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.

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.

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

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

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

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.

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.

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

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

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…

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

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.

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…

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).

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.

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 rational of the smallest symmetric difference is not so clear in the example given in Figure S8.

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.

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.

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 ?

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.

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.

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.

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

Line 383. exemptions
I think the authors mean exceptions ?

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 ?

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

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.