Peer Community In Genomics

Using small fragments to discover old TE remnants: the Duster approach empowers the TE detection

Francois Sabot based on reviews by Josep Casacuberta and 1 anonymous reviewer

A recommendation of:

Agnes Baud, Mariene Wan, Danielle Nouaud, Nicolas Francillonne, Dominique Anxolabehere, Hadi Quesneville. **Traces of transposable element in genome dark matter co-opted by flowering gene regulation networks** (2021), bioRxiv, 547877, ver. 6 peer-reviewed and recommended by Peer Community in Genomics.
10.1101/547877

Submitted: 07 April 2020, Recommended: 26 January 2021

Cite this recommendation as:
Francois Sabot (2021) Using small fragments to discover old TE remnants: the Duster approach empowers the TE detection. *Peer Community in Genomics, 100004.* 10.24072/pci.genomics.100004

Open Access

Published:
18 February 2020

Transposable elements are the raw material of the dark matter of the genome, the foundation of the next generation of genes and regulation networks". This sentence could be the essence of the paper of Baud et al. (2021). Transposable elements (TEs) are endogenous mobile genetic elements found in almost all genomes, which were discovered in 1948 by Barbara McClintock (awarded in 1983 the only unshared Medicine Nobel Prize so far). TEs are present everywhere, from a single isolated copy for some elements to more than millions for others, such as Alu. They are founders of major gene lineages (HET-A, TART and telomerases, RAG1/RAG2 proteins from mammals immune system; Diwash et al, 2017), and even of retroviruses (Xiong & Eickbush, 1988). However, most TEs appear as selfish elements that replicate, land in a new genomic region, then start to decay and finally disappear in the midst of the genome, turning into genomic 'dark matter' (Vitte et al, 2007). The mutations (single point, deletion, recombination, and so on) that occur during this slow death erase some of their most notable features and signature sequences, rendering them completely unrecognizable after a few million years. Numerous TE detection tools have tried to optimize their detection (Goerner-Potvin & Bourque, 2018), but further improvement is definitely challenging. This is what Baud et al. (2021) accomplished in their paper. They used a simple, elegant and efficient k-mer based approach to find small signatures that, when accumulated, allow identifying very old TEs. Using this method, called Duster, they improved the amount of annotated TEs in the model plant Arabidopsis thaliana by 20%, pushing the part of this genome occupied by TEs up from 40 to almost 50%. They further observed that these very old

Duster-specific TEs (i.e., TEs that are only detected by Duster) are, among other properties, close to genes (much more than recent TEs), not targeted by small RNA pathways, and highly associated with conserved regions across the rosid family. In addition, they are highly associated with flowering or stress response genes, and may be involved through exaptation in the evolution of responses to environmental changes. TEs are not just selfish elements: more and more studies have shown their key role in the evolution of their hosts, and tools such as Duster will help us better understand their impact.

References

Baud, A., Wan, M., Nouaud, D., Francillonne, N., Anxolabéhère, D. and Quesneville, H. (2021). Traces of transposable elements in genome dark matter co-opted by flowering gene regulation networks. bioRxiv, 547877, ver. 5 peer-reviewed and recommended by PCI Genomics.doi: https://doi.org/10.1101/547877

Bourque, G., Burns, K.H., Gehring, M. et al. (2018) Ten things you should know about transposable elements. Genome Biology 19:199. doi: https://doi.org/10.1186/s13059-018-1577-z

Goerner-Potvin, P., Bourque, G. Computational tools to unmask transposable elements. Nature Reviews Genetics 19:688–704 (2018) https://doi.org/10.1038/s41576-018-0050-x

Jangam, D., Feschotte, C. and Betrán, E. (2017) Transposable element domestication as an adaptation to evolutionary conflicts. Trends in Genetics 33:817-831. doi: https://doi.org/10.1016/j.tig.2017.07.011

Vitte, C., Panaud, O. and Quesneville, H. (2007) LTR retrotransposons in rice (Oryza sativa, L.): recent burst amplifications followed by rapid DNA loss. BMC Genomics 8:218. doi: https://doi.org/10.1186/1471-2164-8-218

Xiong, Y. and Eickbush, T. H. (1988) Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Molecular Biology and Evolution 5: 675–690. doi: https://doi.org/10.1093/oxfordjournals.molbev.a040521

Reviewed by anonymous reviewer, 2021-01-26 09:18

The authors have substantially modified their manuscript. The modifications and answers to the reviewer comment are satisfactory.

Reviewed by Josep Casacuberta, 2021-01-11 11:26

The authors have adequately addressed my comments.

Revision round #1

2020-06-18

The manuscript has now been seen by two reviewers. While their feedback is encouraging, the manuscript still requiring some corrections before I can recommend it. In particular, some statistical tests and flaws in the description of the controls have to be answered before the full acceptance of the manuscript

Additional requirements of the managing board:

As indicated in the 'How does it work?' section and in the code of conduct, please make sure that: - data are available to readers, either in the text or through an open data repository such as Zenodo (free), Dryad or some other institutional repository. Data must be reusable, thus metadata or accompanying text must carefully describe the data; - details on quantitative analyses (e.g., data treatment and statistical scripts in R, bioinformatic pipeline scripts, etc.) and details concerning simulations (scripts, codes) are available to readers in the text, as appendices, or through an open data repository, such as Zenodo, Dryad or some other institutional repository. The scripts or codes must be carefully described so that they can be reused; - details on experimental procedures are available to readers in the text or as appendices; - authors have no financial conflict of interest relating to the article. The article must contain a "Conflict of interest disclosure" paragraph before the reference section containing this sentence: "The authors of this preprint declare that they have no financial conflict of interest with the content of this article." If appropriate, this disclosure may be completed by a sentence indicating that some of the authors are PCI recommenders: "XXX is one of the PCI XXX recommenders."

All the best,

Preprint DOI: https://doi.org/10.1101/547877

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?

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.

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 workflow 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.

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.

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

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.

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.

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.

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?

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.

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?

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

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

withe 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.

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.

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

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.

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. 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. 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 context can be confusing.

Minor comments

Introduction

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.

Methods

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.

Results

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

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).

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).

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

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.

Author's reply:

Download author's reply (PDF file)