Answer to reviewers are in blue.

Round #1

---

by Nicolas Pollet, 14 Aug 2023 11:39
Manuscript: **https://doi.org/10.1101/2023.05.27.542554** version 2

Dear Rita Rebollo and colleagues,

Your manuscript entitled "Identification and quantification of transposable element transcripts using Long-Read RNA-seq in Drosophila germline tissues" has been reviewed by three colleagues. Globally the reviews are of high quality and positive, and find that your study has many merits, but there are substantial and major criticisms that you need to address before I can finally decide on whether this preprint can be recommended or not.

Especially, I would like to underscore the availability of the data used in your analysis and the availability of re-usable bioinformatics methods.

With my best wishes,

Nicolas Pollet

Dear Editor, please find attached the point-by-point answer to the reviewer's comments. As requested, the raw data is available online at the BioProject PRJNA956863 (long-read testis and ovaries), PRJNA981353 (short-read testis) and PRJNA795668 (short-read ovaries). The treated data is available at https://zenodo.org/records/10277511. We have also updated the different dependency files (genomes, gene and TE annotations) and scripts to a GitLab (https://gitlab.inria.fr/erable/te_long_read/). A major effort has been made in order to build guidelines to replicate this analysis but also to adapt to other biological systems. Therefore, using GitLab/te_long_reads, anyone can reproduce the analysis presented in this article, and adapt it to its own datasets. We thank all the reviewers for improving the manuscript with constructive and clear comments.

Reviews

Reviewed by anonymous reviewer, 18 Jul 2023 17:42

In the proposed manuscript, Rebollo et al explore the use of the long-read Oxford Nanopore (ONT) sequencing technology to describe qualitatively and quantitatively the transcriptional landscape of transposable elements in the Drosophila gonads. The authors generate, sequence

and analyze long-read cDNA libraries (one replicate from each tissue type), as well as compare the obtained results to previously published short-read datasets.

The manuscript is timely, as we can observe a growing interest in the use of ONT technology for transcriptome analysis, especially in the field of DNA repeats. As such, this work will certainly be useful for many researchers, even more so, that it includes clear explanations of all wet-lab and data analysis approaches. The manuscript exposes clearly the strengths, as well the limitations of the techniques used. The manuscript is also quite unique in its detailed comparison between short- and long-read transcriptomic datasets, which shows how these two technologies can complement each other on different levels.

The manuscript could benefit from the following improvements or clarifications:

Major comments:

Regarding “unique” and “unique best” mapping reads vs multi-mappers:

1- According to the text and the Table S1, the “unique best” mapping reads represent 91% and 99% of the sequenced libraries. This however, relates to total reads, only small percentage of which maps to TEs. Thus, effectively, this percentage is true for genes. What are the fractions of “unique” and “unique best” reads for TE-mapping reads only? Supposedly, these could be different than for single copy genes. How does this compare with short-read libraries? To which extend long-reads reduce or overcome the multimapping issue? This would be interesting to show clearly and it is also important for the downstream analysis.

As suggested by the reviewer, we have searched for the long-read fraction mapping to TEs, which amounts to 1 252 reads, including only 50 (4%) mapping to multiple locations with equal alignment scores in ovaries and for testis, 8 138 long-reads map to TEs, including 224 (2.7%) of multi-mapped reads). Regarding short-read libraries, and using the same method for assigning reads to TEs, except that we do not require the short read to cover at least 10% of the TE as we do for long-reads, we find that there are 47 904 reads assigned to TE, including 11 394 (23.7%) mapping to multiple locations with equal alignment scores in ovaries (for testis 52 263 short reads map to TEs, including 11 701 (22.4%) of multi-mapped reads).

If there is a significant fraction of multi-mapping reads among all TE-mapping reads, are these reads taken under account in transcript abundance quantifications? If present, these multi-mappers should contribute to the family-level transcript estimates (Fig 2), and should be taken under account when quantifying copy-specific expression (Fig 3 and 4). If, for example, for a given TE family, there are significantly more multi-mapping reads than unique best aligners, any conclusions on how all the copies contribute the total transcript levels would be impossible to make. The analysis of expressed vs non-expressed TE copies would still hold true, but only for genomic loci that present enough sequence variation to produce transcripts with unique best alignments.

The authors should clarify this by providing the statistics of multi-mapped reads for TEs, performing any additional analysis if necessary and adjusting their conclusions if required.

As stated above, only a small fraction of long reads mapping to TEs are multi-mapped (4% in ovaries, 2.7% in testes). We agree with the reviewer that such reads should be taken into consideration when comparing TE family expression between long and short reads. We have now included them in the data represented in Figure 2, and in the updated Table S3-S6 with two new columns depicting uniquely and multi-mapping counts. We have added a new figure focusing on multimapped *vs* uniquely mapping reads (Figure 3), and have discussed this issue throughout the document. Comparison with short-read dataset can be appreciated in Figure 2D and Figure S17. In addition to the overall contribution of multi-mapped reads to the TE family expression, one of the concerns of several reviewers was the contribution of specific copies to the TE family expression along with their detection. Figure 4A only shows uniquely mapping reads as previously, but we added Figure S18 with total read counts and Figure 3B with a focus on the TE families harboring the most multi-mapped reads. Again, we have thoroughly discussed this issue throughout the text. In summary, except four TE families (*blastopia*, *transib2, M4DM and Burdock),* long-read sequencing can detect single-copy transcripts.

2 -The above point brings up the question of sequence variation between genomic copies of different expressed TE families, which, in the current version of the manuscript, is not much discussed. Expression of evolutionary younger TEs, with lower sequence divergence, would obviously be more difficult to quantify. This would be particularly relevant for the search of full-length TE transcripts (Fig 4), which would carry less informative sequence variation. The authors could include sequence variation of genomic copies (as they do for sequence variants of transcripts in Fig 5) in their analysis or, minimally, they should comment on the limitations that could be related to the potential lack of such variation. Again, this issue will be less relevant if no (or very few) TE-specific multi-mappers are in fact found in the libraries.

Indeed there are very few TE-specific multi-mappers. They are all reported in Table S5 and Table S6. Table S2 also reports for each family, the number of uniquely mapped reads and the number of multi-mapped reads. For instance, POGO has 7 multi-mapped reads and 206 uniquely mapped reads in ovaries. Copia has 21 multi-mapped reads and 78 uniquely mapped reads in ovaries. Following the addition of multi-mapped reads to the analysis, we have now included in Figure 5, now Figure 6, out of the five TE families depicted, only *copia* has multi-mapped reads that can be appreciated in Figure S23.

Regarding mapping reads to features:

3– A substantial number of TEs is located in intronic sequences. Taking under account how the authors assign reads to features, would intronic TEs in expressed genes be taken under account or omitted? This is not clear. Theoretically, in such cases, both features (the gene and the intronic

TE) could be fully covered. In other words, are TEs belonging to the “intron” category if Fig 1E found only (or primarily) in non-expressed genes?

Yes, TEs located in expressed genes are taken into account. We have clarified this point in the description of the methods lines 199-206. Indeed, if a TE and a gene are fully covered (meaning the read spans the exons and the intronic TE sequence), the read will be assigned to the longest feature, so most probably the gene (all exons covered). This allows for read-through transcripts to be removed from the TE count. However, if a read spans the intronic TE and more intronic sequences, the read will be attributed to the TE. We have also added Figure 1C-E to discuss this point in the manuscript. Figure 4C also shows an example of an intronic TE in an expressed gene.

Table S1 and line 215:

Please add read length statistics (median, N50) separately for both samples. Read length will influence some of the downstream analysis, thus it would be important to indicate it.

The table and text were changed accordingly, line 259.

Additional minor comments:

Methods:

Please specify how much total RNA was used as input for the TeloPrime cDNA amplification.

The information was added to the methods, line 139.

Lines 238-240:

The use of percent ranges (e.g. 37-48%) is misleading, when in fact only two samples are analyzed. Replacing with the two obtained values only, would be more accurate.

Indeed, we have modified the text accordingly (line 284-285)

Fig 1B and D:

Transcript coverage in Fig 1B should be plotted as a function of transcript length, similarly as done for the figure panel 1D.

We have added supplementary figures (Figure S10 and S11) representing the coverage depending on the transcript size. Very long transcripts (>5kb) are indeed less well covered. They are also much less expressed (258 reads in ovaries and 1315 in testes). An example of very long transcript well captured by Illumina and not by Nanopore is now given in Figure S14.

Related to the above point (lines 229-230 and Fig 1D), the text of the results sections should not omit the fact that good correlation is achieved only for short transcripts. Although this detailed

explanation is coming later in the text, it would be easier for the reader, if the point of underrepresentation of long transcripts was clarified up front.

We have added this information when describing the Figure 1D results (now Figure 1B), lines 275-276.

Fig 1C:

Please correct, testis > testes

The figure was corrected accordingly, and is now Figure S12.

Line 246-247 and Fig 1F:

The authors to some extend contrast TE transcripts with gene transcripts by stating: “on the other hand, gene transcripts may reach 5kb”. Although this is true based on the data, it should be taken under account that overall genes are much more highly expressed than TEs, increasing the chance of detection of underrepresented long transcripts. If the authors wish to make such comparison, they should include transcript abundance as a contributing factor.

We have now taken into account the whole dataset, multi and uniquely mapping reads, and have shown that a read of 4.5 kb stemming from insertion DOC$3R_RaGOO$24532793$24537308 is present (Figure 1F now). We therefore cannot argue anymore that reads stemming from TEs are shorter than 3 kb while reads stemming from genes may reach 5kb. We updated the text accordingly, lines 292-293.

Overall, due to lack of replicates and important difference on coverage, any conclusions regarding comparison between samples should be made very carefully. The authors do acknowledge this and mostly remain careful in their conclusions (e.x. line 293-294). However, throughout the paragraph (lines 279-294) the authors should avoid direct comparisons of read numbers between female and male libraries. Without any kind of normalization statements such as “but both families with higher transcripts in males” are not very meaningful. Also, regarding the observation that the global proportion of reads mapping to TEs is significantly higher in testes than in ovaries, it was not clear from the text, if this was also true for the previously published data (Larat et al 2017). If so, this would strengthen the obtained result, as it does in Fig 2C for the expression at the family level.

Indeed, the lack of replicates is an important matter and we have therefore toned down the comparisons between testis and ovaries. In response to other comments, we have also performed a subsampled analysis where we show that when taking the same number of long-reads for ovary and testis, there is a higher percentage of reads that map to TEs in testis (Figure 2A).

Finally, are TEs with male-specific transcripts (HETA, TAHRE) enriched on the Y chromosome?

For HETA, out of a total of 743 reads, 183 reads stem from copies located on the Y chromosome, while 560 stem from copies not located on the Y chromosome. For TAHRE, out of a total of 602

reads, 34 reads stem from copies located on the Y chromosome, while 568 stem from copies not located on the Y chromosome. Therefore, it does not seem that male-specific TEs are enriched in the Y chromosome.

Fig 3A:

Please unify: “copy number” = “# of genomic copies”

The figure was changed accordingly (now Figure 4A).

Line 360:

Please remove the abbreviation “v.r.t.”

The abbreviation was removed.

Line 356 and the results section below:

In light of the demonstration that the technical approach taken is strongly underestimating very long transcripts (Fig 1), the section title should be toned down to “are rarely detected” rather than “rarely transcribed”. Also, the authors should remind the reader of this technical limitation here and tone down their conclusion as to whether the detected transcripts are fully reflecting the transcripts presents in the tissues investigated.

We agree with the reviewer and have toned down the title, along with directing the reader towards the technical limitations of the data. In sum, we have added that longer transcripts are less abundant in the cDNA pool, and therefore in the sequencing run. Finally, we added to this discussion the bias of multimapping in very young copies, as suggested by all the reviewers.

Fig 5:

Please enlarge fonts for TE family names

The figure was changed accordingly.

Reviewed by Christophe Antoniewski, 06 Aug 2023 09:11

I think that the article "Identification and quantification of transposable element transcripts…" by Rebollo et al deserves publication because it exhaustively and fairly thoroughly presents an analysis workflow for detecting and quantifying the expression of inserting elements transposable from Nanopore long reads of a Lexogen teloprime library. Nanopore technology is still relatively new. The methods for analyzing the data generated and the bioinformatics tools required are still not standardized. It is therefore clear that the work of the authors will be of interest to biologists involved in the study of transposable elements as well as bioinformaticians having to explore data from long Nanopore reads.

That being said, the manuscript in my opinion suffers from two problems.

Regarding the biological question of the compared transcriptional profiles of TEs in the testes and ovaries, the experiments carried out are not replicated and involve a modest sequencing depth. It therefore comes to pass that, whatever the care taken by the authors in the analysis of the data, the conclusions are rather weakly supported by the observations. In other words, I believe that the statistical power of prediction afforded by the work is very limited, and that other teams using the same approach for the same question are likely to come to substantially different conclusions. I am open to discussion on this problem: it seems to me that the value of the work lies more in its methodology than in the biological significance of the observations made and I would find it a shame to delay its publication to replicate the experiments; on the other hand, many conclusions, in particular about the splicing of ET transcripts, seem too weakly supported by unreplicated observations.

The second and in my opinion main problem is that the analysis workflow developed by the authors is not very transparent and ultimately very difficult to use as it is on other datasets or for other questions. I will detail below my systematic analysis of this problem. But my main message here is that the work cannot be published as is, which is frustrating because (i) I'm confident that the authors' analysis is generally sound (ii) the workflow developed must be usable by others (otherwise, what's the point?).

Issues in the description of the analysis workflow

Line 107 "the European Nucleotide Archive (ENA) under accession number PRJEB50024" This is not the version provided in the zenodo folder. The version actually used in the work has renamed chromosomes and smaller contigs are removed. Moreover, the Y is not taken in the used assembly. This is surprising, especially given that the Y chromosome, although very small, has numerous TE insertions.

We are sorry about this mistake, and have now rerun the analysis and included the small contigs, and the Y chromosome (GCA_927717585.1 assembly). All scripts, genome and annotations used in the analysis are now available at https://gitlab.inria.fr/erable/te_long_read/.

Line 108. "Gene annotation was performed as described in Fablet et al 2022". The reference is an unreviewed preprint. Not surprisingly, the description of the gene annotation workflow in this article is not sufficient to easily check its accuracy.

We have updated the reference since the preprint in question has just been published. We have also described in more detail how the gene annotation workflow was performed (Lines 101-119).

Line 108 "Briefly, we used LiftOff". Which version ?

We have changed the text accordingly (line 106).

Line 109. "(dmel-all-r6.46.gtf.gz)" is NOT the version used for generating zenodo data (6.23)

Again, we are sorry about this mistake and have corrected the analysis and added all reference files or links to the files to the GitLab/te_long_read (https://gitlab.inria.fr/erable/te_long_read/).

Line 109. "with the option -flank 0.2". Give the full command line.

We have changed the text accordingly (line 108).

Line 110. "we produced a GTF file with the position of each TE insertion". First, access to the GTF file is not indicated (as all zenodo data for this manuscript). Secondly, it is mandatory to precisely describe the script used to generate the GTF file, since people not working on D. melanogaster TEs will need it.

The script used to generate the GTF file is now present in the GitLab/te_long_read (https://gitlab.inria.fr/erable/te_long_read/).

Line 111. " We have used RepeatMasker with DFAM dataset from D. melanogaster TE copies". Authors should be more specific. Are they referring to the file from 2006 on the repeatmasker site ? Please identify that file.

We have changed the text accordingly (Lines 101-119).

Line 112 "OneCodeToFindThemAll". Please source the resource and indicate how to use the perl tool.

We have changed the text accordingly (Lines 101-119).

Line 113. SnapGene is a commercial software. Please, provide open-access options.

We have now used blastn to align the expressed copies to the consensus sequence (now Figure 5) and have updated the material and methods accordingly (lines 117-119).

Line 145. Nanoplot v1.39.0. It looks that this version does not exist. The last stable versions of nanoplot are currently v1.33.0, v1.29.1, v1.0.0…

Indeed, we were using a version which was under development. We now use version 1.41.6 which is available here (https://github.com/wdecoster/NanoPlot/releases). We have changed the text accordingly (line 155).

Line 147. Sequencing datasets are not referenced with the same identifiers in BioProject PRJNA956863 and in bam alignments provided in Zenodo by the authors.

Again, we are sorry about this mistake and have corrected the analysis that is now thoroughly described in the GitLab/te_long_read (https://gitlab.inria.fr/erable/te_long_read/).

Line 148. I could not retrieve the Minimap2 2.17-r974-dirty, whose version you will recognize is not very engaging… Please use a stable version of Minimap2 from the GitHub repository: https://github.com/lh3/minimap2/tags.

We now use minimap v2.26 as referenced in the GitLab/te_long_read.

Line 149. "using the splice preset parameter: “minimap2 -ax splice FC30.fastq dmgoth101_assembl_chrom.fasta -o FC30.bam”". The statement is unclear. It looks like a command line to run minimap2 and from this line I would say that the parameter -ax was set to splice. In anycase, the information here is misleading and incomplete since the actual command line used by the authors was: Minimap2 -ax splice --junc-bed dmgoth101.onecode.fixed.bed -o FC29.against_dmgoth.sam dmgoth101_assembl_chrom.fasta FC29.fastq.gz As extracted from the bam alignment file and for FC29.fastq.gz. There, it is obvious that the –junc-bed parameter was also set to dmgoth101.onecode.fixed.bed (by default this parameter is unset), and I am curious to know, as other readers will likely be, how the bed file was generated, and also how it was fixed…

The full command line for minimap2 is now given here: https://gitlab.inria.fr/erable/te_long_read as well as the steps required to prepare the input files (genome and annotation).

From line 150 to line 175. There are a number of statements here that came, I guess, from an analysis of the bam files. I was able to verify, using my own knowledge and tools, that most of them look correct. Not surprisingly however, I could not reproduce exactly the results. The point here is that parsing methodology, small pieces of codes and command lines should be indicated. Otherwise, we have just to trust the authors, and cannot be of any help if we see analysis errors (that always occurs, unfortunately).

The scripts to reproduce the analysis here, along with how to obtain the data, genome and annotations, are now thoroughly described in the GitLab/te_long_read (https://gitlab.inria.fr/erable/te_long_read/).

Line 165. Figure S4. It seems that this corresponds only to ovaries in Fig S4 Line 179. "Then, we discarded all reads that covered less than 10 % of the annotated TE" I guess that this was done using the python script in the ipnb provided by the authors in Zenodo. However, this script is not mentioned in the manuscript. I will come back later on the use of a ipnb file to capture and restitute methodology. Here, I am just saying that if this script is used in support of the line 179 statement,

an explicit reference to the python code block that is ensuring the filtering of the reads should be placed in the manuscript.

The same remark stands for the 3 filtering statements between line 176 and line 182.

The code to reproduce the filtering statements is now clearly indicated in the git repository. It corresponds to the file generate_counts.py lines 308-374.

Lines 191-192. The symmetric difference should be computationally defined (not only graphically). In particular the piece of code in the ipnb file dealing with symmetric difference calculation should be explicitly commented. The rationale of the smallest symmetric difference is not so clear in the example given in Figure S8.

The piece of code used to assign the read to the feature with the smallest symmetric difference corresponds to lines 243-262 of file generate_counts.py. When a read overlaps several features, we compute for each pair read-feature 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. In other words, it corresponds to the size of the union interval minus the size of the intersection interval. Choosing the feature with the smallest symmetric difference is in practice very similar to choosing the feature with the largest Jaccard Index (which corresponds to the size of the intersection interval divided by the size of the union interval). We have now represented this step in Figure 1C.

Line 220. "In order to validate the long-read RNAseq approach, we first determined the read coverage of all expressed genes". How ? There are several options to perform this task. What code/script/command lines were used ? Line 224 and Figure 1B: Same questions as for Line 220.

We have added a paragraph to the methods (line 211-214) explaining how this was performed. Briefly, we mapped reads to the reference transcriptome and computed, for each primary alignment, the subject coverage and the query coverage. We also provide the command lines used and the scripts in the git repository (sam2coverage_V3.py, breadth_analysis.sh)

Lines 224-227. Go term enrichment analysis. The authors obviously forgot to explain how this analysis was performed. I am curious to know in particular since at this stage no Differential Gene Expression analysis was mentioned.

We have indeed forgotten. A section was added to the methods on lines 215-219, and the GO term results are now present in Figure S12.

Line 230. "These correlation coefficients are in agreement with those obtained when comparing direct RNAseq vs Illumina Truseq sequencing (Sessegolo et al., 2019)" I find the agreement cryptic. What is the point of this comparison since the authors performed cDNA sequencing ?

The reviewer is right, the statement was not clear. In previous work, we had shown that direct RNAseq was the best technology to quantify transcripts, and we wanted to recall this fact, but as it was confusing, we decided to remove the sentence.

Line 395. "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." Again an irreproducible procedure. Please expose all the necessary details to reproduce these searches.

We have added a paragraph to the methods (line 226-231) explaining how this was performed.

Line 398. "The remaining cases likely correspond to genomic deletions." If it comes to genomic deletions, I would have expected the presence of D's instead of N's in the CIGAR. However, this may depend on the aligner and I do not know how minimap2 is dealing with gaps in general.

Indeed, when the deletion is too long, it could be interpreted as an intron by the mapper, depending on its scoring scheme, which is why we initially proposed this hypothesis. We now also mention the possibility of non-canonical splicing.

About the python script in the ipnb file available in Zenodo

I appreciate the effort of the authors for providing a notebook of their python code used for the analysis. Using it, I was able to run the notebook and to reproduce some of the results of the authors. Note that it does not mean that the code is correct, just that it executed as expected. However, there are a number of issues here.

- At first, the notebook is not currently mentioned in the manuscript

- A Ipython notebook is a nice tool to develop code, to explore algorithms and

procedures, or to draft an analysis. But I do not think it is the best way to publish

bioinformatics.

- Here the dependencies are not specified (which can be a real issue in the

case of pysam whose methods have considerably evolved over the last years)

- A procedure to install these dependencies should be indicated (at least a pip

dependencies.txt file)

- The code is split between various blocks

    - Most importantly. The code is not "parameterized", which is against the

elementary python guidelines for code testing and reusability. The authors

should derive a python script from their notebook, and make use of optparse

or equivalent python module so that input parameters are generalized (thus

reusable) and not dependent on the authors data.

- The functions and classes in the notebook are insufficiently annotated.

I think that instead of the notebook, or in complement, the authors should provide a python script making use of parameters, along with its requirement.txt file. Possibly in a github or gitlab repo. Currently, it is not clear in the manuscript which part of the results was computed with this script and which part was not. It should be indicated in the manuscript.

The scripts to reproduce the analysis here, along with how to obtain the data, genome and annotations, are now thoroughly described in the GitLab/te_long_read (https://gitlab.inria.fr/erable/te_long_read/).

Style and Typos

line 27, Abstract. "Potentially able" - is a pleonasm.

We do not agree with the reviewer as potentially able means that some TE copies have the ability to, while others don't.

Line 383. exemptions I think the authors mean exceptions ?

Yes! We have changed the text accordingly.

Line 386. "Gypsy copies are able to produce ENV proteins through mRNA alternative splicing, also regulated by piRNAs" This is ambiguous: what is also regulated by piRNAs ?

We have reformulated the sentence on line 529.

Line 414. "But" or "albeit" but not both

We have changed the text accordingly.

Line 489. "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." This sentence is barely understandable. Please rephrase to expose your point.

We have reformulated the sentence and hope it is clearer (Lines 621-625).

Reviewed by Silke Jensen, 13 Aug 2023 14:03

PCI Genomics #250

In this manuscript, Oxford Nanopore technology was used to sequence poly-adenylated capped transcripts to study the transcription of transposable elements (TEs) in an iso-female strain, dmgoth101, which originated from a wild-caught female of *Drosophila melanogaster*. RNA from ovaries and testes was reverse transcribed, amplified by PCR and then sequenced. To study TE transcripts, reads were mapped to the corresponding dmgoth101 genome. Genome annotation was used to identify transcripts containing a TE sequence. The aim was to assign reads to individual genomic TE copies and study their characteristics, the landscape of TE transcription in ovaries and testes and detect putative spliced TE transcripts. One of the main difficulties of this study was that the reads recovered were rather short, less than 2.5 kb long for TEs, less than 5 kb for genes. Despite this difficulty, the authors present interesting results concerning different expression landscapes in ovaries and testes. They also present data that seem to evidence novel spliced TE transcript isoforms. They compared some of these results with those obtained with short-read sequencing data.

I think that this study is interesting but that there is still a lot of work to be done on the manuscript. I very much hope that my comments below will be helpful in this respect.

Major:

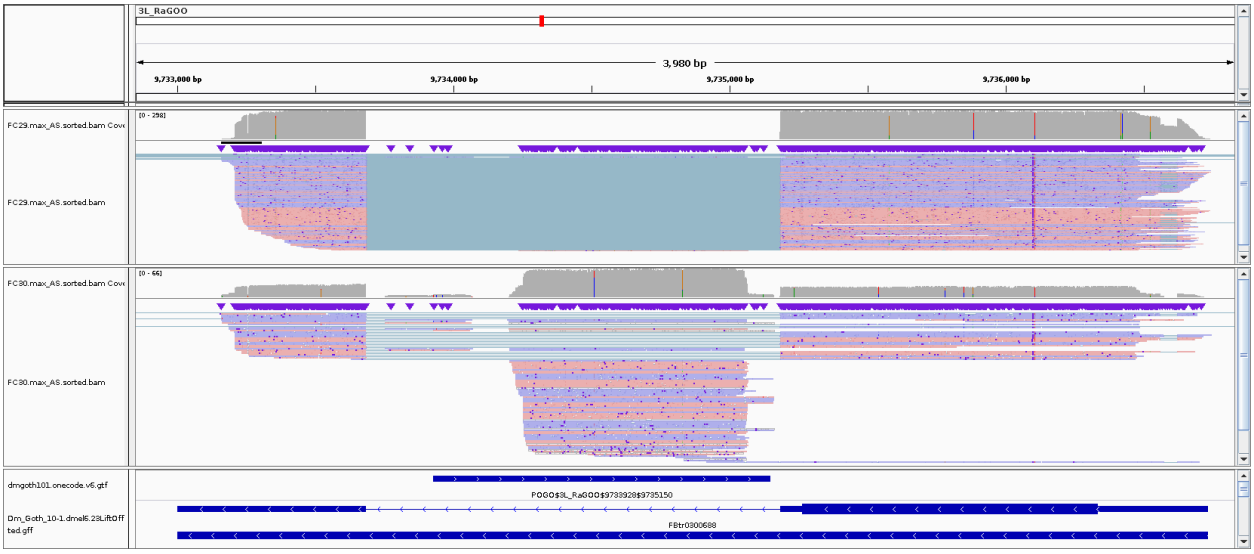
It seems that the dataset corresponding to short-read RNA-seq data for dmgoth101 ovaries, from NCBI BioProject database PRJNA795668 (Fablet et al., 2022), is not accessible. To be verified. There also is a problem with the short-read RNA-seq data for dmgoth101 testes (SRR25004058), which makes a simple IGV visualization after HISAT2 mapping in my hands impossible. Therefore, an analysis of the short-read RNA-seq data to compare with the long reads was not possible within the scope of the review.

I have not found the dmgoth101 genome (line 148) in the databases, only a genome that has not been assembled into chromosomes. This genome should be made available, or an accession number provided, so that the results can be reproduced. In addition, this would enable genomic regions shown in some figures to be visualized in the genomic context of the dmgoth101 line analyzed here (e.g. in figure 3D).

As stated to the previous reviewer, the scripts to reproduce the analysis here, along with how to obtain the data, genome and annotations, are now thoroughly described in the GitLab/te_long_read (https://gitlab.inria.fr/erable/te_long_read/).

The fact that many TE transcripts may be transcribed from promoters located in flanking regions and not from their own promoter is not discussed. For example, when analyzing the pogo-mapping reads shown in Figure 3D and Figure 4B, it appears that most of them also contain sequences other than pogo sequences at their 5'- and/or 3'-end.

Indeed we did not discuss the fact that some TE copies might be expressed from promoters outside of the copy itself, nor that TE copies might be involved in gene-TE chimeric transcripts, but we think this is beyond the scope of this manuscript, although the data could indeed answer this question. Regarding *pogo* copies, Figure 3D is now Figure S20 simply because it is now the second most expressed *pogo* copy in ovaries, the first most expressed *pogo* copy is now present on Figure 4C. When looking at both IGVs the vast majority of reads are included within the TE copy. We have added a column on Table S3 and Table S4 named "Mean_Bases_Outside_TE_Annotation" which corresponds to the mean length of the long-reads that are mapped outside of the TE copy and have plotted this metric on Figure S6. When looking at the three most expressed *pogo* copies in ovaries, we obtain 56, 7 and 4 mean bp of reads outside of the TE copy. If we focus on the 56 bp for the most expressed *pogo* copy in ovaries (POGO$3L_RaGOO$9733928$9735150, see the IGV screenshot below, FC29 = testis, FC30= ovaries, squished IGV on top, and expanded in the bottom) this corresponds to a few reads that are mapping at the 5' end of the copy. From these analyses, it is clear that the majority of *pogo* reads originates and ends within *pogo* copies.



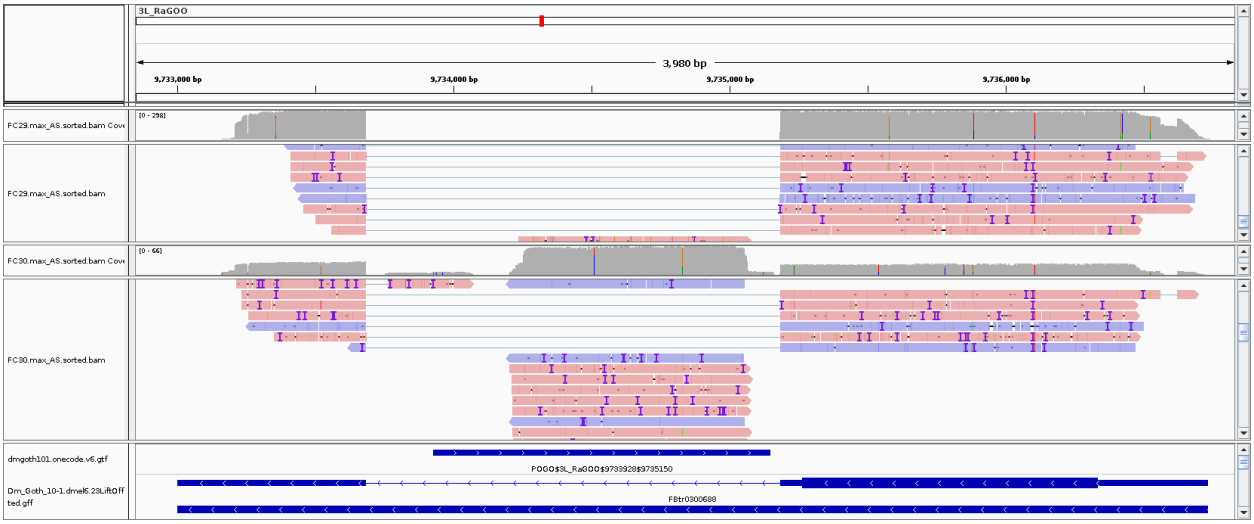


Figure 2A: It would be useful to also present the TE transcriptional landscape obtained with short-read sequencing to compare the results obtained by the 2 technologies, ONT and Illumina sequencing.

The figure can now be appreciated in the supplementary materials (Figure S17) and has also been discussed in the manuscript (lines 352-356).

Figure 3D: It seems that reads that are repeated, which have a mapping quality of zero, are not shown in Figure 3D. But if this pogo element is repeated in strain dmgoth101 and expressed from its own promoter, these reads, even if from the pogo element within the CG12061 gene, would not be visualized. The fact that reads with a mapping quality of zero are not shown or considered should be clearly indicated and discussed, as there may be a significant bias in the detection of transcripts from recently transposed TEs, thus repeated in the genome. In relation to these considerations, it is possible that the filters chosen (lines 176-184 and mapping quality filters?) do not take into account the transcription of young TEs that are repeated in the genome. At the least, this point needs to be clarified and discussed. Notably, without these filters and mapping TE consensus sequences, 1 857 TE-mapping reads are found in ovaries and 11 172 in testes, instead of 1 322 and 8 219 respectively (L237-238).

As stated to the previous reviewer we have computed the number of multimapped reads for each family and copy and have thoroughly discussed the contribution of such reads to TE family expression and to the identification of expressed copies (Figure 3 but also the first section of the manuscript). Figure 1, now takes into account all reads may they be uniquely mapping or multi-mapping). Concerning the copy POGOX_RaGOO21863530$21864880 shown in Figure 3D, now Figure S20, there are 77 uniquely mapping reads, and no multimapped reads. The current *pogo* copy shown in Figure 4C (POGO$3L_RaGOO$9733928$9735150) has 80 uniquely mapping reads and no multi-mapped reads.

In nearly all figures, normalization of the read count would be a good thing for comparison purposes between ovaries and testes.

We have toned down the comparison between testes and ovaries, as suggested by the other two reviewers, due to a lack of replicates. However, we have followed suggestions from readers of the preprint, and have randomly subsampled the reads between the two tissues to match the same read number. We reproduced the GitLab/te_long_reads pipeline in these 1 million subsampled reads and have produced the same plots as for the whole dataset. These can be found in Figure 2A, Figure S15, Figure S16 and Figure S19 and discussed throughout the text.

Figure 4B: This figure is rather misleading as it shows “alignment of transcribed copies against their consensus” together with read counts. This gives the impression that there are, for example, more than 250 ONT RNA-seq reads covering almost the entire Copia element (left), which is clearly not the case, since none of the reads covers the entire Copia element. In fact, the longest Copia read (in testes) covers only 2 254 bp of Copia, and in ovaries, all reads except one (which is ti1 kb long) correspond to Copia copies with a large internal deletion. To avoid this misinterpretation, the RNA-seq reads themselves should be shown in this figure. For the same purpose, paragraph L366-373 should be reworded, as it is difficult to understand the link between TE copies that “covered at least 80% of their consensus sequences” (L368) and counts and TE coverage of the RNA-seq reads.

Yes, we agree with the reviewer, the figure was indeed misleading! We have now changed this entire section to illustrate the expressed genomic copies and not the reads *per se*. Therefore, figure 4B, now figure 5B shows the alignment between the expressed copies (at least 5 reads for both tissues) and their consensus sequences. This allows us to explain that most TE copies found expressed are not full-length copies (due to a technical issue or to a biological factor, all explained in the manuscript now). The reads are now represented in Figure 6, with IGV screenshots.

L395-398: “The remaining cases likely correspond to genomic deletions.” What about retrotransposed spliced transcripts? Did the authors search for such TE copies in the dmgoth101 genome? Such copies would also have GT-AG bordering the putative intron. This question arises especially for Copia where virtually only possibly spliced transcripts are detected (L430-431), while the corresponding putative AG splice acceptor site cannot be clearly identified for most Copia reads (as it seems from my analyses). A possibly retrotransposed copy of spliced genomic Copia could be identified by PCR in case such a copy is located in an unassembled part of the dmgoth101 genome (or by first analyzing the raw genome reads).

Indeed the presence of the retrotransposed copy of a spliced genomic Copia could very well explain the pattern we see. We therefore double-checked this by mapping the raw Nanopore genomic reads (ERR4351625) to both Copia and a spliced version of Copia and we could verify the absence of any genomic read mapping to the spliced version of Copia. For us, this confirms that all genomic copies of Copia indeed contain the full-length element.

It would be necessary to describe in the Materials and Methods section how the GT-AG splice sites were found for the reads with gaps (method, tool, script). Since these putative splice donor

and acceptor sites are located within the intron and therefore not in the putative spliced reads, have they been identified after mapping to TE consensus or to genomic sequences? Another difficulty lies in the fact that splice site mapping is often imprecise due to the high error rate of Nanopore sequencing. For example, I was unable to identify GT-AG for the putative 1.3 kb spliced transcripts for 1731 shown in Figure 8A when mapping to the 1731 consensus. For Copia, the site of the AG splice acceptor cannot be clearly identified. It would be good to show these GT-AG putative splice donor and acceptor sites, e.g. in a supplemental figure, at least for Copia, 1731, Pogo and some other examples.

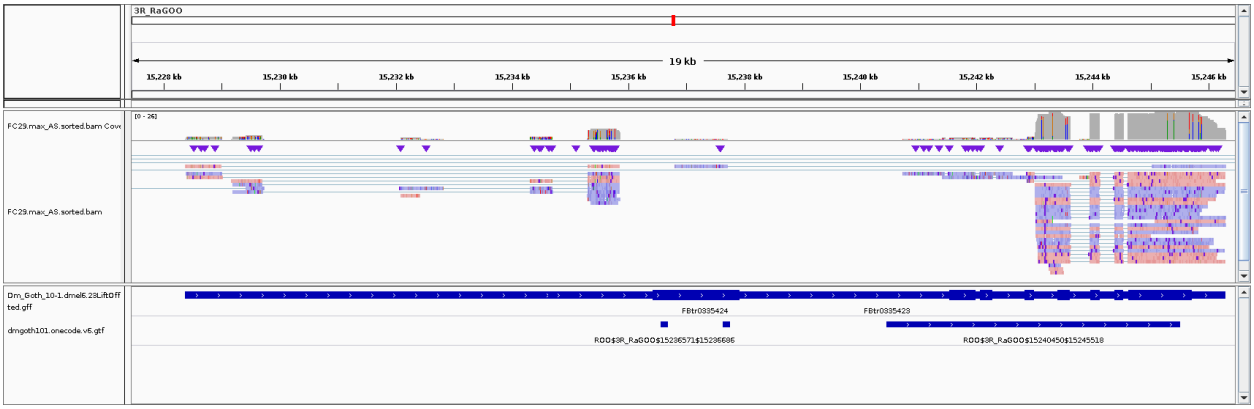
We have now added a section in the Methods to explain this (lines 226-231). Briefly, the reads were mapped to the genomic copies of TEs (not the consensus) and we retrieved the dinucleotides flanking the gap (N's in the CIGAR string). Scripts to perform this have been added to the git repository (SplicingAnalysis.py, splicing_analysis.sh). We also added IGV screenshots showing GT-AG consensus sites at the locations where the reads map for *POGO*, *Copia* and *DM1731*. Those are supplementary Figure S26-S29.

L382-401: "Long-read sequencing unveils novel spliced TE isoforms" When searching for reads that could indicate splicing of transcripts for TAHRE, TART or Roo, which are reported to have a high number of putative spliced transcripts (Figure 5), I found essentially no reads that could correspond to splicing events within these TEs. It seems that most of the reads mapping these elements originate from TE copies which are partially deleted. For Roo, most reads correspond to transcripts containing a Roo solo-LTR as well as other sequences surrounding this solo-LTR. In ovaries, only 5 reads mapping Roo do not correspond to such reads containing the solo-LTR (>30 reads). None of these 5 reads show evidence of splicing within Roo. This is not compatible with the percentage of Roo spliced reads in figure 5. Is it possible that the splicing events detected originate from the splicing of chimeric transcripts that contain flanking genomic DNA as well as TE sequences, and that these splicing events in fact correspond to the splicing of chimeric transcripts within the gene portion of the transcripts? Have the authors verified this point? Were the gaps detected in the part of the transcripts corresponding to the TEs? If not, these splicing events cannot be considered as evidence of "novel spliced TE isoforms". The findings reported in this section of the manuscript on TE splicing need to be re-examined and supported by much stronger evidence. In fact, these observations cast serious doubt about the results presented for putative TE splicing.

We have rerun the splicing analysis using more stringent criteria. We map the reads to the TEs and the transcripts, and for each read, instead of keeping the primary alignment (as we had done for the previous submission), we now keep the alignment with the best alignment score. Keeping the primary alignment was wrong in some cases. Below is an example of a read which maps both to ROO$2R_RaGOO$14213942$14227652 and a gene (CG8180), but where minimap incorrectly flags the ROO alignment as primary, while the read clearly stems from the gene.

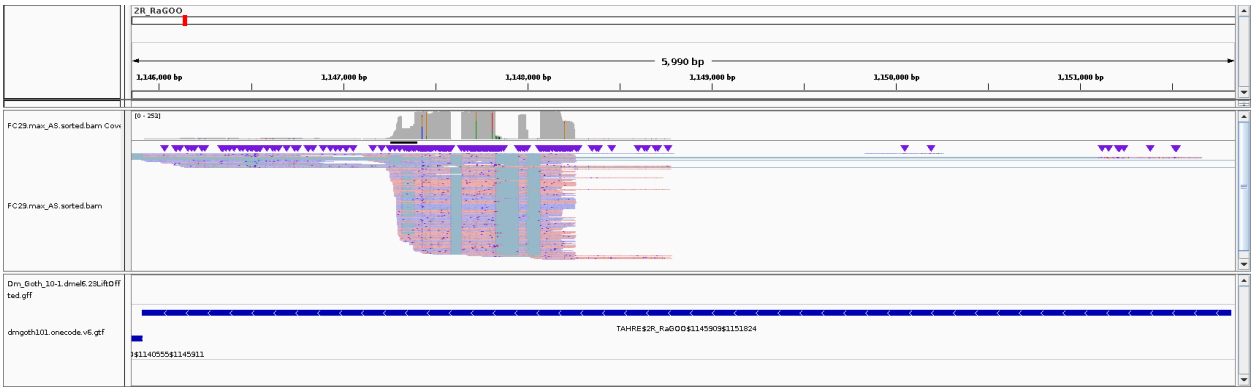
Example of a read which maps both to ROO and the transcript. Minimap considers that the primary alignment is ROO and the secondary alignment is the gene (there are three alternative transcripts for this gene). However, using the alignment score (AS tag) we see that the read aligns better to the transcripts (AS=123) than to the TE (AS=94).

We now do not have any ROO insertion which is considered as spliced and highly expressed. Some insertions, like ROO$3R_RaGOO$15240450$15245518 are supported by 9 reads in testes, 4 of which may be considered as spliced as seen below.



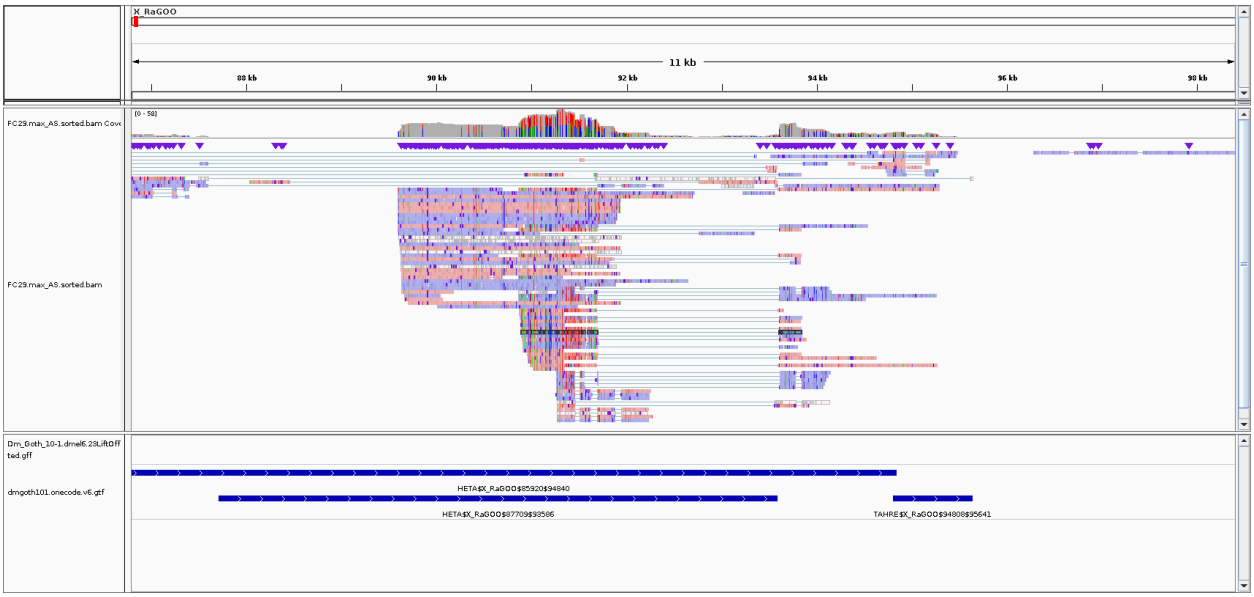
Intron detected in ROO$3R_RaGOO$15240450$15245518. In this case, the TE is included in a gene. Our method assigns reads to the TE, not the gene, because the gene is less well covered by the reads than the TE. A possible explanation of this situation could be that the reads correspond to the use an alternative transcription start site located within the intron of the gene (in the middle of the TE). The end of transcription corresponds to the end of the gene. Introns are spliced because they contain the GT-AG splice sites of the gene.

We however still have evidence of splicing in highly expressed insertion of TAHRE, HETA and TART-A. For instance, the expression of TAHRE$2R_RaGOO$1145909$1151824 is supported by 281 reads in testes, 92% of which are spliced (see figure below). Most reads span the 3 same introns. The coverage of the TE is partial (15%).



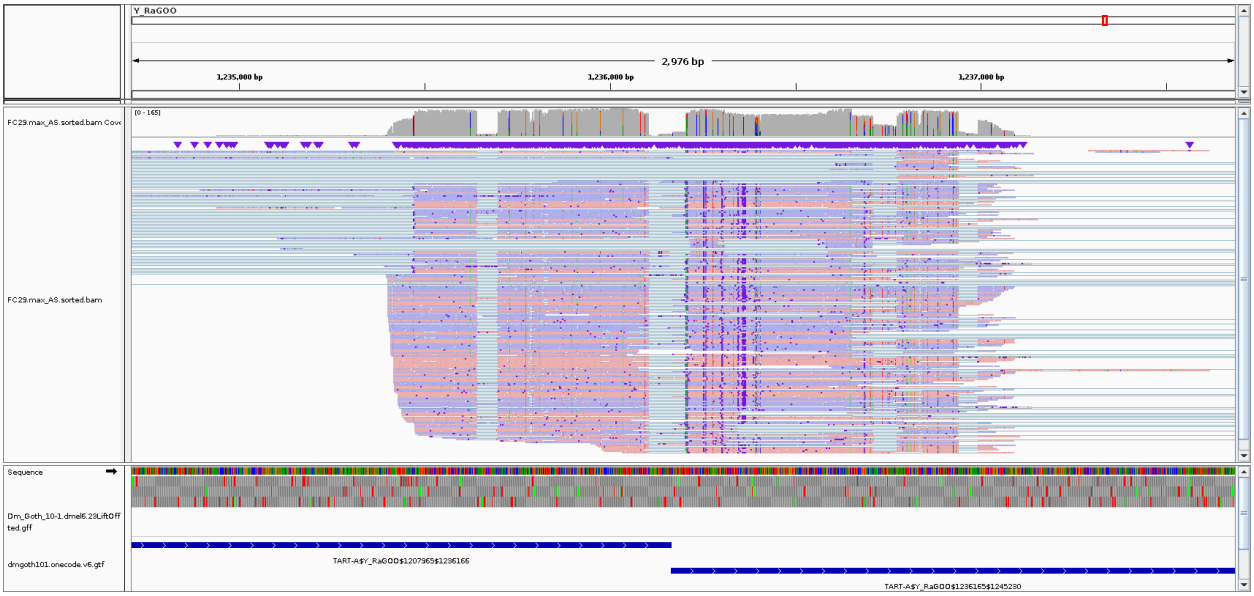
The expression of TAHRE$2R_RaGOO$1145909$1151824 is supported by >200 reads, most of which span several introns. The coverage of the TE is partial (15%).

The expression of HETAX_RaGOO85920$94840 is supported by 41 reads in testes, 87% of which are spliced. Spliced reads span one intron, not always the same (see figure below).



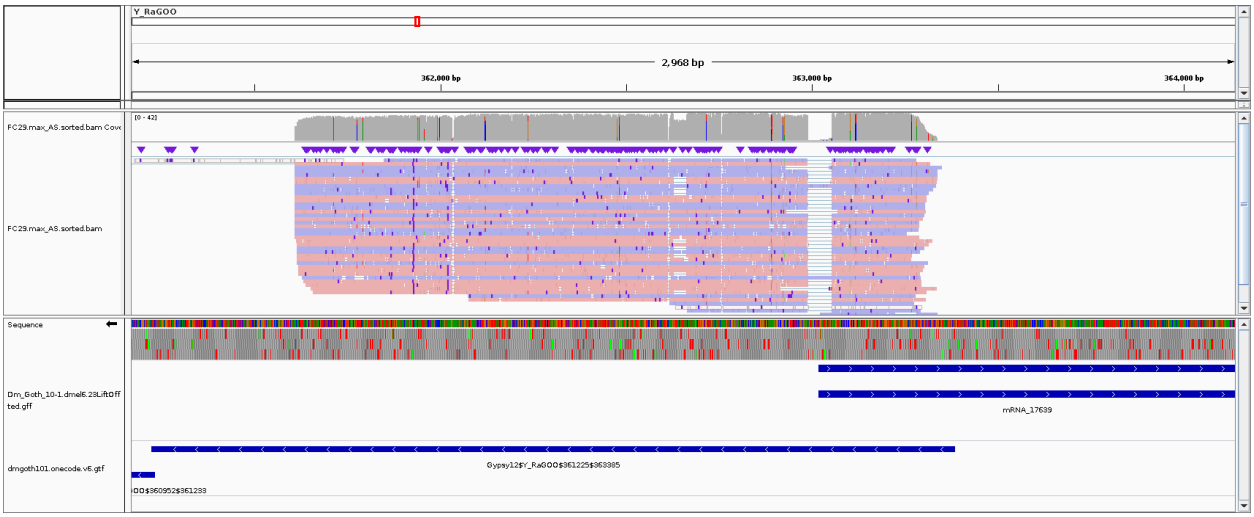
The expression of HETAX_RaGOO85920$94840 is supported by 41 reads in testes, 87% of which are spliced. Spliced reads span one intron, not always the same. The most supported intron (17 reads) is the one with the highlighted read.

The expression of TART-AY_RaGOO1207965$1236166 is supported by 242 reads, 86% of which are spliced and span several introns (see figure below). The transcription unit overlaps two annotated TART-A insertions.



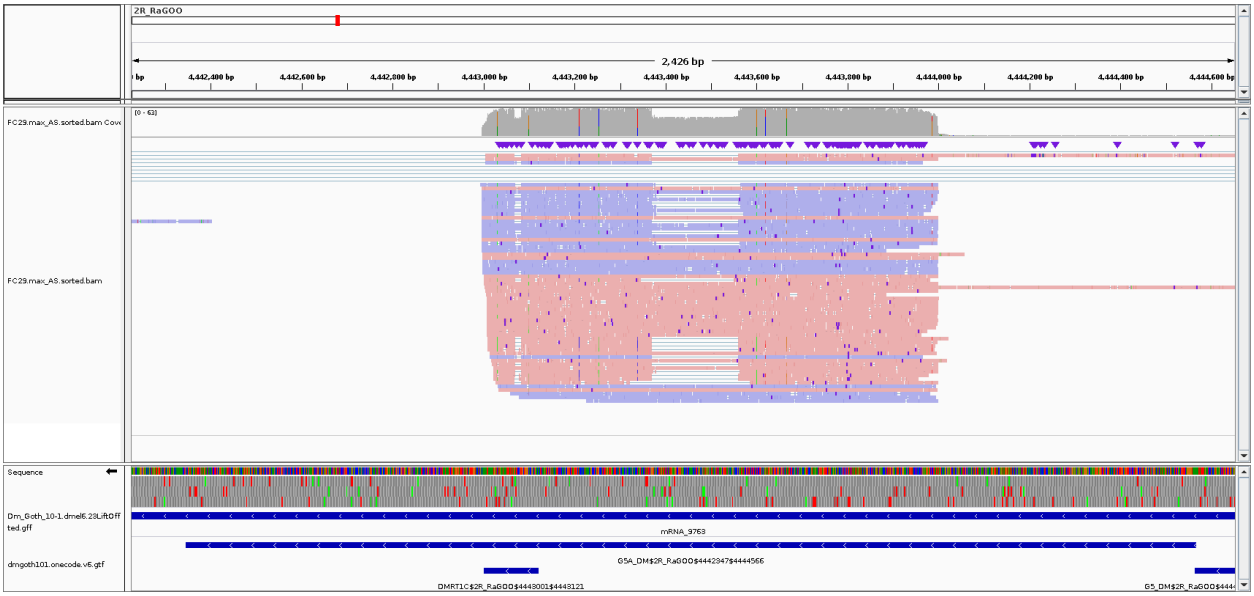
The expression of TART-AY_RaGOO1207965$1236166 is supported by 242 reads, 86% of which are spliced and span several introns. The transcription unit overlaps two annotated TART-A insertions.

The expression of Gypsy12Y_RaGOO361225$363385 is supported by 46 reads, 91% of which are spliced (see figure below).



Gypsy12Y_RaGOO361225$363385 is supported by 46 reads, 91% of which are spliced

The expression of G5A_DM$2R_RaGOO$4442347$4444566 is supported by 64 reads, 32% of which contain gaps, but without GT-AG flanking sites (see figure below). Those could be non-canonical introns, genomic deletions, or mis-alignment of the reads due to a gap in the genomic assembly.



G5A_DM$2R_RaGOO$4442347$4444566 is supported by 64 reads, 32% of which contain gaps, not flanked by GT-AG splice sites. Those could be non-canonical introns, genomic deletions, or mis-alignment of the reads due to a gap in the genomic assembly.

Another problem is that there is a big difference between the TE read counts found by the method adopted by the authors and the read counts found when aligned to consensus TE sequences. Some examples in testes: TAHRE: 590 reads (manuscript supplements_542554_file04) vs. 216 reads (consensus mapping); Nomad: 399 reads (manuscript supplements_542554_file04) vs. 102 (consensus mapping); Roo: 438 reads (manuscript supplements_542554_file04) vs. 390 (consensus mapping). This suggests that many reads may be incorrectly assigned to TEs.

It is complicated to compare the method suggested by the reviewer and the method we have used. Indeed, mapping to consensus might remove reads that map partially to TEs, especially if we do not precisely know how the reviewer has proceeded with the analysis. Here is an update of the three families that have been suggested by the reviewer as an example:

TE family	Total read count - Ovaries	Total read count - Testis
TAHRE	6	603
NOMAD	1	393
ROO	88	486

Given that we filter for reads mapping at least 10% of the TE sequence, we wondered if the reviewer has a filter for reads covering a higher percentage of the TE consensus length. We have not filtered for a minimum of mapping within the read (for instance, a read does not need to match 100% to the TE), and this can also influence the number of reads mapping to the TE. Therefore,

if the reviewer clarifies the methods used, we can then proceed to a proper comparison in order to understand these discrepancies.

L470-492, “Conclusion”: A major problem of this study is that most long reads recovered correspond to transcripts that correspond to ancient, non-functional TE copies. These transcripts seem to be transcribed from promoters that are in genomic regions flanking the TE. This is indeed an interesting result but to my opinion the most interesting transcripts mapping TEs are the transcripts which are produced by functional TE copies. Here the authors state: “Here we demonstrated the feasibility of assigning long reads to specific copies, which remains the biggest issue in TE expression analysis.” It would be a good idea to discuss why the authors think that this is the biggest issue.

We do not agree with the reviewer that the transcripts seem to be transcribed from promoters that are in genomic regions flanking the TEs, although there are indeed cases where this happens but we have not focused this manuscript on this particular cases. In a previous manuscript (Jeffrey R. Adrion, Michael J. Song, Daniel R. Schrider, Matthew W. Hahn, Sarah Schaack, Genome-Wide Estimates of Transposable Element Insertion and Deletion Rates in Drosophila Melanogaster, Genome Biology and Evolution, Volume 9, Issue 5, May 2017, Pages 1329–1340) the authors observed 24 active TE families in *Drosophila melanogaster*. Compared to our dataset, we recovered 11 common TE families (*G-DM, hobo, roo, DM297, blood, burdock, copia, DMCR1A, diver, idefix and pogo*). Some of the TE families were not found as the names might differ between annotations. Out of these families, we have discussed *pogo, copia, burdock, DMCR1A* and *roo* in the manuscript. While some of these families have full-length, young copies being transcribed (Figure 5A), the reads do not correspond to full-length copies, as most of them are either truncated or spliced (Figure 6). We have therefore rephrased the conclusion in order to match the technical problems of recovering long-cDNAs, along with the biological rarity of these transcripts. We have also added that in this manuscript, we have not found reads that span the entirety of a full-length copy apart for *mariner2.* We have discussed this in the conclusion section.

L476-477: “The genome of *D. melanogaster* contains many functional full-length copies but only a couple of such copies produce full-length transcripts in gonads.” It appears that, with the filters applied, transcripts corresponding to functional full-length TE copies (repeated in the genome) can only barely be detected (see below, concerning zero mapping quality). These reads should have their transcription start site inside the TE and should mainly terminate inside the 3'-end of the TE, unless they are transcripts read through the TE poly-A signal. These features should make assignment to specific genomic copies of the TE impossible. It is unclear whether only genome-unique reads were explored in this study. These considerations need to be discussed.

We have now included multi-mapped reads in our analysis and with the exception of a few TE families, most expressed copies can be identified by long-read transcripts. We have added a couple of new paragraphs lines 370-411 along with Figure 3 in order to discuss this potential bias. Concerning reads included within the TE copy, indeed, as stated previously, we have not included these filters in our analysis. A new column named “Mean_Bases_Outside_TE_Annotation” is

present in the tables concerning each copy transcribed for ovaries and testes (Table S6 and S7). We have plotted the frequency of this metric on all the copies where at least one long-read was detected (Figure S6). The majority of the copies have reads that fall within the copy boundaries. We have added a discussion in the methods section.

L481-483: “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.” This seems overstated since there is a near gene, zyd, upstream of this copy of Pogo, whose expression is higher in ovaries than in testes. It is therefore also possible that this zyd gene has an enhancing effect on Pogo expression. This point should be discussed.

The reviewer is right and we have removed this sentence from the manuscript.

L488-489: “Finally, it is important to note that we did not recover TE transcripts longer than 2 Kb, despite gene transcripts up to 5 Kb.” Read number and copy number are higher for genes than for TEs. It is therefore also possible that 5 kb was reached for genes but not for TEs, simply for statistical/probabilistic reasons. See also comments on figure 1F below.

The reviewer is right although with the new analysis we see no differences between genes and TEs.

Minor:

Line 38: “We show that long-read RNAseq can be used to identify and quantify TEs at the copy level.” Quantifying TEs at the copy level can be done with genomic data only. Replacing "TEs" with "transcribed TEs" would be more appropriate.

The text was changed accordingly.

L 220-221: What is meant by “all 220 expressed genes”? Are these all genes that are annotated in the reference genome as expressed in testes or in ovaries?

Sorry for the confusion, we have rephrase it and hopefully the text is clearer (lines 266-276).

L222: “… covering more than 80% of their sequence”. Do the authors mean 80% of the gene or 80% of the longest transcript? To be reworded for clarity.

In the case where a gene has several annotated transcripts, we consider all transcripts. We have also done it for genes which had only one transcript and the result was similar. This is now explained in the Methods section (breadth of coverage). In the special case where one alternative transcript is a substring of another, and only the shorter one is expressed, our method may assign reads to the wrong transcript which may yield to a slight underestimation of the breadth of coverage.

To be reformulated for clarity: L223-224: “Besides, few reads correspond to partial transcripts, as most reads (68.9% in ovaries, 78.6% in testes) correspond to well-covered transcripts (>80% coverage) (Figure 1B).” Any transcript with less than 100% coverage is only partially covered. What is the definition of “partial transcripts”? I think authors more likely wanted to state about partial coverage here.

Indeed, the reviewer is correct and the text was changed accordingly (lines 268-276).

L224-225: To state that “This shows that the TeloPrime protocol was successful in capturing full-length transcripts.”, it would be preferable to indicate the percentage of transcripts with 100% coverage.

Indeed, this was an over-statement. Figure 1A shows that reads are indeed long, but do not always cover the fully annotated transcript. We rephrased the sentence accordingly.

L244-245: “While TE copies range from a few base pairs to ~15 Kb, 75% of annotated copies are smaller than 2 Kb.” By “copies”, do the authors mean “reads”? What does “annotated” mean here, annotated as a gene?

Copies as in “genomic copies” so the copies annotated by Repeat Masker in the genome. We changed the wording in order to clarify this point, line 290.

L252: “We concluded that indeed long and very long transcripts are a minority …” How can the authors be sure that a read, even if it is long, reflects the actual length of a transcript? This would mean that reverse transcription and sequencing would have to cover the entire transcript. It seems tricky to conclude from read length to transcript length. It would be good to rephrase this part. Maybe the authors meant “reads” instead of “transcripts”? More clarity is needed here.

Indeed, we cannot know the actual length of a transcript. We therefore approximate it with the length of its annotation. When we say that very long transcripts are a minority, we mean that, among all annotated transcripts that have at least one read assigned to them, there are only 130 whose annotated length is above 5kb, whereas there are 3603 whose annotated length is below 3kb.

Figure 1A: The cDNA amplification step needs to be indicated here, as RNA can also be read directly without this step by Oxford Nanopore Technology.

This panel was removed as deemed unnecessary.

Legend Figure 1, L257: “The majority of transcripts recovered are full-length.” The term “fulllength” is not appropriate here since this would mean that the reads cover 100% of the annotated transcripts and this is not the case. Furthermore, this is a conclusion that cannot be drawn from Figure 1B. This sentence should be removed from the legend.

The sentence was removed as suggested.

Figure 1F: Replace “Lenght” with “Length”. It is not clear what “TE copies” correspond to. Does it refer to the length of the TEs transcribed in the samples analyzed? “Reads mapping to TEs encompass most TE copy length but lack transcripts longer than 5 Kb, as also observed for reads mapping to genes.” This is not a fair conclusion since this analysis would have to be carried out for each TE separately (analysis of paired data) to be able to draw such a conclusion. To be rephrased. In addition, it seems that there are no TE transcripts >2kb, not 5 kb. This would mean that not any TE >2kb gives a transcript spanning its entire length. To determine whether this is a technical bias or a biological reality, it would be useful to also show the length of the expected gene transcripts in this figure. This would make it possible to check whether long gene transcripts generate the expected long reads.

The text was replaced as suggested. Here length means the length of the genomic copy.

L266-267: “… in agreement with the previous observations using short-read sequencing (Fablet et al., 2022).” This is not correct, since in Fablet et al. 2022, around 0.6% of reads aligned with TEs in ovarian samples. This is five times higher than the 0.11% observed in this study on long reads.

The reviewer is right. Given the differences in read assignment to TE features between long-read and short-read, we have decided to remove this comparison from the text as it is no longer valid.

L271: replace “LTRs” with “LTR elements” or “LTR retrotransposons”.

The text was changed accordingly.

Figure 2A: Replace “LNE” with “LINE” (as in the main text).

The figure was changed accordingly.

Figures 2B and 2C: Normalization, for example to the global read counts or to all TE-mapping read counts, would be a good thing here.

While we can not normalize the long-read dataset given we only have one replicate, we have performed a subsampled analysis and show that the conclusions are maintained.

L285-286: “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.” If there is only one read, it cannot be concluded that their expression is “specific to ovaries”. This makes no sense from a statistical point of view. Delete. Diwo for testes, re-analyze using non-parametric statistical methods would be useful.

We agree with the reviewer and the text has been modified to simply describe the detected families using ONT reads (lines 338-347).

L293: “… and suggests retrotransposons might be strongly and specifically expressed in males compared to females.” This conclusion concerns only LINEs, not all retrotransposons. Replace “retrotransposons” with “LINEs”?

We have removed this sentence in order to tone down the conclusions on testes and ovaries comparisons as *per* other reviewers suggestions.

L315: “… at least one long-read transcript …” What does “long-read transcript” mean? Is it a “long read” or a “full-length transcript”?

Sorry, we have changed the text to “long-read”.

Figure 3A: The threshold of “>1 read” seems rather hazardous from a probabilistic point of view.

We agree with the reviewer but have maintained this threshold to demonstrate what has been recovered in our datasets. We could make another figure with a higher threshold if the reviewer deems it important.

Figure 3B: Normalization of the read count would be a good thing for comparison purposes.

While we can not normalize the read counts given we only have one replicate for each tissue, we have instead performed a subsampling analysis and show it in Figure S19.

Figure 3D: What does “Simplified IGV screenshot” mean?

It means the screenshot was edited to remove all grey areas from the background.

L360: The terme “w.r.t.” needs to be defined. L398: “While most TE copies harbor intronless transcripts (visible in Figure 5, as circles located at 0 in the X-axis), …” What does “most” mean here. It would be useful to indicate the numbers.

The term “w.r.t.” was changed to “with regards to”. The description of Figure 5 has been updated and clarified.

Figure 5: What is shown on the left and right of the figure is not indicated. Replace “% spliced transcripts” with “proportion spliced reads” according to the legend, since the X-axis shows a proportion (0 to 1) not a percentage. Moreover, several reads may correspond to one same transcript.

We have changed the figure accordingly.

Figure 6: It should be noted that it was not possible to align the reads to the assembled genome used in the manuscript, as it appears not to be available. It was therefore impossible to identify the region shown in figure 6.

We hope with the GitLab the reviewer will be able to identify this region. The copy is also present in Table S3 and S4.

L430-431: “Despite the presence of full-length *Copia* insertions in the genome, only spliced transcripts were uncovered in the long-read sequencing (Figure 5 and 7).” Indeed, when mapping to Copia, only 6 reads in testes and 1 read in ovaries correspond to putative nonspliced transcripts. These reads all cover part of the Copia intron and extend until the 3’-end of Copia. This suggests that they potentially correspond to full-length Copia transcripts, but that the reverse transcription step has not been completed. Moreover, this would also explain why full-length transcripts were not captured in general (see figure 1F, TE read size ti 2.5 kb).

Indeed, the following reads (e78971c4-2f89-45a9-b1c3-7c9b34a11a94, 26c75d8c-936f-4089-a697-dca14bc6537d) cover part of the intron of Copia$3L_RaGOO$10022428$10027467 and indeed extend until the 3’end of Copia. However, read 116c5610-5228-49a2-9fec-6f93a1a7ed3a also maps to the intron of this insertion, but without extending to the 3’end of the insertion. Additionally, these three reads map with the same alignment score to several Copia insertions. We therefore cannot know from which insertion the full-length transcript was transcribed. We added Figure S25 to discuss this point and changed the text accordingly (lines 589-591).