Paris, December 18th, 2020

Dear PCI genomics recommender,

Please find below our answers to the two reviewer's questions. We would like to thank them for their careful reading of the manuscript and their detailed comments. We tried to reply as best as possible and we think that it improved the manuscript.

Best regards,

Hadi Quesneville

Reviews

Reviewed by anonymous reviewer, 2020-05-04 07:11

The authors present a new bioinformatic tool to detect TE remnants inside genomic sequences and have tested it on the genome of Arabidopsis thaliana, where they detected 10% more genome corresponding to degenerated TE sequences. In total, the authors propose that the proportion of A. thaliana genome corresponding to TEs is up to 50% and that the new detected sequences would correspond to co-opted TEs to provide functional role as transcription factor binding sites.

The manuscript is very interesting and proposes a new method that could help identify very ancient TE remnant potentially with a functional role for the host. I however have some major comments, starting with the evaluation of the method itself.

Major points :

Q1: their previous tool based on species comparison already increases the detection limit of TEs but the authors considered that it has limitations. It is probably more like a philosophical question but when do we know when to stop?

R1: Genomes are made by DNA from TEs in probably a large proportion. Today, it is difficult to say as we face two limits. First the 4-letter DNA alphabet which scramble the homology signal on long evolutionary time. Second the homology search which is a heuristic relying on kmers and probabilistic model. We worked on this second point to improve the detection. There is probably still room for improvement, so the limit is not reached.

Q2: in introduction, the authors indicate that epigenetic status, nucleotide composition and long term conservation in orthologous position attest of the TE origin of the identified sequences. That means that the authors consider these hallmarks specific to TEs. This is not very easy to comprehend, especially the conservation at orthologous position, since TEs are supposed to move. The authors should elaborate on this point.

R2: The conservation at orthologous position, or overlapping CNS, are evidences showing that these matches are not spurious hits and probably identify functional sequences that are maintained. This is indeed not a TE hallmark. We modify the sentence in the last paragraph of the introduction accordingly (bottom page 2).

Q3: the entire results depend on a new developed tool. But the part corresponding to its description is rather scarce in the manuscript and almost entirely put in supp mat for the strict

algorithm. Since it is particularly important to really assess the results, the program would deserve more description (with a figure explaining the work flow for example), as well as results concerning its testing. This is particularly important since it detects sequences that are beyond recognition using the standards to describe TEs.

R3: We moved the content of the supplementary file into the result section to provide the required details on the algorithm (page 7 to 12). We also add a new figure, now figure 1, providing an overview of the algorithm.

Q4: Following the previous point, a prediction accuracy part is presented in the material and method section but no results are given about it. Also, it seems strange to consider sequences overlapping genes as false positives since TEs or their remnant could be inserted into intron or exon. It may be an approximation but this should be clearly explain and the limit should be stated. Globally, I thus think that the tool in itself lack benchmarking analyses to estimate the *validity of the results.*

R4: The computed accuracy (ACC) was calculated to find the best parameter values for Duster, but not used to compare the performance with other software. We choose to remove it from the Table 1 to simplify the reading.

The referee is right. As mentioned in the material and method section, there is no perfect method to estimate false positive. Gene overlap is the best approximation we have in this situation. The limits of the approximation are indeed that sometimes TEs are part of some genes, but hopefully this remains rare, in particular when we consider coding exons. We considered CDS in our analysis to avoid TEs in introns or in 5' or 3' UTR where they could be frequent. We improved this section by describing this limitation (page 4).

Q5: Parameter -S is not explained nor in the text, or in the supplementary data.

R5: Yes, indeed, it missed. We explained and named all parameters used in the section describing the Duster algorithm (page 7).

Q6: The authors do not describe what types of small-RNAs were used. Indeed, these molecules may correspond to different types that do not target the same things.

R6: As described in dataset GSM277608 from the Gene Expression Omnibus, sequenced sRNA where obtained from bulk RNA which size comprise between 15 and 35 nucleotides. All known kind of small RNA where consequently used for the analysis. We added this information in the material and method section (page 5) as well as in the "Epigenetic profiles" results (page 14).

Q7: Concerning the histone marks, it would be nice to differentiate on the figure between activating and repressing marks for more clarity. H3K27me3 is repressive (contrary to what is written in the manuscript) but usually associated with silenced genes, which is why it is labeled as euchromatic rather than heterochromatic. H3K18ac is usually associated with enhancer, which could be the role of very ancient TE insertions and could explain the observed association.

R7: Yes, we modified the text accordingly at the end of "Epigenetic profiles" section (page 14). We added details on active or repressive mark on figure 3.

Q8: Concerning the histone profiles for TAIR10 and TAIR10-specific, I would have expected the reverse profile between the two datasets. Unless I misunderstood, TAIR10-specific are TEs that are found only in A. thaliana whereas TAIR10 include all TEs classically annotated in this genome. Then the TAIR10-specific should be TEs that still could be active. This is why I would have expected to have them associated with H3K27me1.

R8: TAIR10-specific are annotated TEs neither identified by Duster with Brassicaceae copies nor by REPET with the Brassicaceae consensus. As Brassicaceae dataset also include Arabidopsis thaliana, the TAIR10 copies which escape the REPET annotation are probably Arabidopsis thaliana specific and degenerated copies. They should not be active and should not expected to be associated with H3K27me1.

Q9: the authors explored small RNAs and histone epigenetic modifications. Why not having taken a look at DNA methylation, which is an important modification in plants?

R9: As the Duster identified copies are expected to be very old but still functionally active as shown by epigenetic marks and conservation, we did not expect them to be methylated. In addition, we do not see well what could be the improvement to include C-methylation, as it correlates with some of our analyzed histone marks.

Q10: The process for the orthologous gene comparison is not very clear. On average, for a given gene, how many of these conserved insertions are found? How many genes are concerned in total (those for which shared TEs are observed?). Maybe it would be nice to see an example of alignments as a supplementary data.

R10: We rewrote this part with more numbers to answer the reviewer answers (page 16). Note that one gene may have several overlapping TEs and vice versa. We do not see the benefit to add example of alignments in supplementary to judge empirically by eyes on some selected examples the validity of our claims, as we considered that annotation procedures and the criteria of 50% of identical nucleotides are strong enough arguments.

Q11: I find it puzzling that the very high amount of Duster specific TE found only in the vicinity of genes from A. thaliana could result from horizontal transfers. These events must have been more recent than the separation with A. lyrata since it is not in this last species. How to explain that they are that much degenerated and not recognizable by other means?

R11: We show here Duster-specific copies, that is to say, copies not identified with reference sequences or consensus. Consequently, they probably correspond to TEs for which no reference sequence has been found, and/or TE family cases where consensus sequences have not been built probably because there are not enough copies in the genome. We then identified TEs that had a poor success of invasion following the horizontal transfer. We added this at the end of "TE conservation in flowering plants" section (page 17).

Reviewed by Josep Casacuberta, 2020-04-20 12:37

This manuscript describes a new tool, Duster, which allows annotating sequences from old and degenerated TEs. The authors use this tool to annotate old TE sequences in Arabidopsis thaliana increasing significantly the percentage of the genome annotated as made of TEs. Interestingly, the authors show that the newly annotated TE sequences, which constitute the older TE fraction, are more frequently found close to genes that the previously annotated, and more recent, TEs. This suggests that these sequences close to genes have been specifically retained. The authors explore their possible function in the regulation of gene expression and

show that they overlap with CNS and experimentally determined TFBS, suggesting that they could participate in gene regulation. Annotating TEs and in particular old and degenerated TEs is still a challenge and the development of a tool such as Duster will be of great interest. Also, the results suggesting the retention of old TE sequences close to genes and their potential role in gene regulation are potentially interesting. However, in my opinion, some of the results presented lack statistical support, some of the claims are speculative and some parts of the manuscript need careful revision before the manuscript can be published.

Specific comments

Q1: Page 8 (bottom). " The greater “A-T” richness of TAIR10-specific and Duster-specific copies may indicate that they have undergone a mutation over a longer period and are therefore more ancient.". This does not match with ~~e~~ the data presented in other parts of the manuscript. For example, page 9 (bottom) " the Duster and Brassicaceae TE sequences appeared to be more ancient" (similar sentences are found elsewhere in the manuscript). Please discuss these discrepancies.

R1: The Brassicaceae-specific curve was missing on figure 3A, we corrected this. We mean in this sentence that it is older than TAIR10 TE copies. We add this to the sentence to clarify (bottom of page 13). We remove the text “As the Duster and Brassicaceae TE sequences appeared to be more ancient,” as it can be misleading and not necessary here to understand what we did (page 14, first sentence of “TE conservation in flowering plants” section).

Q2: Page 9. The significance of the overlap of the annotations with CNS is not immediately obvious. A statistical analysis to support the potential enrichment would be very helpful.

R2: Here we meant that we have an important number of TE overlaps with CNS, and not that we have an enrichment. We replaced “significant” with “substantial” to avoid a miss interpretation in the text (bottom of page 14).

Q3: Page 13. The significance of the analysis of the overlap of TEs with TFBS is also difficult to evaluate. There are many more Duster TEs (this is what it seems, although the data is not clearly given) than Brassica or TAIR10 TEs, so it is not surprising that the percentage of the annotations overlapping with TFBS is also higher. Some statistical analysis of the data will help to understand the significance of this result. Similarly, the significance of the numbers (for TFBS, genes, …) given in the rest of this section is also difficult to evaluate.

R3: The differences between Duster, Brassicaceae, and TAIR10 TEs (for example on table 3) can indeed be attributed to the differences between copies numbers in these 3 datasets. We computed the chi square tests to justify our conclusion in the text (page 18). We added occurrences in the table 4. But the remarkable point we want to emphasis is the quite high number of them overlapping TFBS, showing that TEs contribute to provide TFBS.

Q4: In the discussion it is stated that Duster copies are over represented in the 5' regions of GRN for flowering, but this is not obvious from the data in the absence of some statistical analysis.

R4: Duster copies are over represented in the 5' regions of GRN for flowering as shown in table 2. Our statistical analysis compares them through chi-square tests to what is found when all genes are taken. This could be considered as a null hypothesis, i.e., what is the expected proportion if there is no TE insertion effect. The test shows that there is an

enrichment compared to what it is expected if there are Duster-specific insertion in the upstream 500bp, but not for Brassicaceae-specific or TAIR10-specific copies (page 17).

Q5: Discussion. Sensitivity often comes with a cost in specificity. It would be useful that the discussion on the value of Duster also touches upon this aspect. Also, it is not obvious to evaluate the specificity when there is no golden standard (especially for old and previously unknown TEs). So this point should be carefully discussed.

R5: We agree with the reviewer, and we added paragraph in the discussion to elaborate on this (page 22).

Q6: The discussion on TEs regulating genes through the TFs needs revision. As it is it suggests that TEs may regulate gene transcription only because TFs regulate TE transcription. However, TEs that do not contain a promoter and whose transcription is not part of their transposition mechanism, such as MITEs, also contain TFBS and can alter the expression of genes located nearby.

R6: We agree, we added this detail in the discussion (page 23, beginning of second paragraph).

Q7: Also the references cited are all from animals and are relatively old. There are also good examples from plants that could be appropriate, taking into account that the work presented in the manuscript is on Arabidopsis.

R7: We chose to cite here only articles that first show these observation. The two most illustratives example for plants are to our knowledge DAYSLEEPER and ONSEN. They suggest wide effects on many genes in GRN.

Q8: Moreover, the discussion also mixes the domestication of cis elements from TEs (which is what the authors have analyzed here) with the domestication of a transposase into a TF (DAYSLEEPER), which is a completely different issue. Mixing the two without enough contexts can be confusing.

R8: The result of interest related to this work for DAYSLEEPER is that it is not only a TF, but also as such, may bind sequences related to hAT TE. Being domesticated, this suggest that this binding is selected for a functional role, i.e., TE sequences that regulate genes. We developed a bit more this idea (page 23, middle of second paragraph).

Minor comments

Introduction

Q9: Some of the references 1-5 are general, and do not refer to wheat and maize. They would fit better in the previous sentence. Consider eliminating most of the commas of the next paragraph, which seem misplaced.

R9: We move the references and remove the commas as suggested by the reviewer (page 1, first paragraph)

Methods

Q10: Brassicaceae TE copies. This text is a bit confusing, as the previous section describes how Brassicaceae TEs were annotated and it does not necessarily match what is explained here (is the search in each genome done with the Brassicaceae library?). Please revise these two sections.

R10: In the « Genome annotation » section, "Brassicaceae" annotation were obtained from a previous publication. It corresponds to the annotation of the Arabidopsis genome with the TE consensus built for each Brassicaceae genome. We modified the two first sentence expecting to be clearer.

In the following section, we annotated each Brassicaceae genome copies. REPET was improved as well as genome sequences for some species. We rerun this improved release of REPET on all sequences to obtained TE copy annotations for each genome. We also modified the two first sentence expecting to be clearer.

Results

Q11: Page 6 (bottom). I wonder whether "should be more similar to the ancestral sequence" would be more exact than " should conserve the ancestral sequence".

R11: We modified the text as suggested by the reviewer.

Q12: Page 7. " It shows that Duster outperforms standard tools in term of speed". Are there other tools that do exactly the same as Duster (annotating old TE copies)? What are the authors precisely comparing here? Please clarify (check also the rest of the paragraph).

R12: We modified the text adding the name of the software we are comparing (page 11).

Q13: Page 7. Brassicaceae TEs. Although it is defined in the methods section, it would be useful to have a clear definition in the results section too (probably in this paragraph).

R13: We modified the text adding how the Brassicaceae copies have been obtained (bottom of page 12).

Q14: Page 9. Epigenetic profiles. The authors discuss the results in terms of "known TEs" and "unknown TEs" whereas in other parts of the text do it in terms of older and recent TEs. It would be better to homogenize the terms used. Also known and unknown TEs seem odd.

R14: We modified the text replacing them by more appropriate terms.

Q15: Same paragraph. What are " marks copies"? "...appeared to have very few heterochromatic"... (marks, I suppose). Please, revise the whole paragraph, it is confusing.

R15: The word « that » was missing to be able to understand the sentence. We added it (middle of page 14).

Q16: Page 18. TF control TE transcription. Although transcription is the first step of transposition of most TEs, a TE copy (e.g. a defective DNA TE) can transpose without transcription, Therefore "TF control the transpositional activity of TEs" could be misleading.

R16: Even for defective TEs, their mobility is mediated by a transposase resulting from an autonomous TE transcription. We think that the sentence is not misleading here. We also modified it according to the other reviewer comment which precises our idea.