Round #1

Author's Reply:

All the authors would like to personally thank the recommender and the three reviewers for their supportive comments and constructive suggestions which allowed a significant improvement of the clarity and significance of the manuscript.

As detailed in the point by point response below, we have carefully addressed each comment and clarified all the points mentioned.

We hope this new and improved version of the manuscript will reach the standards required for recommendation by PCI Evolutionary Biology.

---

by Ines Alvarez, 2020-06-02 17:27

Manuscript: https://www.biorxiv.org/content/10.1101/2020.04.30.069948v2

This preprint merits a revision

This is a valuable piece of work, very relevant in knot-root nematode M. incognita genome evolution. It is interesting for the community and I think it could be recommendable after addressing reviewers main concerns, other minor questions, and incorporating their suggestions. Three reviewers coincided in the good quality and relevance of the study and therefore, it has a positive feedback, but I also agree that there are several points that should be clarified before its recommendation. Major concerns are related with the recapitulation of the actual knowledge on TEs in nematodes in order to highlight the relevance of the present study, as well as other study cases in which TEs activity directly affect regulatory and coding regions. Other major concern is that TEs activity drives adaptive evolution seems not conclusive here. This is not demonstrated with the data presented here and it should be noticed. There are several paragraphs dificult to follow and understand (see reviewers anotations). Please, respond to each question of all reviewers and make changes in the text accordingly in order to produce a recomendable new version of your manuscript.

We thank the recommender for her encouraging and supportive comments and we totally agree with the main concerns that were raised. We changed the manuscript accordingly and indeed now included a better overview of the actual knowledge on the importance of TE in nematode genomes and examples of functional impact of TE activity at the regulatory or coding level. We also agree that while our current data strongly support an importance of TE activity in the genome plasticity in *M. incognita*

(with possible functional impact), there is so far no clear evidence for an adaptive role. We thus changed the title, introduction and discussion accordingly.

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.

All the supporting data that did not fit in the supplementary material have been deposited in the institutional INRAE dataverse, publicly available at: https://data.inrae.fr/dataverse/TE-mobility-in-MiV3 each dataset has been cited and the associated doi is available in the reference list. Complete metadata accompany and describe each dataset.

-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.

All these details are publicly available in the INRAE dataverse (cf. comments on the previous point).

-Details on experimental procedures are available to readers in the text or as appendices.

All the experimental procedures have been described in detail in the manuscript, the supplement and the accompanying dataverse.

-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.”

We added a conflict of interest disclosure section.

Reviews

Reviewed by anonymous reviewer, 2020-05-31 05:34

In the current manuscript, the authors have surveyed the variation in the presence and frequencies of various transposable elements (TEs) in population isolates of the root-knot nematode Meloidogyne incognita. The authors have performed a comprehensive and careful analysis and reported the results quite clearly. The manuscript may be published as it is. Some minor comments and suggestions are provided below if the authors wish to include them in their analysis or discussions.

Strengths of the manuscript:

1. The root-knot nematode M. incognita is a major pest of agricultural plants. The flexibility of this pest to adapt to various plant host across wide geographical areas, despite being a clonal allopolyploid species without sexual recombination is particularly intriguing. Studies of its genome evolution might provide insights into its biology and potential ways of combating it.
2. Population genetic analysis of sequence evolution is a powerful approach to find genomic regions associated with a given phenotype, with many methods focusing on SNPs. However, the variation in TEs is hard to study from a technical as well as a theoretical perspective. Therefore studies into documenting TE variations are welcome as they may spur the development of new, appropriate methods.
3. The authors have used state-of-the-art computational methods for analyzing their data and used stringent quality filters. The methods are described in sufficient detail and most likely the scripts will be provided as supplementary materials.
4. The authors have experimentally validated some of the TE insertions predicted from their computational analysis.

Open questions / Suggestions:

1. Although authors include “adaptability” in their title, and one of their hypotheses is that TE activity might generate some adaptive variation in M. incognita genomes, very few examples of direct effects on protein-coding genes were observed. The authors have discussed the multiple reasons that could explain this (stringent filters, un-annotated genome, other adaptive effects e.g. recombination). (i) If the authors could discuss results from similar studies in other organisms with respect to the number of protein-coding and regulatory changes caused by TE, it will be a valuable information.

We agree that there was probably too much emphasis in the manuscript on the possible adaptive effects of TE activity in *M. incognita*. In the absence of clear evidence so far, we decided to attenuate these aspects in the manuscript. For instance we removed the term adaptability in the title. We also included in the introduction a series of references illustrating the known effects of TE insertions either in gene expression or gene function, including with clear connections to adaptability.

2. Since TEs can transfer genetic material via horizontal gene transfer, the authors might want to discuss this aspect as a potential contributor to adaptive functions of TEs. This could ve particularly interesting as the authors do observe some examples of TE insertions in Meloidogyne-specific genes.

Could these species- or genus-specific genes arise from HGT via TE insertions?

This is a very interesting point. Indeed, HGT has contributed to plant parasitism in nematodes (Danchin et al. 2010; Haegeman et al. 2011) and is responsible for the presence of some genes that are otherwise absent from the rest of the nematodes. In parallel, TEs are known to jump across species boundaries, including in animals (Peccoud et al. 2017). It is thus tempting to hypothesize that species-specific or genus-specific genes impacted by TEs in *M. incognita* might have been acquired by HGT with a possible contribution of TEs in these acquisitions. To investigate this, we ran an Alienness (Rancurel et al. 2017) analysis on these protein-coding genes (table pasted below). Alienness rapidly detects candidate HGT by comparing the best Metazoan and best non-Metazoan BLASTp hits against the NCBI nr and computes an Alien Index (AI). If the AI is >0, then the non-Metazoan e-value is lower (better) than the Metazoan e-value and this indicates a possible HGT. Conversely, if the AI is <0, then there is a better hit to Metazan than to non-Metazoa and no evidence for HGT. Alienness was able to retrieve all the previously reported cases of HGT in Meloidogyne with an AI score >9 (Rancurel et al. 2017). The new version of Alienness that we used also computes an HGT index, which is based on the same principle than AI, but based on BLAST scores rather than e-values (Boschetti et al. 2012). When no significant hit at all is found against the NCBI's nr library (here with an e-value threshold of 0.01), no AI or HGT index can be calculated.

Gene	**Wormbase Status**	**PCR assay**	**AI**	**HGT-i**
Minc3s00026g01668	**Meloidogyne genus-specific**	**Y**	**None**	**None**
Minc3s00450g12515	**Meloidogyne genus-specific**	**Y**	**None**	**None**
Minc3s00751g16867	**Meloidogyne genus-specific**	**Y**	**None**	**None**
Minc3s00905g18731	**Meloidogyne genus-specific**	**N**	**None**	**None**
Minc3s01127g20975	**Meloidogyne genus-specific**	**N**	**None**	**None**
Minc3s01138g21099	**Meloidogyne genus-specific**	**Y**	**None**	**None**
Minc3s01455g23950	**M. incognita-specific**	**N**	**None**	**None**
Minc3s00909g18773	**Meloidogyne genus-specific**	**Y**	**-9.62**	**-13.9**
Minc3s01827g26567	**Meloidogyne genus-specific**	**N**	**-4.68**	**-49.7**
Minc3s00988g19605	**Meloidogyne genus-specific**	**N**	**-23.76**	**-34.3**
Minc3s00621g15225	**Meloidogyne genus-specific**	**N**	**-59.53**	**-129**
Minc3s00201g07427	Nematode-specific	N	None	None

Minc3s00201g07426	Widely conserved tRNA (non-coding)	N	None	None
Minc3s01318g22714	Tylenchomorpha nematodes-specific	N	-4.7	-50.4
Minc3s00667g15847	Nematode-specific	N	-22.69	-77.8
Minc3s00965g19365	In many nematodes and animals	N	0	-444.1
Minc3s00005g00347	In many nematodes and animals	N	-3.99	-5.8
Minc3s00157g06330	In many nematodes and animals	N	-35.25	-96.3
Minc3s02496g30324	In many nematodes and animals	N	-36.84	-53.2
Minc3s00201g07425	In many nematodes and animals	N	-36.86	-53.2
Minc3s00137g05752	In many nematodes and animals	N	-57.4	-82.8
Minc3s00005g00348	In many nematodes and animals	N	-67.81	-97.8
Minc3s03567g34213	In many animals	N	-112.04	-206.1
Minc3s00199g07364	In many nematodes and animals	N	-114.87	-211.1
Minc3s00199g07365	In many nematodes	N	-164.78	-237.7
Minc3s00905g18730	In many nematodes and animals	N	-176.38	-299.7
Minc3s00301g09724	In many nematodes and animals	N	-268.08	-386.7

As shown in the table above, none of the 12 Meloidogyne-specific genes returned a positive AI or HGT index. Hence, no evidence for a possible acquisition via HGT of non-metazoan origin could be found. Actually of these 12 genes, 8 return no significant similarity at all against the NCBI's nr. Because the genes are conserved in multiple Meloidogyne species and supported by transcriptome data, it is unlikely that these genes originate from overpredictions of gene calling software. The alternative hypothesis is that they might represent Meloidogyne-specific *de novo* gene births. In that perspective, it is interesting to note that TE activity is suspected to be involved in the emergence of species or genus-specific 'orphan' genes (Ruiz-Orera et al. 2015; Wu and Knudson 2018; Jin et al. 2019). The 4 other Meloidogyne-specific impacted genes have slightly negative AI values (ranging from -4.68 to -59.53), the minimal possible negative AI value being -460.5 (Rancurel et al. 2017). Thus, these four genes return better hits to metazoan species than to non-metazoans. However, significant BLAST hits do not automatically imply orthology, and Wormbase orthology is based on Ensembl whole proteomes (not nr) and on automatically generated phylogenetic trees using Ensembl COMPARA. This explains why these 4

genes are considered Meloidogyne genus-specific in Wormbase but do return significant BLAST hits in nr.

We have clarified in the results that the Meloidogyne-specific genes have no further similarity in the NCBI's nr and thus show no evidence for acquisition via HGT and evoked the possible role of TEs in *de novo* gene birth in the discussion section.

We also analyzed with Alienness the rest of the genes impacted by TE (those not specific to the Meloidogyne genus, according to Wormbase). As could be expected, 14 of these 15 genes returned higher similarity to metazoan than to non-metazoan hits, and returned negative AI and HGT index. Only one case (Minc3s00201g07427), returned no hit at all. This gene is nematode-specific, according to Wormbase, and because, apart from *C. elegans* and a few other species, most of the nematode proteins sets derived from genomes in Wormbase and Ensembl are not present in the NCBI's nr, this can easily explain this difference.

3. Some TEs, most famously the P-elements in Drosophila melanogaster and Dropshila simulans have been observed to arise and spread in wild populations incredibly fast (e.g. between 1950s and 1990s). The authors might want to consider a more recent spread of TEs in M. incognita lineage as a potential reason why not many adaptive examples are observed. It could also be informative to analyze if fast invasions of TEs can be diagnosed by some genomic signatures e.g. patterns of TE diversity and genomic hot-spots.

Indeed, a very recent spread of some TEs might be an additional reason for the current lack of clear evidence for adaptive examples. We thank the reviewer for mentioning this point and referred to this possibility in the discussion.

In *M. incognita* the pattern of sequence identity of TE to their consensus suggests a recent burst for some orders. However, in the absence of a molecular clock in the genus Meloidogyne or in close relatives in the clade tylenchomorpha, we did refrain ourselves from dating the events. Although a molecular clock has been estimated in *C. elegans*, the reproductive modes, genome size, effective population size and life cycle duration are totally different and the Caenorhabditis and Meloidogyne genus diverged around 200 My ago.

Checking for genomic hotspots would be very interesting but, unfortunately, the currently available genome is probably too fragmented (N50= 38.6kb) to conduct a proper analysis of TE density along the scaffolds. It would certainly be more sensible to study TE density on future long-read based versions of the genome, hopefully approaching chromosome-scale resolution.

4. Typical lengths of various TE loci in each order : This would be a piece of useful information that can be included in a small table (main text or supplementary)

We thank the reviewer for this suggestion. We produced a boxplot figure available in supplementary data (new sup. Fig 2) summarising the per-order distribution of TE loci lengths.

5. The authors have used %identity of TE loci with corresponding consensus sequences as a key metric. Is it possible to also provide multiple sequence alignments of at least some representative loci within each order, demonstrating various patterns of variation?

We agree that an information about the patterns of variations along TE sequences and comparison across the orders was lacking and would be a useful addition. Such information has also been requested by Reviewer #2. We represented these patterns of variation in a single comprehensive new supplementary Figure 3. In this new figure, for each copy, the percentage of identity to its consensus as a function of the proportion of consensus covered is represented.
Note: Since the supplementary material and files were not available with the manuscript, they could not be reviewed.
All the supplementary figures and tables are now available in as an accompanying single PDF file in bioRxiv and all the supporting data has now been deposited and made publicly available in the INRAE institutional dataverse at the following URL: https://data.inra.fr/dataverse/TE-mobility-in-MiV3 with each dataset being appropriately cited in the text, the corresponding doi being available in the References.

Reviewed by anonymous reviewer, 2020-05-28 14:53

Review for PCI EVOL BIOL of the manuscript entitled “Transposable Elements activity and roles in Meloidogyne incognita genome dynamics and adaptability”.
General comments: In this manuscript, the authors retrace the dynamics of transposable elements (TE) in the genome of the root-knot nematode Meloidogyne incognita, a plant pest that reproduces asexually via mitotic parthenogenesis. The authors re-annotated the latest version of the M. incognita genome for TE and took advantage of population genomics data of a dozen of geographical isolates to study TE polymorphisms as a reporter of TE activity. They showed that TE in M. incognita are mostly DNA transposons and that thousands of TE present very highly contrasted frequencies among isolates, suggesting their transposition activity. Very interestingly, few dozen correspond to neo-insertions (some being experimentally validated) that could possibly impact protein coding genes or their regulation. These results provide evidence that TE could play a significant role in the genome plasticity

and adaptive evolution of Meloidogyne incognita and will be of interest for the community. There are however some points that should be addressed before.

Major points :

- In the introduction, the authors should recapitulate the actual knowledge about TE in nematodes and in C. elegans, in particular. This would further highlight the importance of the present study.

This is an excellent suggestion. We have now recapitulated what is known about TE in nematodes and particularly in *C. elegans*. Actually, outside of *C. elegans* there is no study of the impact of TE in the genome plasticity at the population level in nematodes. This paragraph has been added in the introduction to provide a broader context for the reader.

- lines 181-192. This paragraph is hard to understand and more quantitative details should be given to help the reader. The authors should better explain why they consider that HELITRON and MAVERICK elements do not share a high identity level with their consensus and they should better substantiate their argument that SINE and CLASS2LIKE distributions are similar. Statistical comparisons would clear the picture.

We agree that this section was unclear and totally rephrased this part to clarify the message. The aim of this paragraph was mostly to show that: 1) there was an overall high level of identity between TE copies and their consensuses for all orders, and 2) there were some differences between orders. Indeed too much focus was initially put on point 2, while point 1 was the main message. This has now been corrected. Because our objective was to show there was a general trend for high identity of TEs with their consensuses we did not specifically performed statistical comparisons for differences in these patterns for specific orders. Finally, according to another comment on which we totally agreed, all mentions of Class2 like elements were removed here as in the rest of the manuscript (see our response below to this specific question).

- It is not clear to me why annotations sharing the highest similarity with their consensus are found among DNA-transposon elements.

This was indeed not self-evident. The combination of a higher dispersion in identity among DNA transposons -- i.e the elongated boxes of Fig 2 --, coupled with a peak at high identity values, would put more annotations both at quite low identity values and at very high identity values. We rephrased this section to only focus on the DNA transposon and retrotransposons orders that showed the identity profiles most

shifted towards high values. A shift toward high identity values being considered a proxy for their recent activity.

- The authors should explain what they mean by “sufficient evidence” to consider class*1*like as retros and class*2*like as DNA-transposons. There is no conclusion about class*1*like and class*2*like elements at the end of the section and they are no longer mentioned in the rest of the text. Why are they important?

In the REPET pipeline, consensus sequences are assigned a class (I or II) and an order by the multi-agent classifier PASTEC. Each agent performs a different task such as detecting specific HMM profiles or motifs related to transposable elements or doing homology analysis against databases such as repbase. Each agent gives a score and PASTEC, the main program, integrates these informations to assign an order to each consensus sequence. When there is insufficient support for a given order (or contradictory features), but sufficient support for assignment to class I (retro) or II (DNA-transposon), PASTEC cannot assign an order but is able to assign an order. In these particular cases, PASTEC assigns the “class_1_like” or “class_2_like” classification to the consensus.

As noticed by the reviewer, class1_like and class2_like elements were only mentioned for a descriptive purpose in the beginning of the manuscript but are not analysed in detail afterwards. Because of the uncertainty of these elements, the small number of concerned elements, and the fact this decision would be coherent with our intention to focus on the 'canonical' elements, we decided to discard class1_like and class2_like elements from the analysis. According to that decision, we recomputed and corrected all the values presented in the study without class1_like and class2_like elements to only focus on those that were assigned to a known order.

- Is there a bias in the coverage of the consensus sequence length between retrotransposons and DNA-transposons, and between the different groups therein, that could partly explain the differences of the distributions of the per-copy identity percentages? If yes, the per-copy identity would not directly reflect a biological signal but would be determined by the sequence length.

We thank the reviewer for this suggestion and agree this is an important point that needs to be clarified. To address this point, we plotted for each canonical TE annotation the percentage of identity to its consensus as a function of the proportion of consensus covered (new sup. Fig 3). For each order, we computed Pearson's correlation coefficient (displayed on the figure). All the correlation coefficients were low with non significant associated p-values except for the MITE and TIR orders

where p-values were significant despite very low correlation coefficients (all <0.1), probably by virtue of the high number of elements. We concluded that there is no clear evidence for a link between the percentage of identity a copy shares with its consensus and the length of its alignment on the consensus. Hence, the % identity metrics is not biased by the coverage of the consensus.
We also displayed on the figure the median identity % (horizontal dashed line) and the median consensus coverage (vertical dashed line) for each order. As represented by the crossing of the dashed lines, we can see that TRIM and LINE (Retro-transposons) for instance share similar median identity/coverage profiles with TIR and MAVERICK (DNA-transposons), respectively. No evident difference in these distributions between DNA and retrotransposons can be observed. Moreover, given the diversity of consensus coverage profiles observed between orders and within transposon classes, we conclude there is no bias in the coverage of the consensus sequence length between retrotransposons and DNA-transposons.

- According to Table S2, the TIR annotations, but also the CLASS2LIKE and SINE ones share above 99% identity with their consensus. What is the rationale for highlighting only TIR annotations on the main text?

Indeed, as noticed by the reviewer, similarly to the TIRs, a major part of SINE elements share high identity with their consensuses. However, given the low number of annotations belonging to SINE orders (as illustrated in Table 1 and supp. Fig 3), we considered this tendency could be due to a ''sampling bias'. Hence, we decided not to highlight the SINE order. For the reasons explained previously, CLASS2LIKE elements were removed from the manuscript.

- Figure 4. Why is the Morelos isolate absent from the ML tree in Fig 4A? The authors say that clade 2 is identical in both trees, including branching, but clade 2 contains only two isolates unless I missed something.

The Maximum Likelihood (ML) tree displayed in Fig 4 was retrieved from a previously published study based on SNVs in coding regions where morelos isolate was initially not taken into account (Koutsovoulos et al. 2020). We agree with the reviewer that for a more comprehensive comparison between the SNV-based and TE-based analysis, the Morelos isolate should be included in the SNV-based analysis. To that end, we recomputed the ML SNVs based tree (cf. methods), including the Morelos isolate, and updated Fig 4 and the sup. Fig 5 (trees with branch length displayed). This new result strengthens our conclusion on the similarity between the two topologies and the fact that “most of the phylogenetic signal coming from variations in TE-frequencies between isolates recapitulates the

SNV-based genomic divergence between isolates". Clade 2 contains 3 isolates in both trees (R4-1, R3-1 and R3-4) and the branching order is exactly the same.

- 4 consensuses are involved in 24 out of 33 HCPTE. Is there a bias among MITEs, TIRs and LINEs?

Only 3 TE orders are represented in the 33 HCPTEs loci: TIRs (4 consensuses), MITEs (7 consensuses), and LINEs (2 consensuses). Indeed, 4 consensuses encompass 24/33 HCPTEs: 2 TIRs (8 and 2 copies per consensus) and 2 MITEs (10 and 4 copies per consensus) (cf. table below). We decided to focus exclusively on the consensuses with the two highest number of copies involved in highly heterogeneous polymorphisms: 1 TIR (8 copies) and 1 MITE (10 copies). These two consensuses alone combined, encompass more than half of the HCPTEs. We changed the text accordingly. The following table has been provided in supplementary (sup. Table S7).

consensus	order	nb. of HCPTEs copies
DTX-comp_mincV3XDN-B-R1459-Map20	TIR	8
DTX-incomp_mincV3XDN-B-R11531-Map10	TIR	2
DTX-incomp_mincV3XDN-B-R271-Map10	TIR	1
DTX-incomp_mincV3XDN-B-R3892-Map20	TIR	1
DXX-MITE_mincV3XDN-B-G1048-Map15	MITE	1
DXX-MITE_mincV3XDN-B-G305-Map9	MITE	1
DXX-MITE_mincV3XDN-B-R14125-Map7	MITE	1
DXX-MITE_mincV3XDN-B-R306-Map20	MITE	10
DXX-MITE_mincV3XDN-B-R321-Map20	MITE	1
DXX-MITE_mincV3XDN-B-R3266-Map20	MITE	1
DXX-MITE_mincV3XDN-B-R3611-Map9	MITE	4
RIX-comp_mincV3XDN-B-R6875-Map20_reversed	LINE	1
RIX-incomp_mincV3XDN-B-R4613-Map9	LINE	1

- The 5 HCPTE for experimental validation all impact Meloidogyne-specific genes, but others HCPTE also share this property according to Table S4. What is the rationale for selecting those 5?

We performed experimental validations using DNA leftover from a previous analysis (Koutsovoulos et al. 2020). Given the small amount of available material, we had to

select a limited number of loci to test. We decided to focus on highly heterogeneous TEs (HCPTEs) loci as for these loci the predicted frequencies were close to presence/absence signal in all the isolates, and so more easily testable by PCR.
We based our choice of loci to be validated on the following criteria:

- The element must be predicted to be inserted in a genic or potential regulatory region (max 1kb upstream of a gene) as the most evident criterion for a potential functional impact.
- The element must be short enough to be amplified by PCR and sequenced by SANGER using standard techniques and material (2.5kb max).
- To validate the predicted impacted gene actually exists, it must be supported by substantial expression data in the reference isolate Morelos.
- To maximize the chances of effects on biological traits characteristic of the root-knot nematodes, the impacted gene must be Meloidogyne-specific.

Once all these criteria were applied, we maximized the diversity of TE orders involved and this resulted in the selection of these 5 loci.

- It would be worth mentioning whether or not such highly contrasted polymorphism in TE has already been observed in C. elegans, and to briefly remind the reader of what (if any) has been performed to estimate the impact of TE in C. elegans.

In (Dolgin et al. 2008), the polymorphism of Tc1 and mTcre1 elements has been studied, respectively in *C. elegans* and *C. remanei*. In those species, a high polymorphism for Tc1 elements has been shown. Furthermore, in (Laricchia et al. 2017) a more comprehensive analysis on the whole set of predicted TE dynamics in *C. elegans* populations has been performed. Although the amplitude of variations in TE frequencies within populations was not studied, variations in the pattern of TE presence / absence were shown. These results have now been introduced and discussed.

- In the discussion, when referring to their results, authors should refer to the corresponding figures and Tables. This is especially true for arguments on lines 730 to 736. Based on Figure 2, it is not obvious that the behavior of MITEs and TIRs are different in function of the identity rate with their consensuses. Also, the comment that “TIR neo-insertions are less numerous than expected owing to their abundance in the genome” is supported by Fig S4 but Fig S4 is never cited in the text.

We totally agree that, at this point of the paper, these arguments are not self-evident at all. We now cited the Figures and supplements supporting these conclusions at the appropriate places in the discussion.

- Methods: Diagram recapitulating all the annotation steps would greatly help the reader. A decision tree would also help for the polymorphism characterization.

A diagram recapitulating the TE prediction and annotation, the TE frequency estimation across the isolates, and the TE polymorphism analysis has been provided in supplementary (sup. Fig S7); as well as a decision tree summarising the polymorphism characterisation (sup. Fig S8).

- The bootstrap approach for the TE-frequency NJ tree should be explained, at least briefly.

We acknowledge that because this is not applied to a multiple sequence alignment, the bootstrap procedure deserves more explanation and we have added the following details in methods. To compute the bootstrap values, we used the "boot.phylo" function from the ape-v5.4 package (Paradis and Schliep 2019). The boot.phylo function performs n resampling of the frequency matrix (here the matrix with loci in columns, isolates in row, and values corresponding to the frequencies).

- Data : The authors should also give the project accession number of the M. incognita reference genome they used from Blanc-Mathieu et al. 2017.

We added this information in the material and methods section (ENA assembly accession GCA_900182535, bioproject PRJEB8714).

Minor points:

- A definition of autonomous and non-autonomous TE would be nice and the “autonomous/non-autonomous” status for each order of transposons in the M. incognita annotations could be given in Table 1.

We followed reviewers' suggestions and added a definition of autonomous and non-autonomous TEs in Table 1 caption. We also added each orders' autonomous/non-autonomous status in the first column of the same table.

- Table 1. What is the median of median identity with consensus (%)? Tipo?

Yes, thanks for pointing this out. The concerned column’s header was modified and simplified to “median identity with consensus”.

- Figure 3A. It is very hard to see something at that small size.

We made a new Figure 3 with a white background, better resolution and bigger font size.

- Figure 5. A title (the type of TE locus described) for each panel would help. The number of loci used in each panel could also be given in the panel, not only in the legend.

We thank the reviewer for this suggestion. We modified Fig 5, adding the name of the different categories besides the codes. Also, the count of concerned loci for each panel is now displayed on the figure.

- Table S4 title: correct “ortologs”

Corrected, thanks for pointing it out.

Reviewed by Daniel Vitales, 2020-05-22 18:10

The manuscript “Transposable Elements activity and role in Meloidogyne incognita genome dynamics and adaptability” by Kozlowski et al. presents a thorough study of TE content, frequency variability and activity among several isolates of M. incognita, a parasitic root-knot nematode showing significant impact as an agricultural pest. The whole genome sequencing data used by the authors, which was already analysed in a population genomics study (Koutsovoulos et al. 2020), is here nicely re-analysed to characterize the repeatome of this species, to identify potentially active TE loci and compare their frequencies among some M.incognita isolates with different hosts and geographic origins. Additionally, they study the impact that transposition of some elements could have in some transcriptionally active genes. I would like to say I really enjoyed reading this work, which allowed me to know new methodologies and points of view in the study of repetitive elements evolution.
First of all, I must make clear that I have never used neither the general approaches nor the specific pipelines here employed by the authors to analyse the repeatome. Consequently, although the procedures seem correct to me, I cannot (and will not) evaluate the details of the methodology. In this case, my review will focus on general points related to the hypotheses tested, the results obtained and some of the interpretations carried out. I hope my comments will be helpful for the authors:

1) As stated in the Introduction section, the main goal of this study is to test “whether the TE activity could represent a mechanism supporting genome plasticity and eventually adaptive evolution in M. incognita”. The authors do provide solid evidences to prove the hypothesis that TE played an important role in the intraspecific genomic diversity of M. incognita. They also clearly show the potential impact of some TE in the activity of certain genes. However, while reading the Introduction and the Discussion, my impression was that the most important goal of

the manuscript was on testing the adaptive role of TE in the evolution of the species. The first section of the Discussion is indeed devoted to expose potential effects of TE activity in adaptive evolution or the impact on the function of certain genes in different species. Certainly, some genic regions appear to be impacted by TEs in M. incognita and they were confirmed to be expressed ¬– according to the transcriptome data they got – but I am not sure that the experimental design enables to test here the adaptive role of TE activity. As the authors comment in the first Discussion section, "functional impact itself would need to be evaluated in the future (L591-L592)" and “no evident role in adaptive evolution for the M. incognita genes impacted by TE insertions could be reported so far (L611-L613)”. In my opinion the data and the experimental design here employed works very well to characterize and compare the repeatome content among different isolates of the species, whereas they do not allow to test whether TE activity drives adaptive evolution. I think the message of the manuscript would be stronger if the authors focus the objectives (mainly reformulating the Introduction and the Discussion) on the intraspecific dynamism of TEs in M. incognita, where they could show clear results and obtain interesting conclusions.

We fully agree with the comments of the reviewer, too much emphasis was initially put on the possible adaptive consequences while the actual main message is rather on the possible activity of TE and their role in the genomic plasticity regardless of any adaptive impact. We modified the title, the introduction and the discussion accordingly, by removing most of the interpretations concerning putative adaptation processes, and focusing on the observed activity. The final part of the discussion -- on activity -- was now moved as a first section, with little changes but the addition of a small paragraph emphasizing the fact that what we detect is probably the tip of the iceberg in term of activity, because of the stringency of our filtering and the technical limitations of PopoolationTE2.
The part on adaptation was mostly rewritten, with a focus on the known case of adaptive evolution in plant parasitic nematodes and its putative link with TEs, and what is known on the free-living nematode *C. elegans*. The discussion about putative adaptation was reduced and reframed, emphasizing the fact that in similar contexts, studies on Drosophila or *C.elegans* found no evidence for a strong, large-scale, associated adaptation.
The final part is now the discussion about the conflicting roles of ploidy, hybridization and sexuality on TE load. The title was changed to better reflect the content and focus of the manuscript.

2) At first glance, I was happy to see that the second section of the Discussion was devoted to the “TE-load and composition” of M. incognita. Certainly, the analyses performed by the authors provides remarkable information on the TE content of the species, as well as its intraspecific variability. However, I just found five lines at the

end of this section (L699-L703) commenting the finding that DNA transposons are the most abundant element of M. incognita repeatome, while the rest of the section mainly presents relatively disconnected study cases where the TE content was affected (or not) by hybridization, polyploidy or asexual reproduction. Probably I was expecting that the authors compare the repeatome characterization of M. incognita isolates they obtained with the repeatome of other Meloidogyne species. For this purpose, in case they need a different (broader) perspective of the repeatome landscape of this species (and other they could study) I would suggest to analyse their genomic data with other de-novo approaches (e.g. RepeatExplorer).

As far as we know, to date there has been only one study on the TE-content (and Repeatome) at the whole Nematoda phylum scale (Szitenberg et al. 2016). This study encompassed 42 nematode genomes, including 5 Meloidogyne species and concluded that a higher abundance of DNA transposons seemed to be a general feature of nematode genomes, including in the Meloidogyne genus. The study also concluded that variations in TE abundance and composition was not linked to the life history traits under consideration (e.g. reproductive mode). However, these studies have been undertaken with methods depending on genome assembly and the quality and completeness of the genome assemblies themselves might introduce some biases in the analysis. Therefore, an assembly-free method such as the one proposed by the reviewer would be interesting and complementary. A study in bioRxiv has used dnaPipeTE (Goubert et al. 2015), another assembly-free method to estimate TE contents from genomic reads in various outcrossing and parthenogenetic animals including 5 Meloidogyne species (Jaron et al. 2020). The study also showed that DNA transposons seem to be more abundant than retro-transposon in the Meloidogyne (Figure 4 of that paper). In the same paper, no clear difference in TE loads could be evidenced between the 5 parthenogenetic Meloidogyne species, whether or not meiosis was present (supplementary Figure 8 of that paper). We added these points of comparison in the discussion.
Overall, the scope of our paper was to assess whether TE might participate in the genome dynamics and plasticity in *M. incognita* and not to extensively compare the TE abundance and composition with other Meloidogyne and different nematodes. Nevertheless, we agree that such a study would be interesting. Actually, we are currently performing the repeatome characterisation of the few Meloidogyne species with genome data available using dnaPipeTE to conduct an unbiased comparison of the genomic TE content among the Meloidogyne genus. We are also engaged in massive genome sequencing efforts to generate more genomes aiming at including both selfing, fully parthenogenetic and outcrossing Meloidogyne species as well as hybrids and non-hyrbids, polyploids and diploids. Such a study will need more time and we think will be better highlighted in a separated paper.

3) Finally, regarding the structure of the Discussion, I found that the third section entitled “TE show signs of recent activity in M. incognita and they might still be active” presents some of the most solid and remarkable results of the paper. I would recommend to start the Discussion with the results commented in this section, then going on with the section devoted to explain the impact of TE in the gene(s).

We fully agree, this better fits with the main messages of the paper and modified the Discussion sections accordingly.

Some other points:
L55. This could be a good place to explain for the first time that M. incognita is a triploid, parthenogenic species with a hybrid origin.
We have now clarified this in the introduction, as suggested.
L79. I think the abbreviation of “id est” should be (i.e.)
All the abbreviations have been checked and corrected if needed.
L99. The expression “‘novelty’ / plasticity” could be changed by “genomic novelty or plasticity”
Done. As suggested, we replaced “‘novelty’ / plasticity” by “genomic novelty or plasticity”.
L105. The “arms race” was written between quotation marks three line before.
We removed quotation marks in both.
L115-L117. To me, this statement sounds as a Result or as a Discussion.
We fully agree and rephrased this sentence to explain we have used the distribution of % identity of TE to their consensuses to assess the recentness of their activity without providing the result of this analysis there.
L137. The abbreviation “cf.“ is employed throughout the text to point the methods as a source of information. I think the word (“see XXX”) would be more appropriate.
We replaced the abbreviation “cf.” by “see” in the whole text
L146-L147. There is stated that “Retro-transposons and DNA-transposons respectively cover 0.94 and 3.78 % of the genome”. I understand these values correspond to the retrotransposons and DNA transposons within the canonical TE (not the whole repeatome, which could include “non-canonical TE). If I am right, this clarification should be specified here.
The given values indeed correspond to canonical retro and DNA transposon percentage in the genome of M. incognita. We clarified the sentence by adding the keyword "canonical".
In addition, I think it would be interesting to know whether the composition of canonical TEs is comparable to the composition of the whole repeatome. As commented above, the characterization of the whole repeatome of M. incognita using a de novo approach such as RepeatExplorer could be easy and very informative.

We agree and now provided tables summarising the whole (e.g. unfiltered) TE annotations of *M. incognita* (Table S1) and *C. elegans* (Table S2) in the supplementary for comparison matters. As explained previously, we reserve genome assembly-free de novo methods for a separate paper with extensive comparisons with multiple *Meloidogyne* genomes.

L147-L148. Being the first time the acronyms TIR and MITEs are used, I would suggest to mention them as "Terminal Inverted Repeats (TIR) and Miniature Inverted repeat Transposable Elements (MITEs)...".

Done. As suggested, we replaced “TIR (Terminal Inverted Repeats) and MITEs (Miniature Inverted repeat Transposable Elements)" by “Terminal Inverted Repeats (TIR) and Miniature Inverted repeat Transposable Elements (MITEs)”.

L151-L164. I must recognize I got difficulties to understand some of the information in this paragraph. I got that REPET pipeline estimated a repetitive content in C. elegans genome very similar to that obtained by Bessereau 2006). Were the predictions of 1.8% and 0.2% of MITEs and LTR obtained with or without the filtering protocol? Regarding Bessereau estimations, do they correspond to canonical (i.e. "potentially active" TEs) or to fossil + active TEs? To be comparable, both predicitions should consider the same type (canonical or all kind) of TEs. The authors might consider to rephrase the paragraph for a better understanding.

Indeed, the idea was to show the REPET annotation we performed did recapitulate the results obtained by (Bessereau 2006) on *C. elegans*, as a validation. The concerned paragraph has been rephrased and simplified to focus on comparison of the repeatome between *M. incognita* and *C. elegans*. Tables have been added in the supplementary material (sup. Table 1,2,&3) to detail both draft and canonical TE annotations in the *M. incognita* and *C. elegans* genomes.

L158. Supplementary materials indicate that MITEs compose 0.7% of C. elegans genomes, while here it is mentioned “1.8%”. This should be revised. I would also recommend to construct a table where one could easily compare TE estimations among Bessereau 2006 and your approach.

We thank the reviewer for noticing this misleading information. The value of 1.8% corresponds to the draft unfiltered value (sup. Table 2), while the value presented in the sup. Fig 1 corresponds to the filtered value. Furthermore, the text has been modified to make clearer that we are comparing draft annotations with the literature (not the filtered one).

L174. Please cite a reference for this genome size value.

The reference asked by the reviewer has been added to the manuscript (Blanc-Mathieu et al. 2017).

L184. Should “HELITRON” and “MAVERICK” be written in capital letters?

We thank the reviewer for noticing this mistake. The proper way to write these TE orders indeed is 'Helitron' and 'Maverick' rather than 'HELITRON' and 'MAVERICK' (Wicker et al. 2007). This mistake has been corrected in the core text and also in the figures (main article and supplementary).

L183. I was not able to find A and B parts in Figure 2.
Fig 2 indeed is not composed of panels. We modified the citation to this figure in the text, accordingly.
L206-L208. I was not able to explain why only 6.26% of canonical TEs (i.e. complete elements and suposedly being potentially active) contain a protein coding gene. Is this value something expected? What those canonical elements lacking transposition genes correspond to? Could this percentage be compared to the "autonomous:non-autonomous ratio" of elements found in the repetome of M. incognita?
In the *M. incognita* genome we used in this analysis (Blanc-Mathieu et al. 2017), gene models have been predicted by the EUGENE pipeline (Sallet et al. 2019). In EUGENE, repetitive elements (including TEs) are detected by RED (Girgis 2015) and then masked in the genome, unless there is support from transcription at these genome locations from RNA-seq or EST data (in this case, the region is unmasked). Gene models are then predicted on the unmasked parts of the genome. In the present analysis, we used REPET to annotate TEs in the genome. Protein-coding genes can only be detected by EUGENE in these regions, either because they were not detected as repetitive by RED or were unmasked due to transcriptional support. Therefore, the restricted number of protein-coding genes contained into TE annotations 6.21% (598/9,633) is consistent with the protocol used for gene prediction and annotation. Moreover, this value of 6.21% (598/9,633) takes into account both autonomous (4,320) and non-autonomous (5,313) TE annotations. Discarding non-autonomous elements, supposed to lack functional transposition machinery, and focusing on the 4,320 autonomous TE this value rises to 13.84% (598/4320). Hence, making sense with the predictions all the protein-coding genes in TEs (598) were exclusively found in autonomous TEs. The relatively low percentage being only due to masking options of the EUGENE gene predictor.

L211. I was not able to find this supplementary information on the proportion of LTRs, LINES, TIRs, etc among the 111 canonical TE with transposition genes.
Per-order count of TE with putative transposition machinery and substantially expressed putative transposition machinery has been provided in a new sup. Table 5.

L240. If I understood well, this corresponds only to the 3,524 variable TE loci. Perhaps this could be better explained.
Yes, this is exactly the case, the number of loci used to compute the distance matrix (3,514 loci) has been added in the text.

L238-L250. I would suggest to use "within" for intra-isolate variability and "between" for inter-isolate variables. Or another nomenclature but using it consistently

throughout the text. In my opinion, this will make easier for the reader to understand the values/indexes you are talking about.
The term ‘between’ isolates was already consistently used when referring to comparisons between the isolates, we added the term within-isolates when any confusion could emerge to clearly indicate when frequencies within isolates were specifically considered.

L264-L265. It would be nice to state how many “loci frequencies” values were employed to construct the distance matrix.
The number of loci used to compute the distance matrix (3,514 loci) has been added in the text. Details also have been added in methods concerning the bootstrap values computation.

L276-L277. Here, I would also take into account that the branches leading to R2-1 and R1-6 do not show high support values in TE-based tree. This low resolution obtained in that parts of the TE-based trees - which could be caused by several factors (e.g. hybridisation; polyploidy) - should probably be mentioned in the results.
We now referred to the branch lengths in the results and discussed differences in branch lengths between isolated (particularly R2-1 which show the longest branch in both SNV-based and TE-based trees).

L278-L279. I was delighted to see the beautiful phylogenetic signal the authors obtained from their repeatomic data. However, I would rather expose that "TE-frequencies between isolates contain a valid phylogenetic signal". I would have also enjoyed some comments in the Discussion section about these results. For instance, the authors could consider potential biases caused by the loci selection they performed. This tree was based on the 3524 significantly variable loci, which is only a part from the 9000 canonical TE loci.
We were also delighted to note that the TE-based phylogeny recapitulated almost exactly the phylogenetic signal present in the SNV-based phylogeny. Indeed, for this phylogeny, we only used the variable loci, which do not cover all the canonical loci. Stable loci which do not vary in frequency between isolates do not bring any useful signal for classification of the isolates were eliminated. We did not extensively discuss the TE-frequencies as a valid phylogenetic signal but rather the links between the topology and life history traits. Indeed, although it worked well with our model, we think more extensive tests should be made on other species before assuming TE-frequencies is a general valid phylogenetic signal. If, for instance TE have been massively mobilized as a response to a stress in some lineages, this might not work as clearly as in our study.

L302. As this is important to understand the text below, I would suggest to include the name of the different categories beside the codes A, B, C and D of Fig. 5.

We thank the reviewer for this suggestion. We modified Fig 5, adding the name of the different categories besides the panel codes. Also, as requested by reviewer #2, the count of concerned loci for each panel has been displayed in the figure.

L331-L355. I am not sure to be properly understanding the categorization procedure for polymorphic TEs.
We clarified how we categorized the polymorphic TEs and explained in details the different criteria used in a decision tree in the supplementary material (new sup. Fig S8).
Could the “truncated or diverged versions of TE” be also present in the other isolates (excluding Morelos)? In other words, I was not able to see how authors can be sure that "neo-insertions" do not in fact correspond to "extra-detections" occurring in the rest of isolates (but not in Morelos). Perhaps this part could be better explained.
All the neo-insertions we observed as compared to the reference isolate (Morelos) are either isolate-specific or branch specific. Hence, so far, the most parsimonious hypothesis is that these TE have been inserted de novo, specifically in these branches. The alternative hypothesis of an ancestral presence but loss in multiple independent lineages, including Morelos, is possible but seems less parsimonious. Only one case of neo-insertion shared by 6 lineages belonging to a same monophyletic group could either equally represent a gain in this branch or a loss in the other branch, in terms of parsimony. Actually, only the use of outgroup lineages or a closely related species would allow resolving this. We made clear that these neo-insertions are relative to the Morelos isolate in the manuscript.

L367. Could Figure 6 include the information of non-polymorphic loci?
We believe Fig 6 was already dense and are afraid adding categorization for non-polymorphic loci would blur the message and make the Figure difficult to read and interpret. We provided a table in the supplementary material summarizing the count per orders for both polymorphic (ref-polymorphisms, neo-insertions, extra-detections) and non-polymorphic TEs (sup. Table S6).

L452-L453. The authors might consider to compare the positions of “neo-insertions” with the position of other “polymorphic TE” categories. Is the proportion of elements inside a gene or a regulatory region the same in other “polymorphic TE” categories?
We thank the reviewer for this suggestion. We investigated whether the proportion of polymorphic TE in genic and regulatory regions varied according to the category of polymorphism (i.e. ref-polymorphism, extra-detection, and neo-insertion). In the table below, we can see the proportion is not the same, depending on the category and the differences are significant (Chi-square test, p-value = 1.3e-3).

	ref-polymorphism	extra-detection	neo-insertion
total nb. of elements per type	2091	206	287
nb. of elements intersecting a gene	421 (20.13%)	56 (27.18%)	112 (39.02%)
nb. of elements inserted upstream a gene (< 1 kb)	559 (26.73%)	56 (26.73%)	86 (29.96%)

L470. I was not able to find why specifically these 5 HCPTE were studied but not the rest of them.
A similar question was asked by Reviewer #2 and the list of criteria used to select these 5 genes was explained above at this occasion.

L481. I was not able to find the Supplementary material 4 the authors mention here. Is this the table S4?
We are sorry for this confusion, as explained to the recommender, all the supporting data that did not fit in a supplementary material were deposited in the INRAE institutional dataverse and made publicly available. This has now been updated in the current manuscript and the data is publicly available in the following material (Kozlowski et al. 2020), cited in the reference list.

L482. I missed some more details about the PCR validation results on the rest of HCPTE.
As evoked above, PCR and sequencing validation results (including the original gel pictures) of all the tested HCPTEs are detailed (Kozlowski et al. 2020), cited in the reference list.

L590. According to the transcriptome data, some of these regions were certainly confirmed to be expressed, but can you really state they are functionally important? As the authors comment in the following lines, "functional impact itself would need to be evaluated in the future".
We agree that transcriptional support just informs that the predicted gene is transcribed but does not inform on the functional importance it may have. This criterion was mainly to eliminate possible over-predictions from gene-calling software. The conservation of the genes in multiple Meloidogyne species and their specificity to the Meloidogyne genus reinforce their possible importance.

L722-L723. As the authors admit in L737-747, there should be additional data and stronger evidences to confirm a TE burst during the evolution of the species.
Yes, at this stage the data just suggest a TE burst, and this would be consistent with the recent hybrid origin of this species. We made clear that this hypothesis would need further data to be tested in the future.

L751-L752. I am not sure if the authors can conclude this statement without considering other alternatives. Could these "neo-insertions" be present in the original pre-agriculture genomic pool and later on being fixed or erased from some isolates?
We agree with the reviewer that an alternative hypothesis exists concerning neo-insertion rising in the isolates. However, the results showed each time a group of isolates shared a neo-insertion, all the concerned isolates belonged to the same monophyletic cluster (Fig 4). Moreover, for three out of four cases, a maximum of three isolates out of twelve shared the neo-insertion. Hence, we consider the most parsimonious scenario is that these neo-insertions occurred in *M. incognita*, after the separation of the different main clusters but before the diversification of the phylogenetically-related isolates, within a cluster, in a common ancestor. This has been clarified in the revised manuscript.