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).

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.

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.

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).

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

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 <1 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.

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).

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.

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.

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.

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.

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.

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.

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.

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.

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?

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.

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.

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.

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?

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.

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.

Legend Figure 1, L257: “The majority of transcripts recovered are full-length.” The term “full-length” 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.

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.

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.

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

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

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.

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. Ditto for testes, re-analyze using non-parametric statistical methods would be useful.

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”?

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

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

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

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

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.

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.

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.

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 non-spliced 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 < 2.5 kb).