Pipeline to detect the relationship between transposable elements and host's coding genome

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

10.1101/2021.02.25.432867
Submitted by Marie Lefebvre 2021-03-03 15:08

Abstract

Understanding the relationship between transposable elements (TEs) and their associated genes in the host genome is a key point to explore their potential role in genome evolution. Transposable elements can regulate and affect gene expression not only because of their mobility within the genome but also because of its transcriptional activity. Gene expression can be suppressed, decreased or increased and cellular signalling pathways can be activated through the act of the nearby TE expression itself or subsequent TE replication intermediates. We implemented a pipeline, which is capable to reveal the relationship between TEs and adjacent gene distribution in the host genome. Our tool is freely available here: https://github.com/marieBvr/TEsgenesrelationship_pipeline

Keywords: Transposable element, Gene, Genome, Bioinformatics, Pipeline

Round #1

by Emmanuelle Lerat, 2021-04-05 12:28
Manuscript: **10.1101/2021.02.25.432867**

Revision needed

I have received the comments of two reviewers for your manuscript. As you will see, they both consider your work interesting. However one reviewer points out that already some known tools exist that could perform similar analyses. I would thus recommend you to perform comparative analyses with other similar tools to evaluate the added value of your pipeline. Similarly, reviewer 2 points out the fact that you should make it clear that your pipeline may be used with other genomes.
Sincerely,
E. Lerat

The points raised by the two reviewers have been adressed as follows.
All typos and grammar mistakes have been addressed and can be seen with trackchanges.

Reviews

Reviewed by anonymous reviewer, 2021-04-01 18:53

Download the review (PDF file)

Reviewed by anonymous reviewer, 2021-03-23 09:37

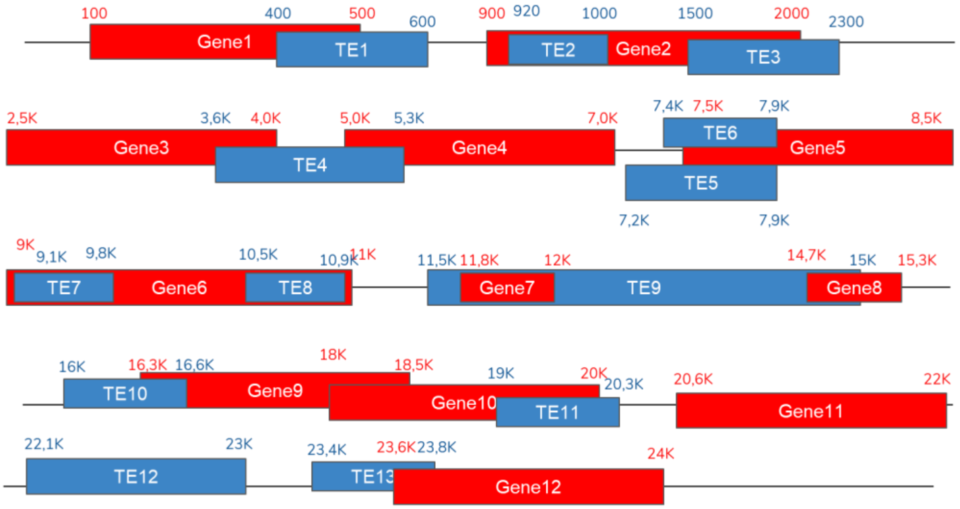
Title : Pipeline to detect the relationship between transposable elements and adjacent genes in host genome

REVIEWER 1

In this work, C. Meguerditchian and colleagues propose a pipeline to retrieve, from a list of transposable elements (TEs) coordinates and a list of gene annotations, a list of overlapping and closest upstream / downstream genes. It is true that this step is probably the first one to apply when searching for TEs with a potential impact on gene expression. However, different tools are already available to manage such analysis, and I would only cite the Bedtools suite (https://bedtools.readthedocs.io/en/latest/), with the tools "intersect" or "closest".

The bedtools intersect is used to report an overlap between two or more sets of genomic features so long as there is at least one base pair overlapping. However, it will not be very useful if there is no overlapping between the sets.

The bedtools closest can report the distance in base pairs between the intervals of set A and the closest intervals of set B. However, the default parameters do not report all scenarios that could occur between TEs and the closest genes as depicted in the figure below (Figure 1 in the paper).



Although the bedtools closest proposes different options to deal with the transcription's orientation, these should be used with caution. In the case where there are different TEs overlapping at the same location, the bedtools closest cannot report all the schemes. The proposed pipeline is simple to use and provide a complete and straightforward tool without being obliged to execute with several options. It also offers a graphical representation under R.

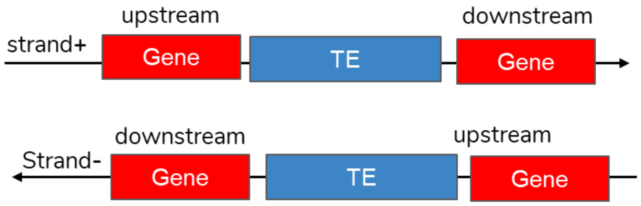
In addition, and contrary to the postulate of the authors that "none existing tools can reveal the relationship between TEs and host coding sequences", several works attempted to address this question and even went further by taking into account expression data or functional data such as GO terms (for example, LIONS (Babaian et al. 2019), or GREAM (Chandrashekar et al. 2015)).

Thus I do not think that this tool, in spite of the fact that it seems to function, constitutes neither a novelty nor a useful adding to the existing programs. I still list some propositions to improve the manuscript.

LIONS is a well-developed tool to address the association between TEs and genes at transcriptome level and not at genomic level. This tool is suitable for detecting and quantifying transposable element initiated transcription from RNA-seq.
Our pipeline is to report the structural organization between transposable elements and neighboring genes at the genomic scale in the host genome.
GREAM, 'Genomic Repeat Element Analyzer for Mammals', is a web-server designed for screening and selecting potentially important genomic repeats, however, it was developed for Mammals and proposed only 17 mammalian species. It is not currently adapted to plant datasets or customer datasets.
We have modified the Abstract and the Introduction to clarify this point.

Major comment
- The authors retrieve upstream and downstream genes of TEs, but how do they deal with non-coding TEs, for instance MITEs ? Does it have a sense to distinguish downstream and upstream genes in such case?

The genes located upstream or downstream of TEs reported in R scripts were identified based on the TEs' strand for all types of elements, including non-autonomous elements, as described in the figure below. However, in the python output file, this information was based on the physical structure and the TEs' orientation was reported, as well as the four calculations of coordinates: <Start TE - End Gene>; <Start Gene - End TE>; <End Gene - End TE> and <Start Gene - Start TE>. This will help users to define by themselves the relationship between genes and TEs such as upstream or downstream for each type of elements of interest, depending on autonomous or non-autonomous elements. This can also allow users to generate the graphics by themselves to adapt to their biological questions.



Minor comments

- In the title, the kind of "relationship" between TEs and genes should be precised. Similarly, the authors should precise what they mean by "TEs associated genes" in the abstract, as well as in the sentence "We implemented a pipeline which is capable to reveal the relationship between TEs and adjacent gene distribution in the host genome".
- Introduction: Please provide references for human and maize genome TE coverages.

- The sentence “Due to their role in transposition […], TEs can regulate [...]” should be rephrased → “Due to their transposition...”
- “will help determine the important role of TEs” → “will help determine the role of TEs”

- How do the authors define “upstream” and “downstream” parts of TEs ?
Please see Figure 3 and detailed information explained above.

Typos and grammar
- abstract: because of its transcriptional activity → because of their transcriptional activity
- Implementation:
in an downstream location → in a downstream location
this function returns gene with → this function returns genes with
this function searches for gene, which is… → this function searches for genes, which are…
what type of TE are present → what types of TEs are present
the number of TE → the number of TEs
- Conclusion:
running on two different TE annotation software → running on two different TE annotations
- Fig 3: specie → species
- Ref 1: M. Barbara → B. McClintock

REVIEWER 2
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:
<u>2 - Introduction:</u>
"... 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.

The sentence has been reformulated.

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.
The link to the raw data has been added to the sentence as well as to the section Software and Data availability.

" ... 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 ?
The Figure 1 has been added to clarify this point and examples are given in the Implementation section.

"... 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.
The Figure 1 has been added to describe all scenarios of TE-gene relationships.

3.2 Implementation

All following typos and sentences rephrasing have been addressed.
"... 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.
The processing step is composed of three functions that are executed independently. The Figure 2 (Figure 1 in the previous version), describing the workflow, has been modified to clarify this point.

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.

The Figure 1 has been added to clarify this point and examples are given for each category in Implementation section (upstream, upstream-overlap, downstream and downstream-overlap).

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.

The github documentation has been modified to address this point (see commit c630555).