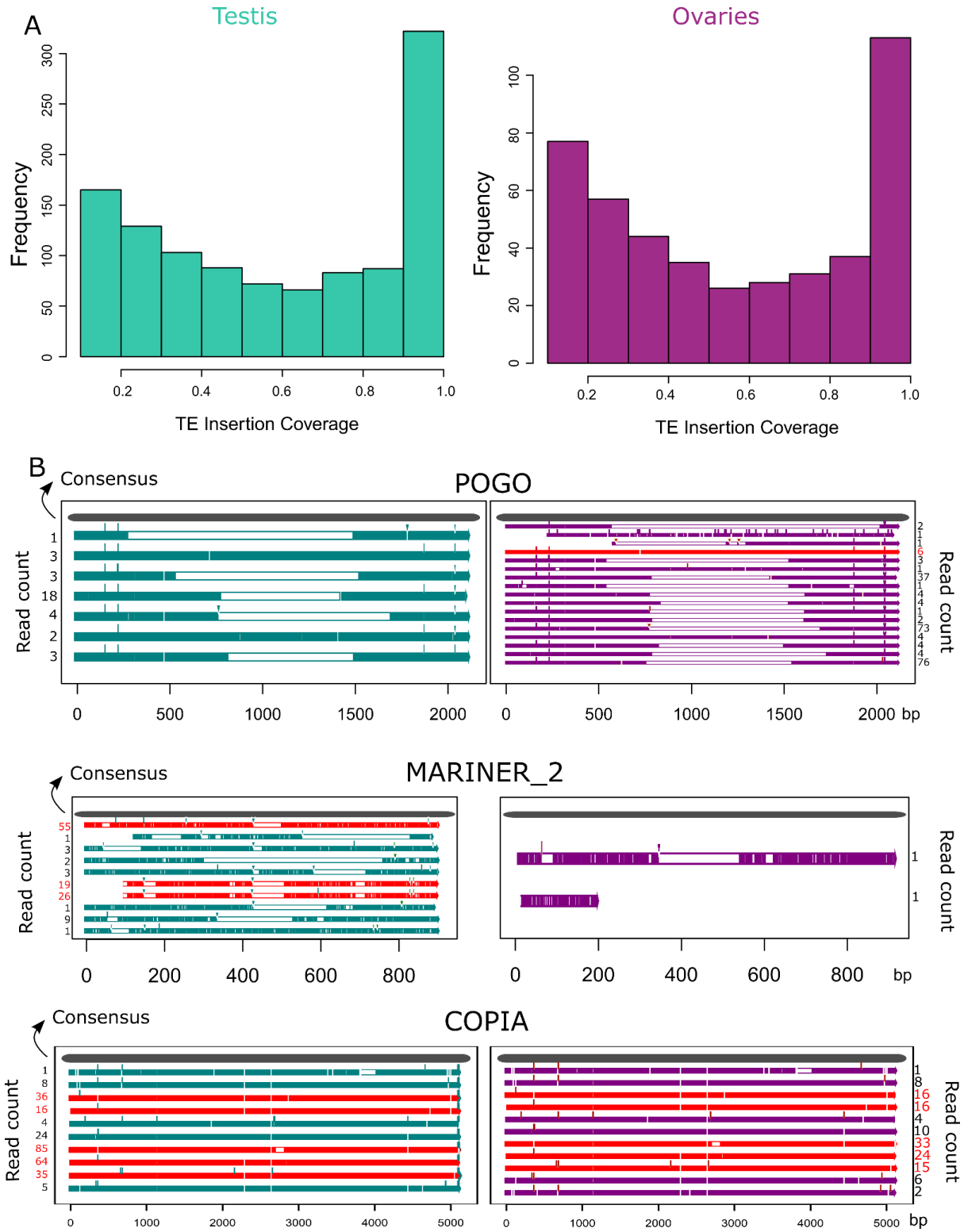

Figure 4. TE transcripts stem mostly from deleted or truncated copies. A. Coverage of ONT reads on TE insertions. One-third of copies are covered for at least 80% of their length. B. Alignment of transcribed copies against their consensus for three major subfamilies: POGO, Mariner2 and COPIA, all harboring potentially full-length expressed copies. Read count is shown for each copy and full-length copies (> 80% coverage) and their long-read count are depicted in red.

Long-read sequencing unveils novel spliced TE isoforms

While most TEs do not harbor introns, there are a couple of exemptions 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, also regulated by piRNAs (Pélisson et al., 1994; Teixeira et al., 2017). 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). Such gaps can correspond to either deletions or reveal the presence of an intron sequence. Given the mapping strategy and most precisely the filtering steps providing only uniquely mapping reads, the gaps correspond most probably to introns. Indeed, a read harboring a deletion, would have mapped elsewhere in the genome with a higher mapping score. While most TE copies harbor intronless transcripts (visible in Figure 5, as circles aligned at 0 in the X-axis), there are a significant number of TE transcripts, both in testis and ovaries, that harbor potentially spliced transcripts.

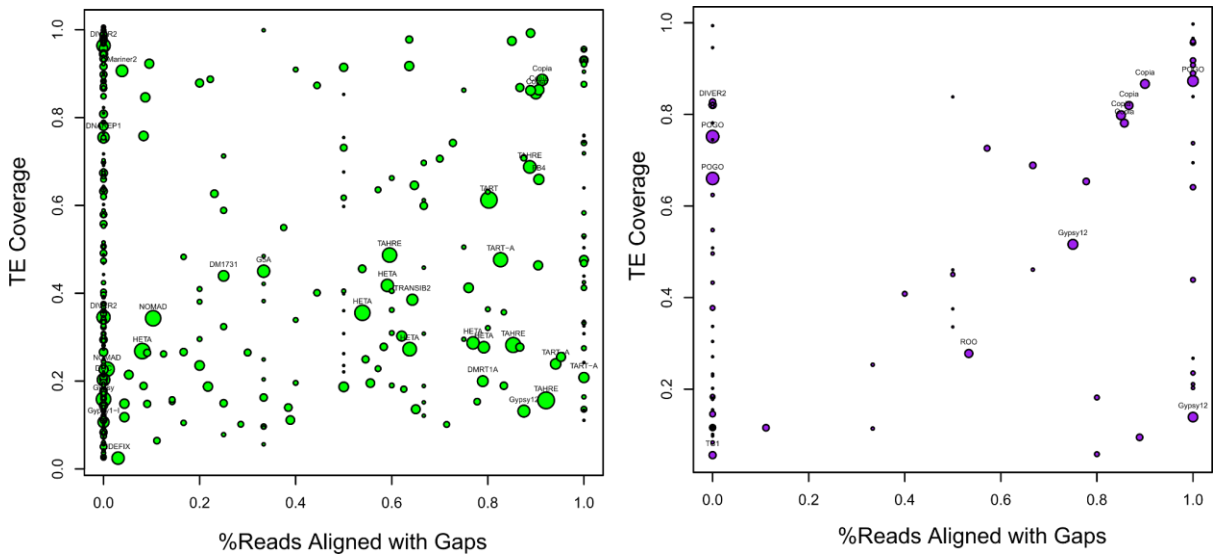

Figure 5. TE spliced transcripts are ubiquitous. Each circle depicts a TE insertion and their size is proportional to the expression level of the insertion. The TE family name is written for highly expressed insertions (>30 reads in testis, >10 reads in ovaries). The X-axis represents the proportion 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.

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 17 out of 25 copies producing capped poly-A transcripts corresponding to 224 long-reads, while only 7 expressed copies with a total of 34 long-reads are observed in testes. While we previously noted that only two full-length copies are transcribed in ovaries (and in testis also but albeit with a lower number of reads), there are many truncated or deleted copies that are transcribed (Figure 4B). 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 6A). 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. POGOX_RaGOO21863530$21864880 (Figure 3D) also contains a large genomic deletion, but which overlaps the intron, which explains why we do not see spliced reads for this insertion.

Despite the presence of full-length Copia insertions in the genome, only spliced transcripts were uncovered in the long-read sequencing (Figure 5 and 6B). In contrast, with Illumina short-reads, we see both spliced and unspliced transcripts (Figure 6B). The full-length 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 S12). 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.

While other examples of spliced transcripts are observed for different families that we have manually curated (FB4, TART-A, TAHRE and others), one particular case involves the 1731 retrotransposon. Indeed, the copy DM1731_LTR$3L_RaGOO$24209333$24213898, encoding a GAG and POL open reading frames, produces a 1.8 kb internal transcript, corresponding to a truncated GAG protein and a truncated reverse transcriptase, and a spliced 1.3kb transcript, spanning the full 4.5 kb genomic copy but splicing a 3.2 kb intron, and producing the RNAseH protein. Both transcripts are present in testis, while no expression of this copy is observed in ovaries.

Figure 6. Simplified IGV screenshot of spliced TE copies. POGO (A), Copia (B) and 1731 (C) show spliced transcripts using long-read datasets. In green, testis coverage and an excerpt of mapped reads, in purple the same information for ovaries. For panel B, short-read data is also present in addition to ONT reads. 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, which remains the biggest issue in TE expression analysis. 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 testis and ovaries, and we demonstrate a different TE transcriptional landscape regarding TE subclasses. Interestingly, some insertions like POGOX_RaGOO21863530$21864880, located in the intron of a gene, are expressed only in ovaries and seem to have a silencing effect on their host gene. 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.

Finally, it is important to note that we did not recover TE transcripts longer than 2 Kb, despite gene transcripts up to 5 Kb. While the sequencing coverage might indeed play a role in the detection of rare transcripts, it would be important to verify if such long transcripts necessitate different RNA extraction methods. 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.

Acknowledgements

This work was performed using the computing facilities of the LBBE/PRABI.

Data, scripts, code, and supplementary information availability

Data are available online at the BioProject PRJNA956863.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having 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

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

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

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, Noûs C, Rebollo R, Vieira C (2022) A quantitative, genome-wide analysis in Drosophila reveals transposable elements' influence on gene expression is species-specific. https://doi.org/10.1101/2022.01.20.477049

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, 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

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