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.

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. TART-AY_RaGOO1207965$1236166 in the authors' response, see also the attached file "TE-annotation-analyses. 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.

<u>Examples of TE annotations shown in "author-reply_16dec2023.pdf":</u>

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:
HETAX_RaGOO85920$94840
no annotated gene or transcript

Figure 5:
TART-AY_RaGOO1207965$1236166
no annotated gene or transcript

Figure 6:
Gypsy12Y_RaGOO361225$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).

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.

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.

Figure 5: TART-AY_RaGOO1207965$1236166

citation from "author-reply_16dec2023.pdf":

"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 annotated sequence TART-AY_RaGOO1207965$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:1.235.400-1.237.100. 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-AY_RaGOO1207965$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.

Figure 6: Gypsy12Y_RaGOO361225$363385

Gypsy12Y_RaGOO361225$363385 is only a Gypsy12 LTR (76% identity), not a full Gypsy12.

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 non-canonical 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.

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-annotation-analyses.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.

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.

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.

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.

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 short-read 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).

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"

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.

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.