Authors response are in blue.

Revision round #2

Decision for round #2 : ***Revision needed***

Invitation to revise your manuscript

Dear Rita Rebollo,

The three reviewers have made available their comments on your revised manuscript. The referees' comments indicate that they are very satisfied with the modifications made. Yet, one reviewer has critical and well-argued comments that need to be answered in detail because they can potentially prevent recommendation.

I would especially invite you to answer the points raised on novel spliced TE isoforms :
  - TE insertions refer to recent insertions mediated by a functional TE

- TE annotation impact on your analysis

- discuss the impact of chimeric cDNA (most likely derived from coligation, something that is well known but insufficiently quantified) on your results. Possibly trying to figure out a quantification of chimeric cDNAs based on easily recognizable events from long reads derived from abundant transcripts.

I will send you the two tables mentioned by the reviewer in a separate e-mail.

With best wishes,

Nicolas Pollet

*by **Nicolas Pollet**, 20 Jan 2024 19:13*
Manuscript: **https://doi.org/10.1101/2023.05.27.542554**

version: 3

Dear Nicolas,

You will find attached the answers to the reviewers. We apologize for the delay as we redid all the figures of the manuscript since we have changed the TE annotation file as asked by one of the reviewers. The conclusions remain absolutely the same. Regarding the three main points you can find a brief answer below:

1) TE insertions refer to recent insertions mediated by a functional TE

TE insertions correspond to all the annotated TEs in the genome. While we do understand that for a *Drosophila* point of view, many of the insertions studied by the field are full-length, to consider TE insertions only those subsets is to ignore a big part of the literature, especially on other species. Therefore, TE insertions are all annotated TEs, whether they are fragmented or full-length. We did not explicitly add a sentence explaining this, but throughout the article we have distinguished between full-length TE analysis and old TE fragments.

2) TE annotation impact on your analysis

Of course, TE annotation impacts this and any TE analysis. We have changed the annotation to match the latest *Drosophila* TE annotation file, from Josefa Gonzalez's group (doi: 10.1038/s41467-022-29518-8) so as to have the most accepted *Drosophila* TE annotation but our conclusions remain the same.

3) discuss the impact of chimeric cDNA (most likely derived from coligation, something that is well known but insufficiently quantified) on your results. Possibly trying to figure out a quantification of chimeric cDNAs based on easily recognizable events from long reads derived from abundant transcripts.

Our data indeed contains chimeric cDNA, as we had acknowledged in the previous versions of this manuscript. This affects 20% of the reads. Our way of dealing with this issue is to restrict our analysis to the primary alignment of the read and ignore the supplementary alignments. When the read corresponds to a concatenation of several mRNAs, this will correspond to analysing the longest. To assess the impact of this strategy on our results, we did the following:

1- we recomputed the expression levels of TE copies discarding all chimeric reads. We found the quantifications were similar (Figure S5). No TE copy was more affected than others.

2- we computed, for each TE copy, the mean percentage of soft-clipped bases (i.e. bases that are located in the non-aligned portion of the read). We find that this percentage is 18% on average and is constant across all TE copies, indicating that the chimeric read issue equally affects all TE copies. We now output this metric in our pipeline and as a column in Table S3/S4. We hope this metric will be useful to the community. We also agree that the issue of chimeric reads is likely common and under-reported. A particularly high value of this metric for a specific copy could be an indication that the chimeric reads are not associated to a library preparation issue, but to a structural variation absent from the reference genome.

Review by anonymous reviewer 1, 11 Jan 2024 11:10

Rebollo et al., Identification and quantification of transposable element transcripts using Long-Read RNA-seq in Drosophila germline tissues.

– revised manuscript

The authors have put substantial effort in revising the manuscript, which, in my opinion, is now significantly improved. They have addressed sufficiently all my comments, especially by clarifying the issue of single- vs multi-mapping TE reads and by softening their conclusions on the biological significance of the non-replicated data. Thus, I maintain my feeling, that the strength of the manuscript lays in the technology used and the developed analysis tools. These, largely thanks to the comments of the other two reviewers, should now be much more reproducible and merit their sharing with the community.

Below, I have suggested a few last modifications:

2, 252, 266, 303 and throughout: For clarity, in a final version of the manuscript, the authors should better avoid using “long-read RNA-seq” and replace with “long-read cDNA-seq”, in order to avoid any confusion with ONT direct RNA sequencing approaches.

We agree that the term RNAseq is ambiguous as it refers both to dRNA-seq and cDNA-seq. We however prefer to keep it in our title so that readers interested in long-read RNAseq in general can find our work. We now clearly state in the abstract that we produce cDNA, which we think should prevent any confusion.

258: The statement “between ~1 to ~3 million reads per tissue” when there are only two samples in total (one per tissue) is misleading. Please rewrite.

Indeed. We have rewritten the sentence.

Fig 1A and S10-11: I strongly feel that the transcript length bias, which is well explained in the text, is very important for this study and for others that might want to perform similar type of analysis. Thus, graphs from figures S10 and S11 should be moved to the main fig 1 (either in addition to or replacing the panel 1A).

We have changed the figure as suggested.

288: Replace “to ensure” with “to check if”.

We have changed the text accordingly.

294: Move “(as suggested by the cDNA profile, Figure S1)” at the end of the sentence.

We have changed the text accordingly.

349: Please add a caution note reminding the reader that definite conclusions on the difference between sexes would require replicating the results.

We have added a sentence as suggested.

371: Suggestion: replace “single-copy” with “copy-specific”.

We have changed the text accordingly.

393: Should be: “unambiguously mapping”

We have changed the text accordingly.

. Fig 3B: The y-axis range for Pogo element graphs is too high. Add: “Each dot represents a unique genomic copy”.

We have changed the figure and text accordingly.

487: Should be: “ONT long-read sequencing detects”

We have changed the text accordingly.

488: Please add that, knowing the poor recovery of long transcripts, expression of longer TEs copies might be underestimated. This statement, present in lines 510-512, can be moved up.

We have moved up this statement accordingly.

501: Please change to “may unvail” (in regard to lack of replicates, frequent low read support and more extensive splicing analysis).

We have changed the text accordingly.

544: Should be: “only one or two”

We have changed the text accordingly.

602: Should be: “to specific copies of transposable elements”.

We have changed the text accordingly.

616: Should be: “retrotransposed”

We have changed the text accordingly.

610: Should be: “TE transcripts are spliced”

We have changed the text accordingly.

614: Please add: “While our results suggest that TE splicing could be prevalent, additional studies with biological replicates, high sequencing coverage and mechanistic insights into the splicing machinery will be needed to confirm our observations.” Or a similar statement.

We have added this statement as suggested.

We thank the reviewer for the comments and suggestions.

Review by Christophe Antoniewski, 13 Jan 2024 11:45

First of all, my apologies to the authors for taking a long time to re-evaluate the manuscript. I have read the revised version in detail and find that the authors have done an excellent job and addressed the vast majority of my comments satisfactorily. I particularly appreciate the effort on rewriting the methods and depositing them in a GitLab repository.

I think that the article is now reproducible and above all that it can be useful to a large community of biologists, well beyond Drosophila researchers, including researchers working in human genomics.

It was worth it, wasn't it?

Yes it was ! We thank the reviewer for the comment and the suggestions.

Review by Silke Jensen, 10 Jan 2024 02:56

The supplementary excel files that are cited in the review document, destined to the authors, will be sent by e-mail.

PCI – Round # 2 Rebollo et al.

The authors have addressed most of my concerns correctly and changes have been made to the manuscript accordingly. Unfortunately, with the new examples of TE annotation and putative TE splicing given in the authors' response, I have now identified a significant problem with TE annotation. I sincerely regret not having identified this problem during the first revision cycle. Indeed, the TE annotation method seems to be wrong for many TE sequences. This erroneous annotation leads to significant errors in the interpretation of the results, particularly with regard to the putative splicing of transcripts that have been erroneously assigned to TEs.

We have now changed the TE annotation file to the latest annotation provided by Josefa Gonzalez's group (doi: 10.1038/s41467-022-29518-8). We have checked all the instances where the reviewer had a problem with the annotation and have compared the different annotation files with the previous analysis. There are not a lot of changes and the overall annotations are very similar. In addition, the results of the manuscript remain the same. The reviewer should bear in mind that the TE copies studied here are all copies and not only young, full-length copies, extremely well annotated. The copies are also old copies, fragmented, interrupted, and decayed. Those copies are usually harder to annotate and from *Drosophila* to humans, no TE annotation file will be perfect. What is important here is to be able to differentiate between a TE that is contributing to a transcript, may that TE be in itself a transcription unit, or may it be part of other transcripts. This manuscript focuses on all those TEs. We did a subset analysis on younger full-length copies in the last two sections, as already seen in the previous versions of this manuscript, which we think corresponds to the point of view of the reviewer. Therefore, all the comments below that address TE annotation, should take into consideration the fact that in the manuscript we consider all TE copies, while the reviewer only considers young well-annotated TE copies.

Putative TE splicing events - my answers concerning the examples given in the IGV figures in the "author-reply_16dec2023.pdf" document:

I am sure the authors are convinced by their results on what they think is TE splicing and I regret to say that the data do not support the splicing events they attribute to TEs. The results of my investigations concerning the examples shown in the authors' reply are detailed below and in the supplementary attached documents.

Looking at the examples of TE annotation given in the response to reviewers, I finally discovered that the essential problem of the manuscript is the annotation of TEs. TEs were annotated using "RepeatMasker with DFAM dataset from D. melanogaster (-species Drosophila) TE copies (Dfam_3.1) and then used OneCodeToFindThemAll (Bailly-Bechet et al., 113 2014)" as indicated in the Methods. As shown in the attached excel file with CENSOR (Repbase, https://www.girinst.org) and RepeatMasker analyses of the example sequences given for TE splicing in the response to the reviewers, these annotated "TE insertions" present problems ("TE-annotation-analyses.xlsx"). It is clear that the annotation of transposable elements obtained with the method used here is not sufficiently precise to allow deduction of the splicing of transposable elements. Indeed, regions annotated as transposable elements not only contain regions corresponding to the respective TE, but may also contain other TE fragments or, worse still, gene segments or unknown sequences. It is possible that the use of the OneCodeToFindThemAll tool has made the annotation even worse, as there are examples of TE annotation where distant fragments of TEs from the same family have been merged to give the impression that they are one large TE, which is clearly not the case (e.g. TARTA$Y_RaGOO$1207965$1236166 in the authors' response, see also the attached file "TE-annotationanalyses. xlsx"). In most cases, when authors write in the manuscript or supplementary material that there is a "TE insertion", when I inspected the corresponding RepeatMasker results, there were only fragments of remnants of ancient TE invasions, highly mutated, rearranged and with deletions. These ancient TE fragments cannot be considered as "TE insertions". In my opinion, the term "TE insertion" suggests that a genomic element corresponds to a recent insertion of a presumed functional TE. However, this is clearly not the case for the elements cited as examples. The terms 'TE fragments' or 'TE remnants' would be more appropriate for all TEs shown in the figures in the response to reviewers. In addition, each TE fragment should be annotated separately, especially to draw conclusions about TE splicing. This would avoid considering regions that are intermingled with TE fragments as being TEs.

As previously stated, the term “TE insertion” which can also be “TE copy” corresponds to any TE annotated sequence in the genome. Therefore, there are TE insertions that are fragments and have annotations that are at the limit of the tools used. The terms TE fragments and TE remnants are, in this manuscript, considered TE insertions /copies. When talking about young or intact copies, we addressed them as young or full-length copies. Instead of using One Code, we have used RepeatCraft to merge the TE annotations from Repeat Masker. We have also used the latest TE annotation file from the Gonzalez’s lab (doi: 10.1038/s41467-022-29518-8).

Examples of TE annotations shown in "author-reply_16dec2023.pdf":

The IGV figures in the answer to authors show several annotated putative TEs and gene transcripts:

Figure 1: POGO$3L_RaGOO$9733928$9735150 FBtr0300688: Dmel\CG10809-RB, CG10809 is at R6 3L:9,857,383..9,859,889 [-]

Figure 2: ROO$3R_RaGOO$15240450$15245518 FBtr0335424: Dmel\CG3992-RG, = Dmel\srp-RG, srp is at R6 3R:15,986,152..16,004,085 [+] FBtr0335423: Dmel\ CG3992-RF, = Dmel\srp-RF, srp is at R6 3R:15,986,152..16,004,085 [+]

Figure 3: TAHRE$2R_RaGOO$1145909$1151824 no annotated gene or transcript

Figure 4: HETA$X_RaGOO$85920$94840 no annotated gene or transcript

Figure 5: TART-A$Y_RaGOO$1207965$1236166 no annotated gene or transcript

Figure 6: Gypsy12$Y_RaGOO$361225$363385 mRNA_17639: I didn't find the corresponding record for this annotated mRNA

Figure 7: G5A_DM$2R_RaGOO$4442347$4444566

Analyses of the regions shown in Figures 1 to 7 in the answer to reviewers:

I analysed the regions shown in the seven figures of the response to raters using CENSOR (Repbase, https://www.girinst.org) and RepeatMasker (https://www.repeatmasker.org). Detailed results can be found in the attached document "TE-annotation-analyses.xlsx" sheet "author-reply examples".

Figure 1: region of POGO$3L_RaGOO$9733928$9735150 I cannot agree with the following statements of the authors in the answer to reviewers concerning the reads that map POGO$3L_RaGOO$9733928$9735150. Citation: " 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." In fact, the IGV image in the figure shown does not reveal all the information about the parts of the reads that map outside the annotated regions. When I analysed 26 reads that map to

POGO$3L_RaGOO$9733928$9735150, I found that 8 reads also map to other distant regions of the genome over hundreds of base pairs, while 18 reads only map to POGO$3L_RaGOO$9733928$9735150 (see attached excel file "TE-annotation-analyses.xlsx"). It seems that these eight reads are chimeric reads, i.e. fusion products of different cDNAs from different genomic regions, see also below for other reads. So I don't quite understand how the authors found "56, 7 and 4 bp average reads outside" this copy of POGO. Therefore, it would be wise to check the method used to measure the number of bp corresponding outside a TE copy, and in particular to check how many bp do not correspond to the analysed region at all (i.e. soft- and hard-clipped sequences). Tables S3 and S4 show the mean number of bp of reads mapping outside TE copies, but it is not clear what the data correspond to and how they were generated (columns G, H and I of the tables).

Indeed, the reviewer is correct. Our dataset includes chimeric reads, as previously stated in the older versions of the manuscript, which likely derived from coligation during library preparation. We now more clearly quantify this issue. We added a column in Tables S3 and S4 which provides, for each TE copy, the mean percent of soft-clipped bases. For this specific POGO copy, we obtain 17.5%, which is comparable to what is obtained for other TE copies (the average is 18.5%). We report separately the number of bases which are aligned at the genomic location of the TE copy, but that lie outside of the TE annotation. This is the column “Mean bases outside TE”. This column does not count soft-clipped bases. We have now clarified the meaning of each column of Table S3/S4 in the git repository.

There are many reads which map distinct non-contiguous regions of the assembled genome (GCA_927717585.1.contig_named.fasta). These reads are composed of 2 or more parts that map the assembled genome with the same high mapping quality (MAPQ 60) at distant loci. This suggests that the assembled genome does not contain the entire corresponding genomic region or that they are chimeric reads generated by ligation of cDNAs of different origin during library preparation. The protocol used should be checked by the authors to assess whether this is a possible event. When 8 of these potentially chimeric reads were mapped to the raw reads of strain dmgoth101, none mapped to the full length of a genomic read, suggesting that these are indeed chimeric reads from different fused transcripts/cDNAs (see attached excel file "chimeric-reads-analyses.xlsx"). The apparent occurrence of chimeric reads is an important finding as it presents a particular challenge for data analysis. I think the authors should discuss this problem somewhere in the manuscript to inform future users of this cDNA sequencing technique. Is there any adapter clipping that could solve this problem? I have not found any mention of adapter clipping of the cDNA reads in the manuscript.

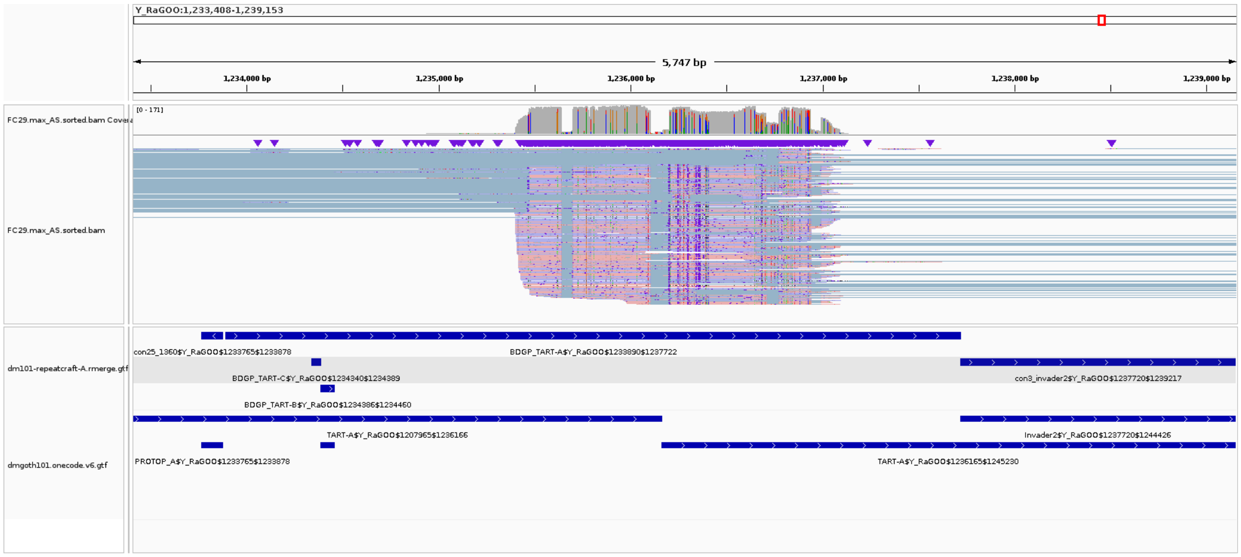
Indeed, our dataset contains chimeric reads likely derived for coligation during library preparation. We agree that this poses an additional challenge for data analysis. We address this by focusing on the primary alignment of each read. In practice, if a read corresponds to the concatenation of several mRNAs, our strategy results in analyzing the longest mRNA and disregarding the others. We also performed our analysis without chimeric reads (instead of keeping the primary alignment and neglecting the others) and the results were unchanged (Figure S5). We also computed the average number of soft-clipped bases for each TE insertion and found that no TE copy exhibited a particularly higher number, which would be an indication that the reads are stemming from another TE copy, absent from the reference genome.

Figure 2: ROO$3R_RaGOO$15240450$15245518 The worst example of an erroneous TE annotation is ROO$3R_RaGOO$15240450$15245518 : ROO$3R_RaGOO$15240450$15245518 overlaps several exons of an annotated gene transcript. This is impossible because TEs and genes are distinct genetic elements. In fact, there are only too small fragments of 77 bp and 47 bp in this annotated "ROO" sequence, with a sequence divergence of 23% and 12.8% compared with the Dfam reference ROO sequence (RepeatMasker analysis). No ROO-type sequences were found by CENSOR. It is certainly not an ROO. In addition, the small sequences that RepeatMasker found to be linked to the ROO are different from the regions where the introns are located. The splicing events detected clearly correspond to splicing of the transcript of the annotated gene shown in the figure, and not to an ROO. I would like to point out here that in fact the 13 sequences that are annotated as "ROO" and map more than 10 unique reads contain only very small fragments of ROO-like sequences, the largest ROO-like fragment being 367 bp long (see RepeatMasker analyses in the attached "TE-annotation-analyses.xlsx" sheet "TEs with over 10 uniq reads"). None of these sequences can therefore be considered as ROO. This is also true for ROO$2R_RaGOO$14213942$14227652, which is discussed in the author's response.

ROO$3R_RaGOO$15240450$15245518 and ROO$2R_RaGOO$14213942$14227652 are not present anymore in the new annotation.

Figure 5: TART-A$Y_RaGOO$1207965$1236166 citation from "author-reply_16dec2023.pdf": "The expression of TART-A$Y_RaGOO$1207965$1236166 is supported by 242 reads, 86% of which are spliced and span several introns. The transcription unit overlaps two annotated TART-A insertions." The annotated sequence TART-A$Y_RaGOO$1207965$1236166 is a 28.2 kb sequence, which is flanked by some mutated and rearranged TART-A fragments, but 23.4 kb of this sequence is not TART-related at all. Reads suggesting splicing events are found in the region Y_RaGOO:1.235.4001.237.100. There are indeed also TART-A-related sequences, but these are fragments of ancestral TART-related elements. Furthermore, only one of the putative introns has splice donor and acceptor sites: on the negative strand, the Y_RaGOO:1,236,941-1,236,993 region; GT-AG being in the opposite orientation to TART. It is more likely that the reads shown in the figure come from regions with TART-related sequences that are absent from the genome assembly, and which have small deletions, here resembling introns in the IGV images. In fact, HeT-A, TART and TAHRE are mainly located in telomeric heterochromatin, which is generally absent from genome assemblies (because it is difficult to sequence and difficult to assemble). Consequently, all analyses of these telomeric elements are heavily biased by their low presence in genome assemblies. On closer inspection of the read mapping shown in the figure for TART-A$Y_RaGOO$1207965$1236166, it can be seen that almost all reads assumed to be spliced are also mapped elsewhere in the genome (clipped sequences on the left and right of the reads), indicating that these parts of the reads do not originate from the region shown in the figure. Further investigation is required to determine the origin of these reads. It may be useful to map the cDNA reads to the raw reads obtained from genome sequencing.

The annotation of TART-A$Y_RaGOO$1207965$1236166 has changed. The new annotation BDGP_TART-A$Y_RaGOO$1233890$1237722 coincides better with the aligned reads as can be seen in the figure below.


Y_RaGOO:1,233,408-1,239,153
5,747 bp
1,234,000 bp
1,235,000 bp
1,236,000 bp
1,237,000 bp
1,238,000 bp
1,239,000 bp
FC29.max_AS.sorted.bam Covera
[0 - 171]
FC29.max_AS.sorted.bam
dm101-repeatcraft-A.rmerge.gtf
con25_1360$Y_RaGOO$1233765$1233878
BDGP_TART-A$Y_RaGOO$1233890$1237722
BDGP_TART-C$Y_RaGOO$1234340$1234389
con3_invader2$Y_RaGOO$1237720$1239217
BDGP_TART-B$Y_RaGOO$1234386$1234460
TART-A$Y_RaGOO$1207965$1236166
Invader2$Y_RaGOO$1237720$1244426
dmgoth101.onecode.v6.gtf
PROTOP_A$Y_RaGOO$1233765$1233878
TART-A$Y_RaGOO$1236165$1245230

Interestingly, as the reviewer pointed out, the consensus sites flanking the gap in the alignment are CT-AC, and not GT-AG. This suggests that the transcript is antisense. Our long read data is unstranded, but we could verify, using our stranded short read data, that this was indeed the case. Examples are presented in Figures S32 and S33. We now discuss separately CT-AC and GT-AG cases in the manuscript. We agree that the analysis of telomeric TEs is delicate because telomeres are often poorly assembled. For the specific case of BDGP_TART-A$Y_RaGOO$1233890$1237722, we find that the percent of soft-clipped bases is 14%, which is similar to what is obtained for non-telomeric TE copies. We cannot exclude that the reads stem from a non-assembled region (although we should probably see a higher mismatch rate), but we do not think that our interpretation of the splicing of the antisense transcript is compromised.

Figure 6: Gypsy12$Y_RaGOO$361225$363385 Gypsy12$Y_RaGOO$361225$363385 is only a Gypsy12 LTR (76% identity), not a full Gypsy12.

Indeed, this is not a full-length TE. We did not include it in Figure 6 of our paper, which is focused on full-length TEs. We however include it in Figure 7, because it is transcribed and spliced (with a GT-AG consensus).

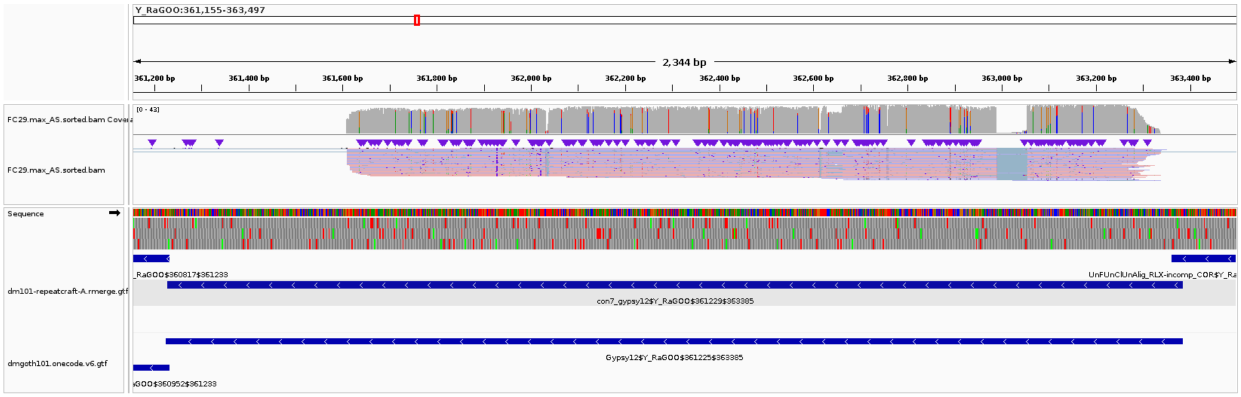


Figure 7: region of G5A_DM$2R_RaGOO$4442347$4444566 Citation answer to reviewer: "The expression of G5A_DM$2R_RaGOO$4442347$4444566 is supported by 64 reads, 32% of which contain gaps, but without GT-AG flanking sites (see figure below). Those could be noncanonical introns, genomic deletions, or mis-alignment of the reads due to a gap in the genomic assembly." Indeed, I agree with the authors. Reads that do not map to the full length on the assembled genome and have gaps (putative introns) may also come from related TEs or related repeat sequences that are not in the assembled genome: The Drosophila line analysed here was not isogenic and can be very heterogeneous with multiple structural variations. Minimap2 will then display the best alignment on the assembled genome, but this will not necessarily be the correct genomic sequence from which the transcript originated. A genome assembled de novo from a non-isogenic lineage always contains only part of the true genomic sequences. But I don't think this point is addressed in the manuscript, although it seems important.

We consider the lines to be nearly isogenic because they are 30 generation sib-mated.

Analyses of some putative chimeric reads are shown in the attached "chimeric-reads-analyses.xlsx".

Analyses of all TEs in the supplementary file "media-1.xlsx" sheet "TableS3_testes" with more than 10 uniquely mapping reads by RepeatMasker can be found in the attached sheet "TE-annotationanalyses.xlsx" sheet "TEs with over 10 uniq reads". These analyses show that many annotated TEs contain sequences that do not correspond to the annotated TE.

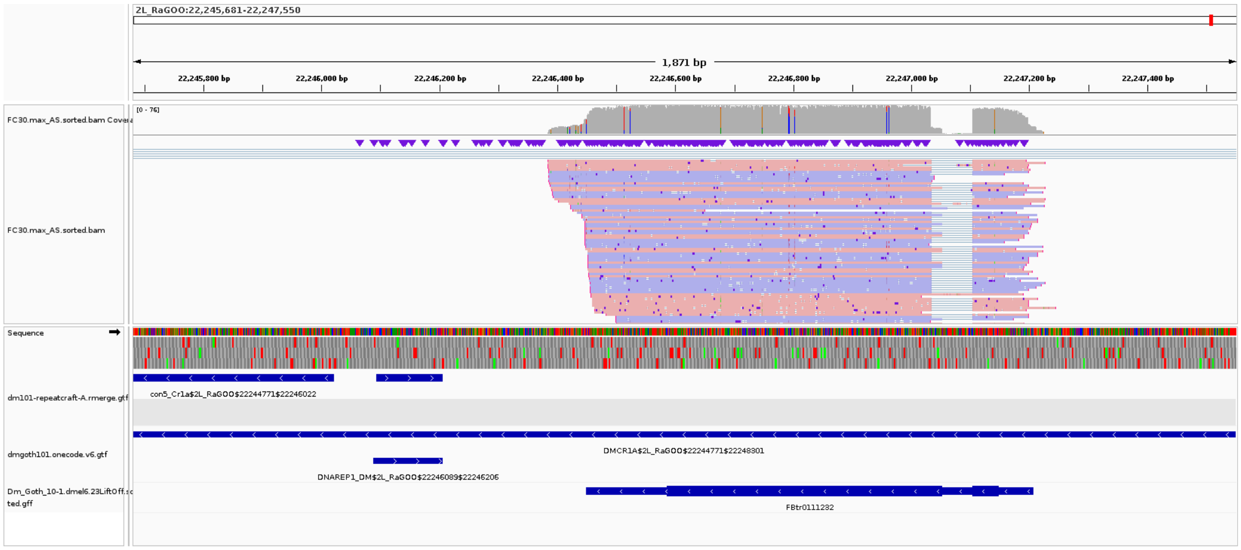
Examples of TE annotation from supplementary "media-2.pdf":

Figure S3: "Figure S3. A. Example of a read mapping to four locations on the genome. These four locations are insertions of Gypsy7. The read aligns to these four locations with a score of 861, 859, 859, 853." Firstly, it is not clear what these scores correspond to since these are not mapping quality scores ("MAPQ") of Minimap2. Secondly, Dfam GYPSY7 is 5486 bp long. In the figure the putative Gypsy7 insertions are only around 4.5 kb long. I inspected the region shown in Figure S3 (3L_RaGOO:25,720,001-25,744,000) using RepeatMasker and CENSOR and I found the following (see attached "TE-annotation-analyses.xlsx"): The four Gypsy7-like elements are in fact tandem duplications of an ancestral copy of Gypsy7 with low sequence identity with Gypsy7 (93% to 95%). Each duplicated Gypsy7-like copy is incomplete, all 4 copies have the same internal deletions and only one LTR for each copy. The only read mapping to this region also maps another genomic location on Chromosome 2L.

We agree that these four Gypsy7-like copies are not full-length. Our point here is that one of these copies is expressed because we have a read that maps to these four locations. In order to decide which location is most likely expressed, we consider the alignment score (AS tag in the minimap2 output). It turns out that the highest score is 861 and corresponds to a mapping quality of 1. The other scores are 859, 859 and 853 and the mapping quality is 0. We note that this read also generates a supplementary alignment at position 2L_RaGOO:11979158. This supplementary alignment has an alignment score of 538 and a mapping quality of 60. It corresponds to the complete transcript of the gene CG6770. This gene is highly expressed (850 reads). We think that read 1bf5274e-6531-4146-b59e-fb6221889f4c is a chimeric read generated during library preparation and resulting in the ligation of two independent transcripts, one from gene CG6770 located at position 2L_RaGOO:11979158 and one from a Gypsy7-like copy located at position 3L_RaGOO:25724750-25729257.

Figure S7: The annotated TEs correspond to diverse fragments of ancestral TE sequences. The annotated gene transcript FBtr0111232 corresponding to Dmel\CG40439-RA is located approximately at 2L_RaGOO: 22,246,400-22,247,200. There is no TE-like element detected at the location of this annotated FBtr0111232 transcript. Thus, there is not conflict between TE-mapping and transcript-mapping. The reads clearly stem from the gene.

Indeed, using the new annotation, there is no conflict anymore at this locus because the annotation of DMCR1A is not present anymore in the new annotation as shown in the figure below.


2L_RaGOO:22,245,681-22,247,550
1,871 bp
22,245,800 bp
22,246,000 bp
22,246,200 bp
22,246,400 bp
22,246,600 bp
22,246,800 bp
22,247,000 bp
22,247,200 bp
22,247,400 bp
FC30.max_AS.sorted.bam Covera
[0 - 76]
FC30.max_AS.sorted.bam
Sequence
dm101-repeatcraft-A.rmerge.gtf
con5_Cr1a$2L_RaGOO$22244771$22246022
dmgoth101.onecode.v6.gtf
DMCR1A$2L_RaGOO$22244771$22248301
DNAREP1_DM$2L_RaGOO$22246089$22246206
Dm_Goth_10-1.dmel6.23LiftOff.sc
ted.gff
FBtr0111232

We now have another example for Figure S7 to illustrate how our method assigns reads in the case of overlapping features.

Figure S8: Figure S8 shows a possible conflict of assigning reads to an annotated TE or to a gene. But in fact, this conflict can be avoided by different TE annotation. My CENSOR and RepeatMasker analyses of the concerned region shows the following (see attached "TE-annotation-analyses.xlsx"): The annotated TEs correspond to diverse fragments of ancestral TE sequences. The annotated gene transcript FBtr0111232, corresponding to Dmel\CG40439-RA, is located approximately at 2L_RaGOO: 22,246,400-22,247,200. There is no TE-like element detected at the location of this annotated FBtr0111232 transcript. Thus, in fact there is not conflict between TE-mapping and transcript-mapping. The reads clearly stem from the gene.

Indeed, using the new annotation, there is no conflict anymore at this locus. We now present another example in Figure S8 to illustrate how our method assigns reads in the case of overlapping features.

To avoid this type of conflict, each TE fragment must be annotated separately. Merging distant TE fragments of the same element type makes no sense in this type of analysis. TEs must not overlap gene exons. They can only be located inside introns (which happens quite often). The best thing to do would probably be to use only the annotation of exons to assign reads to genes and the separate annotation of each TE fragment, allowing TE fragments to merge only if they are contiguous or separated by a very small distance. Such a method will allow reads to be clearly assigned to genes or TEs without conflict in most cases, and it is very likely that most of the splicing events detected here will then be assigned to genes or unannotated regions and not to TEs, as illustrated by the case of ROO$3R_RaGOO$15240450$15245518, TAHRE$2R_RaGOO$1145909$1151824 in the author reply and others (see my attached excel file "TE-annotation-analyses.xlsx" for detailed analyses).

Conclusions concerning TE annotation:

In conclusion, it is not sound at all to conclude for any "novel spliced TE isoforms" from such a imprecise and even erroneous TE annotations.

The results of mapping cDNA reads to repeat-rich regions of a de novo assembled genome are very complex. Even if the genome is of high quality, especially repeat-rich regions are not fully assembled, which may lead to unexpected mapping results. In addition, it seems that there are multiple chimeric reads resulting from fusion of different transcripts mapping distant loci. It is not easy to draw conclusions about the origin of reads and splicing without closer inspection of the mapping results.

In summary, the only convincing splicing events that I can find in the manuscript are the ones shown in Figure 8 and in supplementary Figures S26-S29 (see also below and RepeatMasker analyses in the attached "TE-annotation-analyses.xlsx"). The suspected TE splicing events clearly need more investigation due to the erroneous TE annotation that I detected in most cases shown in the author reply. I sincerely regret, but to my opinion, most of the analyses of splicing events assigned to TEs should be deleted, notably the ones in Figure 7, or re-done with different, more accurate annotation of TEs and gene exons (not of the entire transcripts), avoiding redundancy between TE and exon annotation. The apparent occurrence of chimeric reads originating from different transcripts is an additional challenge that should be considered and discussed.

We are now using a new annotation, which we think addresses the points raised by the reviewer. We regret to have used ROO$3R_RaGOO$15240450$15245518 as an example of splicing in our previous response to reviewers. It was a poorly chosen example and we think this choice generated a lot of confusion. The cases of splicing we present now clearly correspond to cases that are attributed to TE copies and not genes. We specifically require that the gap is located within the annotated TE boundaries to be considered as a potential intron. We further make the distinction between CT-AC and GT-AG cases. We chose to restrict Figure 7 to GT-AG cases as we think those are the easiest cases to understand. CT-AC cases are also more delicate because they mostly concern telomeric TEs, which are not trivial to study as telomeres are hard to assemble. We provide as Table S7 and Table S8 the full list of TE copies containing all suspected introns.

We acknowledge that mapping long reads stemming from repeats is a challenging task when dealing with incompletely assembled genomes. Genome graphs may be interesting in such cases. We also clearly present the issue of chimeric reads, which is an issue for the community. Our strategy of focusing on the primary alignment could be improved. Ideally, this issue should be addressed at the base calling step. Identifying adapters after the base calling is not trivial, because they are not located at the end of the read.

My comment concerning Figure 2A:

citation "author-reply_16dec2023.pdf" document: my comment: "Figure 2A: It would be useful to also present the TE transcriptional landscape obtained with shortread sequencing to compare the results obtained by the 2 technologies, ONT and Illumina sequencing." author reply: "The figure can now be appreciated in the supplementary materials (Figure S17) and has also been discussed in the manuscript (lines 352-356)." That is a good thing but I didn't find the discussion in lines 352-356 (and no changes tracked in blue in this part). The problem with Figure S17 and others is the impact of the problems of erroneous TE annotation highlighted above. In relation with Figure S17, in lines 240-250 in "track_changes_16dec2023.pdf": spelling of the tool "TEtranscript" should be always the same (without space).

The spelling is corrected and the annotation has been addressed on the other points of this rebuttal.

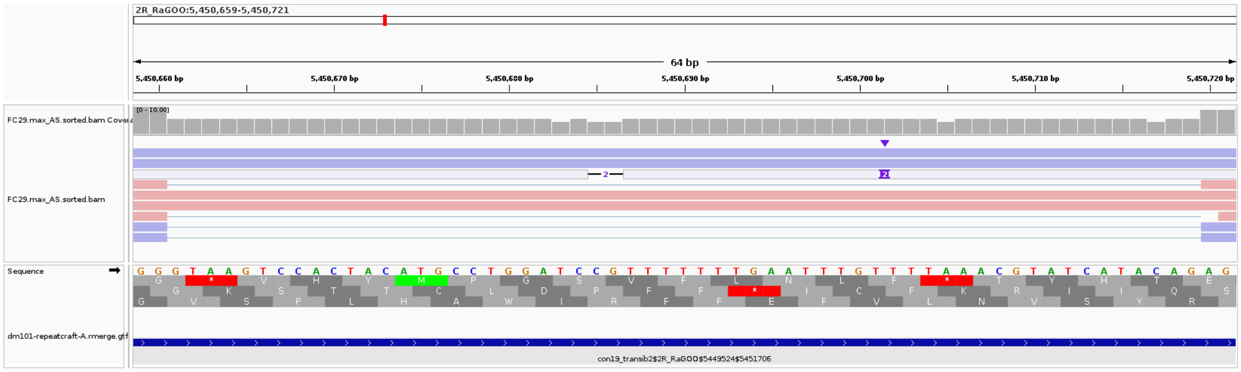
New supplementary Figures S26-S29:

I thank the authors for these new figures. The splicing events shown in these figures are indeed convincing. The problem is that they concern only Copia elements, one POGO and one 1731 copy. The legend to Figure S26 is incomplete in the downloaded version of the "media-2.pdf" file: "Figure S26: Zoom on the donor site and acceptor site of the intron of 9"

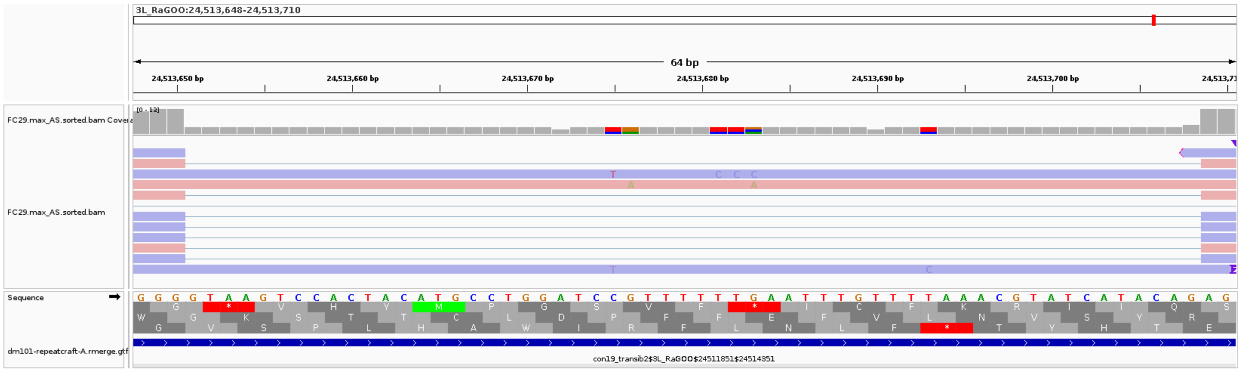
Indeed, the legend was truncated and it is now fixed.

Figure 6: I thank the authors for this new figure which is quite informative and joins two more examples of putative TE splicing with the shown MAX and Mariner-2 copies.

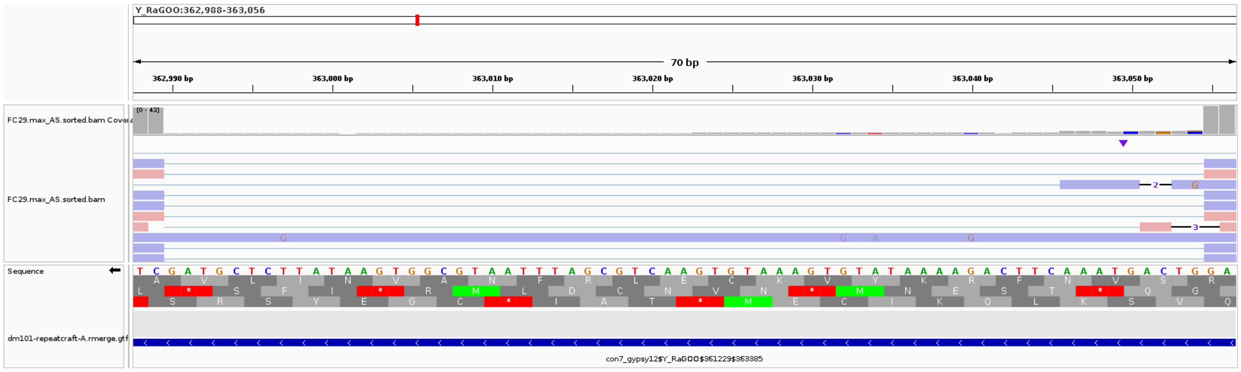
We understand that the reviewer is primarily interested in introns located in full-length TEs. We would however like to argue that identical introns detected in older TE copies may also be interesting since they are an indication of the presence of the intron in the ancestral active copy. For the few cases (*con19_transib2*, *con7_gypsy12*) where two independent copies of the same TE family have introns, we can verify that the intron is indeed the same in the two copies. Below are examples of GT-AG introns located in two independent copies of *con19_transib2*. The length and the sequence of the intron is the same. We then present the same result for two GT-AG introns located in two independent copies of *con7_gypsy12*.


2R_RaGOO:5,450,659-5,450,721
64 bp
5,450,660 bp
5,450,670 bp
5,450,680 bp
5,450,690 bp
5,450,700 bp
5,450,710 bp
5,450,720 bp
FC29.max_AS.sorted.bam Covera
[0 - 10.00]
FC29.max_AS.sorted.bam
2
2
Sequence
G G G T A A G T C C A C T A C A T G C C T G G A T C C G T T T T T T T G A A T T T G T T T A A A C G T A T C A T A C A G A G
dm101-repeatcraft-A.rmerge.gtf
con19_transib2$2R_RaGOO$5449524$5451706

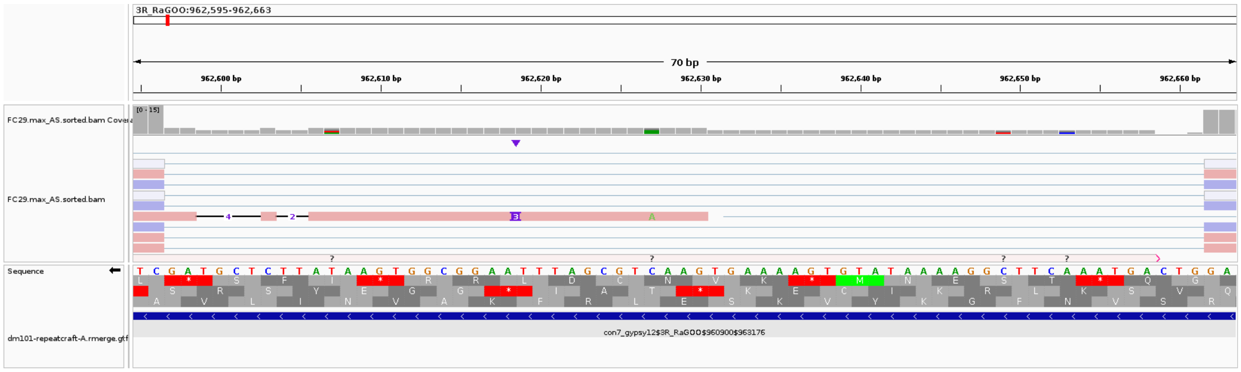
Intron detected in con19_transib2$2R_RaGOO$5449524$5451706


3L_RaGOO:24,513,648-24,513,710
64 bp
24,513,650 bp
24,513,660 bp
24,513,670 bp
24,513,680 bp
24,513,690 bp
24,513,700 bp
FC29.max_AS.sorted.bam Coverage
[0 - 13]
FC29.max_AS.sorted.bam
Sequence
GGGGTAAGTCCACTACATGCCTGGATCCGTTTTTTGAATTTGTTTTAAACGTATCATACAGAG
dm101-repeatcraft-A.rmerge.gtf
con19_transib2$3L_RaGOO$24511851$24514351

Intron detected in con19_transib2$3L_RaGOO$24511851$24514351


Y_RaGOO:362,988-363,056
70 bp
362,990 bp
363,000 bp
363,010 bp
363,020 bp
363,030 bp
363,040 bp
363,050 bp
FC29.max_AS.sorted.bam Coverage
[0 - 43]
FC29.max_AS.sorted.bam
2
3
Sequence
TCGATGCTCTTATAAGTGGCGTAATTTAGCGTCAAGTGTAAAGTGTATAAAAGACTTCAAATGACTGGA
dm101-repeatcraft-A.rmerge.gtf
con7_gypsy12$Y_RaGOO$361229$363385

Intron detected in con7_gypsy12$Y_RaGOO$361229$363385


3R_RaGOO:962,595-962,663
70 bp
962,600 bp
962,610 bp
962,620 bp
962,630 bp
962,640 bp
962,650 bp
962,660 bp
FC29.max_AS.sorted.bam Coverage
[0 - 15]
FC29.max_AS.sorted.bam
4
2
3
Sequence
TCGATGCTCTTATAAGTGGCGGAATTTAGCGTCAAGTGAAAAGTGTATAAAAGGCTTCAAATGACTGGA
dm101-repeatcraft-A.rmerge.gtf
con7_gypsy12$3R_RaGOO$960900$963176

Intron detected in con7_gypsy12$3R_RaGOO$960900$963176

Minor comments: References like FBtr0114142 and FBtr0346695 (Example in Figure S8, and as in all IGV figures shown in the supplementary) do not correspond to genes but to transcripts. Please correct this in the corresponding text and legends.

Corrected.