Review of manuscript entitled
"Pipeline to detect the relationship between transposable elements and adjacent genes in host genome"

Caroline Meguerditchian, Ayse Ergun, Veronique Decroocq, Marie Lefebvre, and Quynh-Trang Bui

General comments:

The manuscript entitled "Pipeline to detect the relationship between transposable elements and adjacent genes in host genome" by Caroline Meguerditchian, Ayse Ergun, Veronique Decroocq, Marie Lefebvre, and Quynh-Trang Bui describes a pipeline destined to report TE and adjacent gene distribution in a host genome. The pipeline needs as input a gff annotation file of the analysed genome and a TSV file containing information about TE annotation. The results are provided in a TSV file. In addition, three R scripts create graphs and CSV files from the latter TSV file, giving some statistical outputs. Examples of 2 graphs are presented in Figures 2 and 3. The workflow of the pipeline is shown in Figure 1. The pipeline and a short manual are freely available at https://github.com/marieBvr/TEs genes relationship pipeline.

The manuscript is quite well written with some edits to be made according to my comments below. I estimate that the pipeline is useful for researchers who study transposable elements and/or gene expression if it can be used for all genomes. I guess this is the case but it should be clarified. In my opinion the manuscript can be published in PCI Genomics with some minor edits and clarifications according to my comments below.

Comments concerning the text:

2 - Introduction:

"... the TEs can regulate gene expression by modifying the closest reading frames into pseudogenes, ..."

I'm not sure of what the authors mean here. They should explain : Do they mean TE insertion into reading frames ? Moreover, to my opinion, genes may become pseudogenes but simple reading frames cannot become pseudogenes.

3 - Materials and methods

3.1 General workflow

Please give a reference or link for the "Apricot dataset".

Even if they have been given elsewhere in the manuscript, please give the references for the different published tools in the "Materials and methods" section as well.

" ... sorting out TEs in order to increase information retrieved from their position in the genome."
It's not clear what this means, please be more precise.

" ... subset and superset genes."
Do the authors mean overlapping genes here ? It seems not very clear to me. Could you please define "subset" and "superset" genes ?

"... to visualize TE-coding sequence relationships ..."
Do the authors mean TE-gene relationships here ? This would be more consistent with the description of their method before.

3.2 Implementation

"... in an downstream location ..."
Please replace "an" with "a".

"... overlapping the the downstream part of the TE."
Please replace "the the" with "the".

"... searches for gene, which is either a subset or a superset of the TE."
As above, I'm not sure of "subset" and "superset" mean here. Please explain it at the first use of these terms. Please replace "gene" with "a gene".

"... the distance between TE ..."
Please replace "TE" with "a TE".

"... how many TEs have an overlap with genes, both upstream and downstream."
This is not clear: If genes are up- or downstream, they should not overlap the TE. This is confusing, please specify what is meant here.

"... the number of TE ..."
Please replace "TE" with "TEs".

4 - Use case

"... on Figure ..."
Please replace with "... in Figure ...".

5 - Conclusion

"... the ability to change their position within the genome."
As transposable elements comprise retrotransposons, which move through a copy-and-paste mechanism, "change their position" should be replaced with "move".

"These mobile elements play an important role in gene regulation ..."
I think this should be tuned down. In the abstract, the authors write "Transposable elements can regulate and affect gene expression ..." which seems more adequate to me. Please replace "play" with "can play" for example.

"This pipeline could be useful to reveal potential effects of TEs on gene expression as well as on the study of specific gene function."
This is overstated. The pipeline doesn't "reveal" effects on gene expression since it reports TE-gene relation within the genome. There are no expression data analysed here. Please replace "reveal" with "subsequently analyse" for example. Also, it doesn't "reveal potential effects of TEs on the study of specific gene function" but allows to subsequently analyse "potential effects of TEs on specific gene function". Please delete "the study of" or reformulate the sentence.

References

Reference [1] "M. Barbara", please replace with "B. McClintock".

Comments concerning the figures:

Figure 1:

According to the workflow in Figure 1, upstream genes are sorted out first then downstream genes and at last "superset/subset" genes. The authors should explain what this means: Does it mean that genes that were found upstream of a TE will not be considered further in the search for downstream genes ? Or maybe these different steps proceed with the same input files and are not subsequent steps. This should be clearly stated and the figure 1 changed accordingly if these are not subsequent steps using the output of step 1 for step 2 etc.

Figures 2 and 3:

It is not clear what the "downstream overlapping gene and upstream overlapping gene" correspond to. Either a gene is overlapping or it is upstream or downstream, but not both overlapping and upstream or overlapping and downstream. Please specify what is meant here. Then, minus and plus strands are considered here and the authors should explain what it means. Is the TE sense-oriented either on the plus strand or on the minus strand to define what up- and downstream means ? Please make this clear in the legend or in the text. What is the relevance of presenting plus and minus strands ? It would be more interesting to know 1) whether a gene close to a TE has the same or opposite orientation, whether it may be on the plus or minus strand is not important to my opinion; and 2) whether the gene is upstream of the TE or downstream (the TE being considered in its sense orientation to define "up-" and "downstream"). Maybe this is what the authors intended to show but it has to be clarified.

Legend Figure 2:
"Figure 2: Number of TEs with a downstream overlapping gene and upstream overlapping gene."
This is not clear: Either a gene is up- or downstream or overlapping.

Legend Figure 3:
"... the Prunus specie Mandshurica"
Please replace "specie" with "species".

Comments concerning the manual of the pipeline
(https://github.com/marieBvr/TEs genes relationship pipeline)

It is written:
"... Long Terminal Repeat (LTR) that are type of TE."
This is incorrect. Long Terminal Repeats or LTRs are the identical sequences at 5'- and 3'-ends of "LTR retrotransposons" which frame the internal sequences containing the ORFs.

It is not clear in the manual whether the pipeline can only be used for the Apricot genome and LTR retrotransposons or also for other species. The authors should clarify this.