Cell Genomics Transparent Peer Review Record

Ancestral genome reconstruction enhances transposable element annotation by identifying degenerate integrants

Wayo Matsushima, Evarist Planet, Didier Trono

Summary

Initial submission:	Received : Apr 26, 2023
	Scientific editor: Sara Rohban
First round of review:	Number of reviewers: 3 Revision invited : Aug 09, 2023 Revision received : Nov 17, 2023
Second round of review:	Number of reviewers: 3 Accepted : Jan 06, 2024
Data freely available:	YES
Code freely available:	YES

Referees' reports, first round of review

Reviewer #1: This manuscript was one of the more enjoyable papers I have read recently, and I recommend publication as a short report. The authors make use of recently published ancestral genome reconstructions to improve the annotation of TEs in the human genome. In doing so, they attribute an additional 10% of the human genome to TE sequences. By comparing these newly annotated "degTE" insertions with the binding profiles of transcription factors, they show that many cis-regulatory elements are likely derived from ancestral TEs. Overall, the paper is short but sweet, and the methods seem to be sound. I have no serious issues with technical aspects of the paper, but a few questions, and I feel that there are several obvious analyses that could improve the paper. That being said, these are not essential for the scientific validity of the paper, so I make them as general comments rather than requests for additional work.

1. My biggest concern with the method itself is that the RAGs are not annotated de-novo, but using existing DFAM libraries with RepeatMasker and the "-species human" parameter. I suspect that this will heavily bias the method towards finding sequences that are similar to existing human TEs, at the expense of TE families that may otherwise be well represented in the RAGs, but present at very low copy number in the present-day human genome. It would be very interesting to see how this approach compares to one using de-novo TE annotations of the RAGs (e.g. with RepeatModeller), and it would be more in keeping with the species-agnostic approach of the progressive cactus aligner. On the other hand, if people are strictly interested in annotating the human genome, the method described here is likely best.

2. On a related note, it would be good to see a few sentences more on the limitations of the RAGs themselves. The genomic fraction attributable to TEs in RAGs is low (Fig 2B), and Fig 2D shows a striking lack of young TE insertions in the RAGs. While it is unavoidable that many "young" TEs present in the true ancestral genomes will be absent from the RAGs if they are not inherited by at least some of the species used for reconstruction, it is surprising that there are so few young families/insertions in the RAGs. As the authors mention at lines 57-60, TE insertions in the RAGs should - in theory - be closer to the original consensus sequences. I believe Fig S2b addresses this, but bit scores are (for most people) a less intuitive measure than divergence from consensus. If degTE insertions in the RAGs are indeed closer to consensus than those in hg38, it would be interesting to hear the authors' thoughts on why young insertions seems to be depleted in Fig 2D.

3. The paragraph starting line 125 is difficult to parse. In addition, I have a pedantic note referring to the use of the term "homology" (e.g. line 125, where the authors state that they "tested how much homology degTEs retained"). Strictly speaking, since homology refers to a binary state, "sequence conservation" or "similarity" would be more appropriate here.

4. I found the section on KZNFs very interesting. I wonder if the authors have compared the age of the KZNFs targeting degTEs to those targeting more recent TE expansions such as L1PA subfamilies in human. If KZNFs targeting degTEs are significantly older than KZNFs targeting more recent TE expansions, it would lend further support to the idea that KZNFs facilitate the integration of TEs into novel gene regulatory networks.

Reviewer #2: Here, Matsushima et al present a novel method for annotating transposable elements based on ancestral genomes. When applied to the human genome, the methodology expands the number of bps annotated as TEs by 10%.The authors characterize the overlap with functional genomic elements, concluding that the newly discovered elements play an important role in the evolution of gene regulation. Finally, they also demonstrate that 250 of these elements are part of human genes active during early embryonic development.

Overall, I find the manuscript well written with few typos and it is easy to follow the authors' line of reasoning. TE annotation remains a challenging and important problem as these elements are by far the largest class in most genomes. Although the work presented here provides an interesting advance, I do not think that it could be published in Cell Genomics in its current form. The following major issues need to be resolved;

1. It is my understanding that the authors rely on a combination of RAGs and RepeatMasker for identifying TEs. Although I agree that RepeatMasker is the most widely approach for finding TEs, it is not the only method out there. It would be useful if the authors could apply other methods for TE detection (e.g. from Goerner-Potvinet al , Nat Rev Gen, 2018, Ou et al, Genome Bio, 2019) to help understand how sensitive the results are to RepeatMasker.

2. The authors mention FDR for RepeatMasker and other TE finding approaches in the discussion. I understand that the lack of ground truth regarding TEs and degTEs makes it hard to provide these estimates for the results presented here. Nevertheless, it would be helpful to try or at the very least add to the discussion on how one could in principle calculate such estimates.

3. Although the authors publish their new TE annotations, this is limited to humans. There is a large community of biologists studying other genomes who might be interested in improved TE annotation. I see no reason why the authors could not make their software available as a package so that others can run it on other genomes. I understand that it may take time to run, but as long as this is clearly stated upfront, then users can decide on their own if they want to wait for results.

4. The authors present several pieces of statistics showing overlap of newly discovered TEs with various functional elements. However, there are no null models presented to allow for statistical tests determining whether or not these overlaps are larger than expected than by chance. The authors should include suitable null models, hypothesis tests and p-values for these comparisons.

5. In the last section, the authors investigate human RNAseq data to identify genes containing parts of TEs, suggesting that they have functional roles. Although tantalizing, this analysis leaves much to desire. First, the authors need to report on the expression levels of these genes - how many are above meaningful thresholds? Second, although the example shown in Fig 5 is convincing, it would also be useful to know how many of them constitute the dominating isoform. Third, what about the function of the 250 genes? Are they enriched for certain categories or pathways. Fourth, what about the function of the new exons? The example in Fig 5 looks like it is a 5' UTR, so this probably has no impact on the protein (unless a new ORF is created). What about others? How long are they? Are they in-frame? Do they disrupt annotated protein domains? Taken together, more evidence is required to support the case that these chimeric TE transcripts are functionally important.

The following minor issues need also be addressed:

It would be helpful if the authors' could comment on the computational resources required to produce the new TE annotation as well as for some of the other analyses presented in the manuscript.

There is no legend for the color scale in Fig 4c.

Reviewer #3: Matsushima and colleagues present the first attempt to identify additional reference transposable element (TE) insertions using multiple reconstructed ancestral genomes (RAGs). They identified >10% more reference TEs (called degTEs) that are missing from the current hg38 reference genome due to accumulated mutations over the evolutionary time. Subsequently, using published data, they identified new candidate cis-regulatory elements, transcription factor binding sites, and chimeric transcripts in human embryos associated with the degTEs. The concept of improving existing TE annotations using shared ancestry of several hundred genomes is novel. Although the authors were mostly applying different existing tools, rather than presenting a novel methodology, they have successfully demonstrated the utility of the improved TE annotation through multiple lines of downstream analyses.

I recommend the authors to share the code so that other researchers can customize the process and update the degTE annotations with the upcoming new release of genomes. Although the authors shared the degTE dataset, the overall confidence of the dataset is uncertain given the lack of false discovery rate estimate. In some cases, the conclusions drawn from the figures are not fully supported, so additional explanation and information are needed.

Major points:

1. The authors need to establish an estimate false discovery rate of their method, as is done by other methods (P-clouds, RepeatMasker) mentioned by the authors.
2. More explanation on the rationale of choosing these 6 RAGs is needed, since the Armstrong et al. Nature paper used different clades.

3. Provide the table of 250 non-canonical exons with degTEs.
4. Fig 2d: Alu, L1, SVA elements are not labeled, so cannot draw the conclusion in line 90-92 "Furthermore, TE divergence profiles revealed that a peak corresponding to evolutionarily young LINE-1, Alu, and SVA elements was only seen in hg38 and the simian RAG and was absent in the older RAGs".
5. Fig 3e: In addition to the scatterplot, add a plot or a table of sum of gained bases. Can't directly draw this conclusion by looking at Fig. 3e, because only certain subfamilies are labeled.
6. Line 92-93: " Together, these results confirm that precise reconstruction of TEs was achieved in the RAGs." is not supported by sufficient evidence. The findings of Fig 2 are consistent with our expectations, but do not support this claim.
7. Fig. 4a: PLS, pELS, etc are not defined. More details need to be provided at least in Methods.
8. Why is there no ZKSCAN1 in Figure 4c, but only ZKSCAN2/3/5?
9. In the section on degTEs contributing to chimeric transcripts, the authors might want to perform GSEA or other gene set enrichment analysis with all the genes containing degTEs to further probe the effect of TcGTs on human development

Minor points:
1. Provide a README file to explain the files shared within "Enhanced hg38 transposable element annotation and associated RepeatMasker outputs of hg38 chromosomes and reconstructed ancestral genomes (RAGs)"
2. Fig 2c: Why are total numbers of bases of TEs born in Eutheria larger in Boreoeutheria, and even larger in Euarchontoglires, and so on?
3. Fig 3a, explain the different rows in each track.
4. Fig 4b: add DBD in parentheses in the figure legends
5. Need a reference in line 161 "previously identified"
6. Line 155: "ZMYM2 has been reported to be involved in TE silencing18,19, which may explain their frequent association with degTEs." Needs more explanation why.
7. Line 165: "which significantly binds L2 integrants" Provide more information on the significant binding, e.g. P value.

Authors' response to the first round of review

Reviewers' comments (in black) with our responses (in blue):

Reviewer #1: Comments enter in this field will be shared with the author; your identity will remain anonymous.This manuscript was one of the more enjoyable papers I have read recently, and I recommend publication as a short report. The authors make use of recently published ancestral genome reconstructions to improve the annotation of TEs in the human genome. In doing so, they attribute an additional 10% of the human genome to TE sequences. By comparing these newly annotated "degTE" insertions with the binding profiles of transcription factors, they show that many cis-regulatory elements are likely derived from ancestral TEs. Overall, the paper is short but sweet, and the methods seem to be sound. I have no serious issues with technical aspects of the paper, but a few questions, and I feel that there are several obvious analyses that could improve the paper. That being said, these are not essential for the scientific validity of the paper, so I make them as general comments rather than requests for additional work.

We appreciate the reviewer's recommendation on the publication of this manuscript. We would like to leave the decision to the editorial team on which article format is best suited.
We provide our responses to the reviewer's questions and suggestions below.

1. My biggest concern with the method itself is that the RAGs are not annotated de-novo, but using existing DFAM libraries with RepeatMasker and the "-species human" parameter. I suspect that this will heavily bias the method towards finding sequences that are similar to existing human TEs, at the expense of TE families that may otherwise be well represented in the RAGs, but present at very low copy number in the present-day human genome. It would be very interesting to see how this approach compares to one using de-novo TE annotations of the RAGs (e.g. with RepeatModeller), and it would be more in keeping with the speciesagnostic approach of the progressive cactus aligner. On the other hand, if people are strictly interested in annotating the human genome, the method described here is likely best.

We appreciate the reviewer's suggestion on the use of a *de-novo* TE annotation method. We totally agree that our reference-based method might fail to annotate TEs that are highly

degenerated in present-day genomes and therefore are not included in TE databases. We have added a new paragraph discussing this limitation along with the issue raised in the point 2.

However, we do not think that the use of the RepeatMasker parameter "-species human" itself introduces the mentioned bias. According to the RepeatMasker help documentation (https://github.com/rmhubley/RepeatMasker/blob/master/repeatmasker.help), the parameter only limits its search space to TEs that are found in human and are predicted to have emerged in its ancestors. Since it uses the same TE consensus sequences built based on multiple genomes, this parameter does not affect the detection sensitivity of a given TE regardless of its abundance in the genome used in the parameter. In our setting, for instance, it excludes TEs found only in clades that are not ancestral to human (e.g. rodentspecific TEs) but still finds TEs quite rare in the human genome with the same sensitivity as for abundant TEs, as indicated by the identification of a lot of rare ancient DNA transposons as degTEs (Fig. 3D, Table S2). Thus, our current approach is lineage-restricted but speciesagnostic.

We chose a reference-based approach to achieve an enhanced TE annotation that requires little downstream curation and is directly comparable to those already present in a large number of publicly available genomes. Although we would like to stick to this aim, we have included the use of *de-novo* annotation in Discussion as one of the future perspectives as shown below.

> *Another potential approach to RAG-assisted enhanced TE annotation is to construct TE libraries de novo probing RAGs. This may discover previously unidentified TE clades that have long been extinct and are highly degenerate in many present-day genomes. (lines 287-289)*

2. On a related note, it would be good to see a few sentences more on the limitations of the RAGs themselves. The genomic fraction attributable to TEs in RAGs is low (Fig 2B), and Fig 2D shows a striking lack of young TE insertions in the RAGs. While it is unavoidable that many "young" TEs present in the true ancestral genomes will be absent from the RAGs if they are not inherited by at least some of the species used for reconstruction, it is surprising that there are so few young families/insertions in the RAGs. As the authors mention at lines 57-60, TE insertions in the RAGs should - in theory - be closer to the original consensus sequences. I believe Fig S2b addresses this, but bit scores are (for most people) a less intuitive measure than divergence from consensus. If degTE insertions in the RAGs are indeed closer to consensus

than those in hg38, it would be interesting to hear the authors' thoughts on why young insertions seems to be depleted in Fig 2D.

We thank the reviewer for raising this excellent point. We have been puzzled by this observation likewise. We speculate that this may be due to the derivation of HMM models for these ancient TE subfamilies based on highly degenerate and domesticated sequences in present-day genomes. While this may accurately capture the diversity of sequences for a given TE subfamily, the resulting consensus might not necessarily reflect their original state, as it may include highly prevalent later-occurring mutations important for domestication. Thus, the ancestral sequences predicted in RAGs may appear “diverged” from the consensus, despite closer to the original sequence.

Accordingly, we have added a new paragraph in Discussion to discuss the limitation on the use of reference-based TE annotation and provide our hypothesis on the observation raised by the reviewer.

> *We present a reference-based TE annotation in RAGs, since we believe this provides the most comprehensible annotation of TEs. However, the use of consensus sequences constructed based on present-day genomes to probe ancestral genomes may in itself harbour limitations. It is likely that the consensus sequences generated this way contain mutations that are prevalent in modern genomes because they were important for the domestication of particular TEs but do not necessarily reflect their original sequences. This may explain why so few “young” or less diverged TEs were discovered in the RAGs in our study (Fig. 2D). This might be because the reconstructed TE sequences either appeared diverged despite being closer to the original state or were missed because they deviate significantly from the consensus. Ultimately, what might be required to further improve the quality of annotation in RAGs is reconstruction of full-length TE progenitor sequences for a number of TE families as was done for LINE-1[37]. (lines 304-315)*

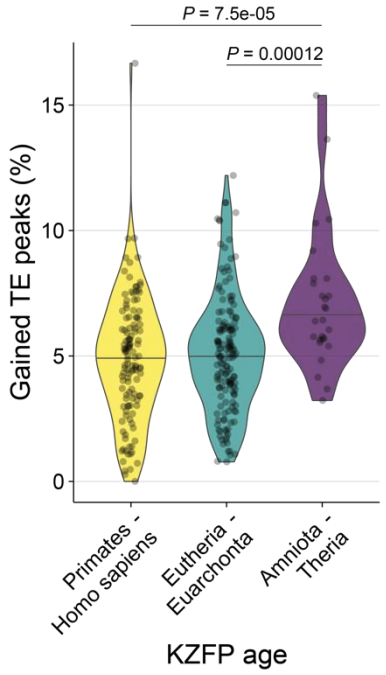
3. The paragraph starting line 125 is difficult to parse. In addition, I have a pedantic note referring to the use of the term "homology" (e.g. line 125, where the authors state that they "tested how much homology degTEs retained"). Strictly speaking, since homology refers to a binary state, "sequence conservation" or "similarity" would be more appropriate here.

Following the reviewer's suggestion, we have updated the corresponding paragraph as below.

> *Next, we tested how much sequence similarity each degTE retained, compared with the corresponding inserts found in the RAGs. For this, we searched degTE sequences across the Dfam TE library and calculated homology scores without the inclusion threshold imposed when annotation was performed with RepeatMasker. This analysis revealed that more than 20% of the degTE integrants displayed significant homology to the same subfamily as the one found for the corresponding sequences in the RAGs, whereas the remaining elements were either assigned to a different TE subfamily or showed no significant homology to any TEs (Fig. S3A). The homology scores obtained for degTEs were expectedly lower than those of the corresponding TEs in the RAGs (Fig. S3B). Also, we observed that degTE L2c elements exhibited similar coverage as those already annotated in hg38. The observed 5' truncation is known to be a typical feature for LINE integrants[16] (Fig. S3C). (lines 137-147)*

4. I found the section on KZNFs very interesting. I wonder if the authors have compared the age of the KZNFs targeting degTEs to those targeting more recent TE expansions such as L1PA subfamilies in human. If KZNFs targeting degTEs are significantly older than KZNFs targeting more recent TE expansions, it would lend further support to the idea that KZNFs facilitate the integration of TEs into novel gene regulatory networks.

We appreciate the reviewer's positive comment on our KZFP analyses. We have added a new supplementary figure that compares the fractions of peaks overlapping with degTEs among different KZFP age groups (Fig. S5). This confirms that the older KZFPs indeed recognise significantly more degTEs than their younger counterparts.



Reviewer #2: Here, Matsushima et al present a novel method for annotating transposable elements based on ancestral genomes. When applied to the human genome, the methodology expands the number of bps annotated as TEs by 10%.The authors characterize the overlap with functional genomic elements, concluding that the newly discovered elements play an important role in the evolution of gene regulation. Finally, they also demonstrate that 250 of these elements are part of human genes active during early embryonic development.

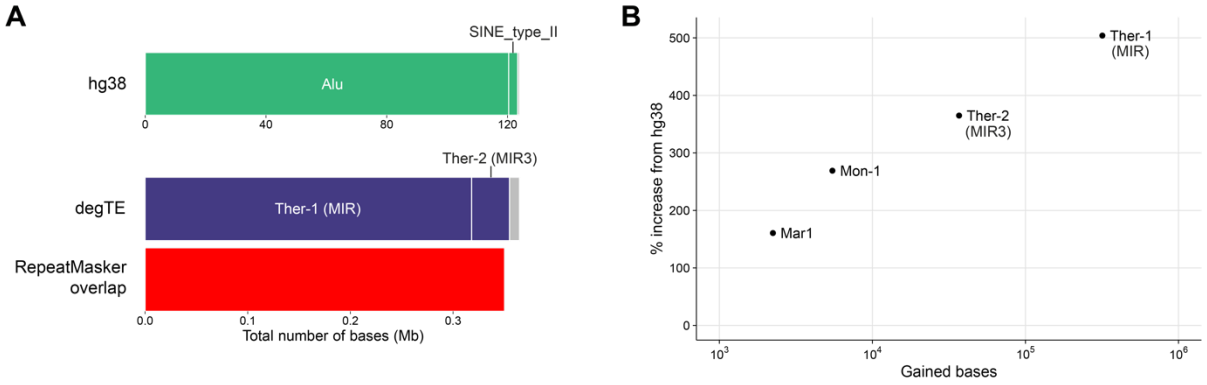
Overall, I find the manuscript well written with few typos and it is easy to follow the authors' line of reasoning. TE annotation remains a challenging and important problem as these elements are by far the largest class in most genomes. Although the work presented here provides an interesting advance, I do not think that it could be published in Cell Genomics in its current form. The following major issues need to be resolved;

We thank the reviewer for providing a positive evaluation on the quality of our manuscript and appreciating the advancement made by our approach. We take the issues raised by the reviewer seriously and address them as point-by-point replies below.

1. It is my understanding that the authors rely on a combination of RAGs and RepeatMasker for identifying TEs. Although I agree that RepeatMasker is the most widely approach for finding TEs, it is not the only method out there. It would be useful if the authors could apply other methods for TE detection (e.g. from Goerner-Potvinet al , Nat Rev Gen, 2018, Ou et al, Genome Bio, 2019) to help understand how sensitive the results are to RepeatMasker.

We appreciate the reviewer for raising this point. We realise that we may have not been clear enough regarding the goals of our methodology. Our claim is that, for a given TE annotation software tool, use of RAGs significantly enhances TE detection compared to examining only a single genome. We used RepeatMasker for our demonstration, but it was solely as an example, not to make any statement regarding its superiority over any other existing software packages. Since it is the use of RAGs, and only this, that provided the enhancement, other tools can be used instead of RepeatMasker depending on one's research questions, including tools optimised to detect specific TE families.

To further clarify that our approach is software-independent, we have employed another repository-based TE annotation method to test if similar enhancement can be achieved. We used SINEBase, a database specialised for SINE annotation, to perform the enhancement to the human TE annotation. Although the total increase in the annotation coverage is less pronounced because of highly abundant young Alu elements in the human genome, we achieved enhancement adding abundant SINEs mostly ancient Ther-1 (MIR) and Ther-2 (MIR3) elements. Importantly, 96% of the degTEs identified with SINEBase overlap with SINE degTEs discovered with RepeatMasker. The increase is up to five-fold for the most abundant element, showing that this TE annotation enhancement method is robust to different software tools.



In addition to the above analysis, we added new lines to Discussion to clarify that another software tool may be used.

> *Thus, although we provide a pipeline that employs RepeatMasker for TE annotation and six RAGs ranging between primate and Eutherian common ancestors, these should be adjusted to research questions asked. For instance, if enhanced annotation of a certain TE class is desired, there are several other software tools that are optimised for this purpose[32,41]. Also, if evolutionary trajectories of TE subfamilies at a higher resolution in a specific clade are sought (e.g. Alu subfamilies in primates), more RAGs corresponding to nodes within the clade should be screened. (lines 263269)*

2. The authors mention FDR for RepeatMasker and other TE finding approaches in the discussion. I understand that the lack of ground truth regarding TEs and degTEs makes it hard to provide these estimates for the results presented here. Nevertheless, it would be helpful to try or at the very least add to the discussion on how one could in principle calculate such estimates.

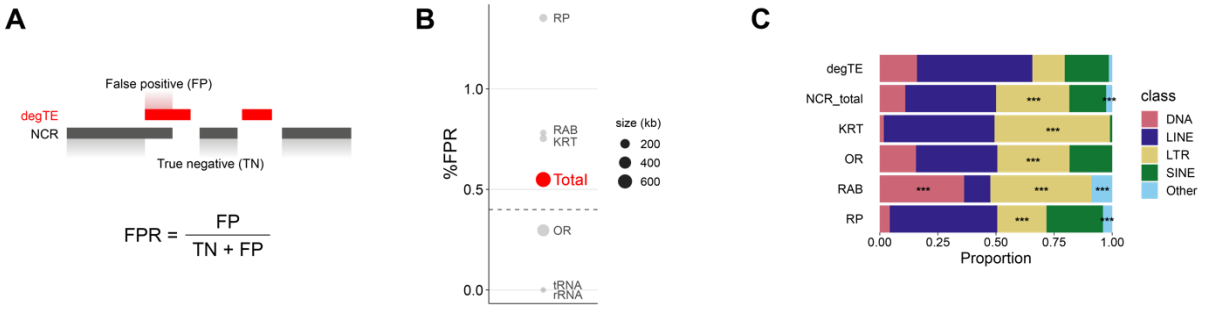
We thank the reviewer for raising this critical point. For any methodology, accuracy estimation is an important point to discuss, though, as the reviewer points out, it is generally challenging for TE annotation due to the lack of ground truth. In the paper where the falsepositive rates (FPRs) of RepeatMasker and P-cloud are reported, which we discussed in this manuscript, a computational algorithm, GARLIC, was built to construct background sequences as negative controls for the human genome and these sequences were used to estimate an FPR for several TE annotation tools.

For our method, however, we cannot simply apply GARLIC to estimate an FPR. Our approach comprises two major processes, ancestral genome reconstruction and TE annotation. GARLIC, or other similar methods, can calculate an FPR from the latter, but, to our knowledge, there is no prior study on estimating how likely it is to reconstruct a TE-like sequence from non-TE sequences.

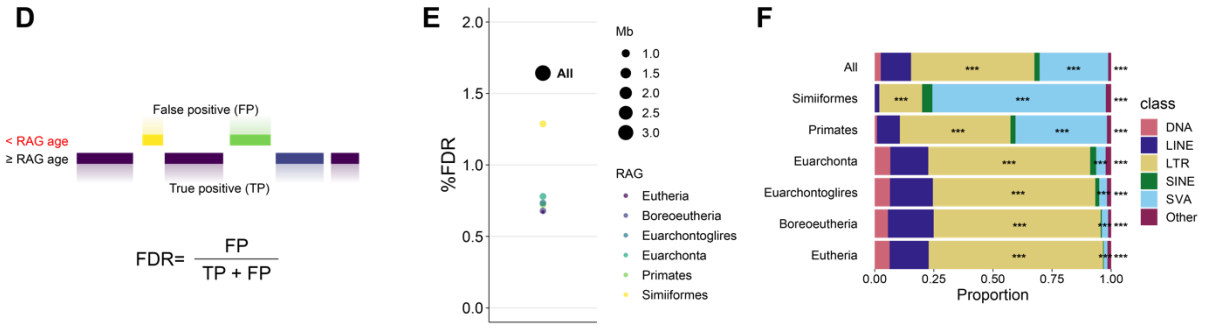
We therefore tackled this issue by defining negative control regions (NCRs), which are least likely to be of TE-derived (Fig. S2A-C). As an FPR is defined as the ratio between the number of false positives and the total number of negatives, we calculated an FPR of our method by dividing the bp overlaps between degTEs and NCRs by the total bp of NCRs. NCRs were chosen from the most abundant gene families (both coding and non-coding) in the human

genome that satisfy the following features: evolutionarily older than the mammalian common ancestor, highly conserved structures among paralogues, and high genomic coverage.

With this approach, we estimated the FPR of our method to be 0.56%. Since this figure for RepeatMasker alone is 0.4%, this analysis indicates that the ancestral genome reconstruction contributes to only an additional 0.16% of false positives.



In addition, we also approximated a false-discovery rate (FDR) by focusing on the ages of degTEs found in the RAGs (Fig. S2D-F). As TEs annotated in a RAG should have an age that is the same or older than that of the RAG, we calculated an FDR by dividing the coverage of TEs that are younger than a RAG (false positives) by total coverage of degTEs annotated in the RAG (total positives). With this, we estimated an FDR to be 1.6%.



We would like to note that multiple lines of results from other analyses also, though indirectly, suggest a low FDR of the method.

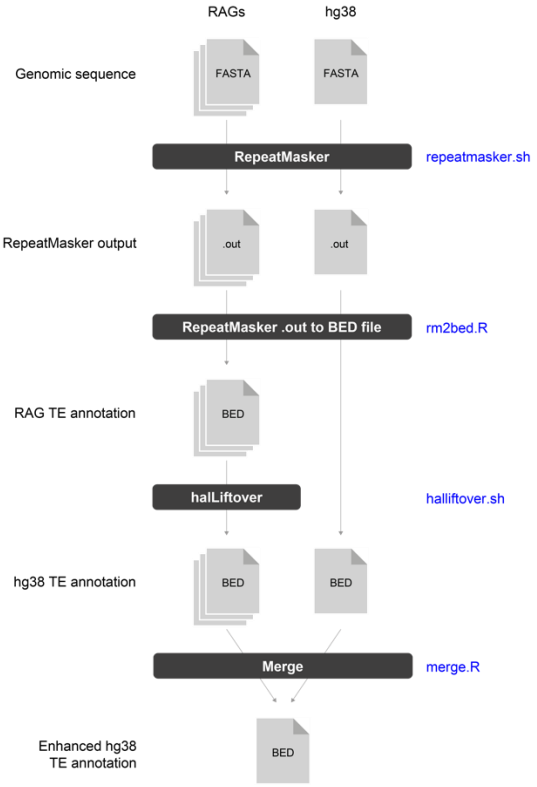
- The vast majority of degTEs belong to TE classes that are evolutionarily old (Fig. 3D).
- The coverage of degTEs identified as L2c in a RAG is similar to that of L2c already annotated in hg38 (Fig. S3C).

All of the above would not be observed if false positives of the method are frequent.

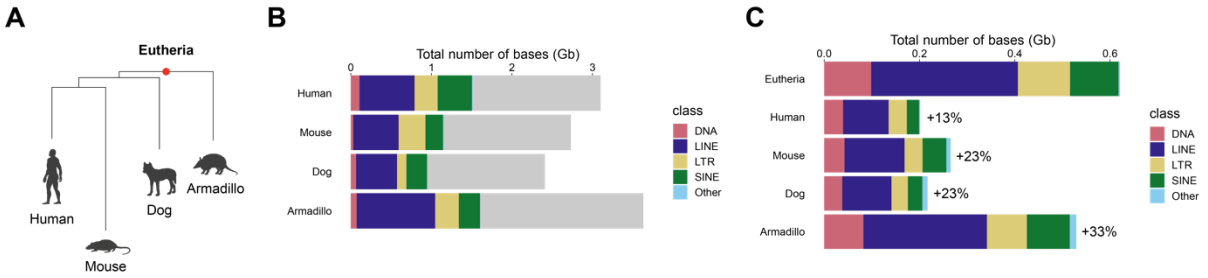
3. Although the authors publish their new TE annotations, this is limited to humans. There is a large community of biologists studying other genomes who might be interested in improved TE annotation. I see no reason why the authors could not make their software available as a package so that others can run it on other genomes. I understand that it may take time to run, but as long as this is clearly stated upfront, then users can decide on their own if they want to wait for results.

We agree that the human genome indeed is just one of many out there and are eager to provide our enhanced annotation to members of the research community working on a wide range of other species. One significant challenge in achieving this is that it would require TE annotation software/RAGs adapted to one's research question, and even the human enhanced TE annotation we share in this manuscript could be suboptimal for examining particular points. For instance, if one is interested in the evolution of primate-specific TEs, use of RAGs older than the primate common ancestor is uninformative, and more RAGs corresponding to multiple primate clades should instead be used. Also, the use of RepeatMasker may not always be the best depending on the TEs and genomes of interest.

Thus, instead of providing the pipeline as a fixed package, we share our scripts as separate modules on Zenodo so peers can easily modify them in accordance with their research questions. We also added a new supplementary figure (Fig. S8) that visually summarise the computational process for an easier execution.



Also, to further prompt the readers to perform the enhanced TE annotation on non-human genomes, we have performed the enhanced TE annotation on three additional species: mouse (*Mus musculus*), dog (*Canin lupus familialis*), and armadillo (*Dasypus novemcinctus*). As shown in the new Fig. S5, each genome gained a significant proportion of new TE annotation, indicating our method can be performed species belonging to distinct clades than primates.



4. The authors present several pieces of statistics showing overlap of newly discovered TEs with various functional elements. However, there are no null models presented to allow for statistical tests determining whether or not these overlaps are larger than expected than by chance. The authors should include suitable null models, hypothesis tests and p-values for these comparisons.

As degTEs are highly degenerate, we expect only a minor fraction of them to be associated with functional elements. This is already inferred from the fact that less than 80,000 cCREs are identified in 1,452,810 degTEs. Thus, we do not intend to claim that “degTEs are significantly associated with functional elements” but rather that “of the identified degTEs, a certain fraction is associated with functional elements”.

5. In the last section, the authors investigate human RNAseq data to identify genes containing parts of TEs, suggesting that they have functional roles. Although tantalizing, this analysis leaves much to desire.

We realise that we should have provided more background information on TcGTs and our aim in this section, which is to investigate whether the identified degTEs are included as transcripts in addition to serving as non-transcribed regulatory elements as shown in previous sections. Our claim here is that our new method facilitates the discovery of TcGTs, which have been implicated to play various biological roles in previous studies. We neither claim that the newly discovered degTE-derived TcGTs are functionally distinct from the previously found TcGTs in embryos nor that all of them are functionally important as any transcripts can simply be products from spurious transcription. To clarify these points, we have updated the corresponding paragraph as below:

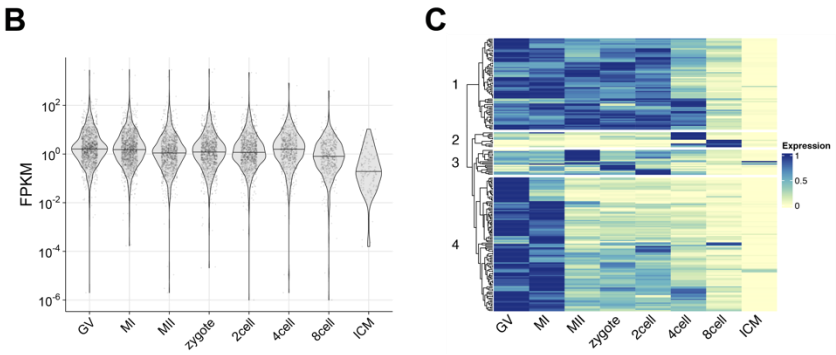
> *In addition to regulatory elements discussed above, TEs, including transpositionincompetent integrants, are known to contribute novel sequences to transcripts of the host. Non-canonical chimeric transcripts between TEs and genes, or transpochimeric gene transcripts (TcGTs), have been detected in various cell types[24]. A TE-derived sequence in a TcGT can initiate transcription[25], insert a non-canonical coding sequence[26], and serve as microRNA targets[27]. Especially, in the mammalian embryos, TEs have been reported to define stage-specific expression of the host genes by forming TE-driven TcGTs[6,7]. (lines 205-211)*

Also, we have significantly updated the main and supplementary figures to answer the raised issues. Details of the changes are shown below as point-by-point responses.

First, the authors need to report on the expression levels of these genes - how many are above meaningful thresholds?

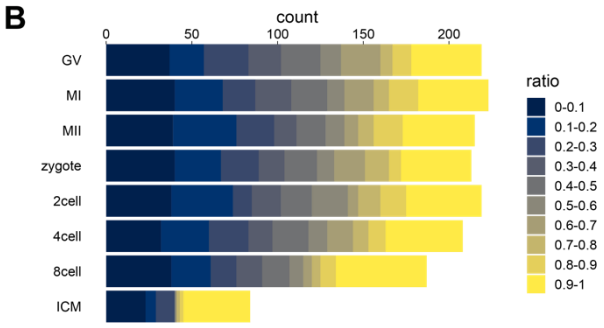
Following the reviewer's suggestion, we have imposed a new threshold to filter TcGTs that are expressed at >5 FPKM in at least one of the embryonic stages.

We also have added new plots summarising the distribution of the expression levels of TcGTs (Fig. 5B) and their relative changes during embryonic development (Fig. 5C). The source data are summarised in Table S4 as well.



Second, although the example shown in Fig 5 is convincing, it would also be useful to know how many of them constitute the dominating isoform.

Following the reviewer's suggestion, we have added a new plot reporting the relative expression levels of TcGTs among non-TcGT transcripts for a given gene (Fig. S6B).

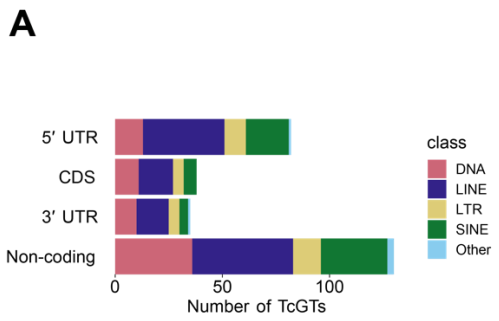


Third, what about the function of the 250 genes? Are they enriched for certain categories or pathways.

We do not expect it to be the case, since the newly identified TcGTs simply add to a large sum of TE-driven gene transcripts previously described by other (Göke et al. 2015). Also, as discussed below, many TcGTs are predicted to be non-coding. Of note, we anyways have run the suggested analysis but indeed no gene ontology terms/pathways were significantly enriched.

Fourth, what about the function of the new exons? The example in Fig 5 looks like it is a 5' UTR, so this probably has no impact on the protein (unless a new ORF is created). What about others? How long are they? Are they in-frame? Do they disrupt annotated protein domains? Taken together, more evidence is required to support the case that these chimeric TE transcripts are functionally important.

Following the reviewer's suggestion, we have performed an ORF prediction on the TcGTs with TransDecoder and summarised the positional information (UTR, CDS, etc.) of degTEderived exons (Fig. 5A). Other requested information is provided as a new supplementary table (Table S4). It is important to note that, other than altering protein sequences, a TEderived sequence in a transcript can exert biological functions though initiating transcription, serving as miRNA targets, or serving as long non-coding RNA (lncRNA).



The following minor issues need also be addressed:

It would be helpful if the authors could comment on the computational resources required to produce the new TE annotation as well as for some of the other analyses presented in the manuscript.

We have added a description of the computational resource required to complete our method in the Limitation section.

> *While this method is purely computational relying only on non-commercial software, use of a high-performance computing cluster is necessary to complete the analysis within a reasonable timeframe. Especially, whole-genome TE annotation with RepeatMasker is computationally intensive, and it took us approximately ~72 hrs with 24 CPUs (~1,728 CPU hrs) for each genome, though this may change proportional to genome sizes. (lines 317-321)*

There is no legend for the color scale in Fig 4c.

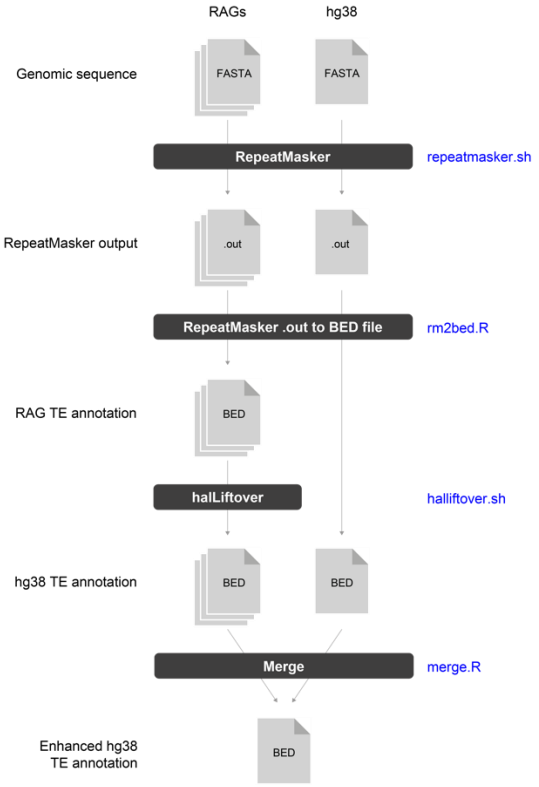
The colour scale of the TE and KZFP ages was provided in the figure legend, but it probably was too small. We have improved its visibility as below.

Reviewer #3: Matsushima and colleagues present the first attempt to identify additional reference transposable element (TE) insertions using multiple reconstructed ancestral genomes (RAGs). They identified >10% more reference TEs (called degTEs) that are missing from the current hg38 reference genome due to accumulated mutations over the evolutionary time. Subsequently, using published data, they identified new candidate cisregulatory elements, transcription factor binding sites, and chimeric transcripts in human embryos associated with the degTEs. The concept of improving existing TE annotations using shared ancestry of several hundred genomes is novel. Although the authors were mostly applying different existing tools, rather than presenting a novel methodology, they have successfully demonstrated the utility of the improved TE annotation through multiple lines of downstream analyses.

We thank the reviewer for appreciating the novelty of our study and for the interest of the improved TE annotation achieved through our method.

I recommend the authors to share the code so that other researchers can customize the process and update the degTE annotations with the upcoming new release of genomes.

Following the reviewer's suggestion, we have uploaded scripts used for the enhanced TE annotation to Zenodo. Also, we have added a new supplementary figure summarising our computational pipeline for an easier reproduction of our pipeline (Fig. S8).



Although the authors shared the degTE dataset, the overall confidence of the dataset is uncertain given the lack of false discovery rate estimate. In some cases, the conclusions drawn from the figures are not fully supported, so additional explanation and information are needed.

We have addressed these concerns as part of our point-by-point responses below.

Major points:

1. The authors need to establish an estimate false discovery rate of their method, as is done by other methods (P-clouds, RepeatMasker) mentioned by the authors.

Please see our response to the same issue raised in Reviewer 2's point 2.

2. More explanation on the rationale of choosing these 6 RAGs is needed, since the Armstrong et al. Nature paper used different clades.

We suppose that the reviewer refers to the L1PA6 analyses done in several RAGs in the Armstrong et al. paper. It is indeed true that they used a different set of RAGs from ours, but this was supposedly due to the fact that the L1PA6 family is Catarrhini primate-restricted and is not present in older RAGs. As our aim is to recover a wide range of TEs as much as possible,

we used the oldest available RAG and a few other RAGs in between. We added the following sentence to clarify our choice of RAGs as per the reviewer's suggestion.

> *We chose six RAGs ranging between the oldest available, Eutheria, and the simian common ancestor: Simiiformes, primates, Euarchonta, Euarchontoglires, Boreoeutheria, and Eutheria common ancestors (Fig. 2A). (lines 85-87)*

3. Provide the table of 250 non-canonical exons with degTEs.

Following the reviewer's suggestion, we have added a new supplementary table (Table S4), which includes information on the genes involved in newly found TcGTs.

4. Fig 2d: Alu, L1, SVA elements are not labeled, so cannot draw the conclusion in line 90-92 "Furthermore, TE divergence profiles revealed that a peak corresponding to evolutionarily young LINE-1, Alu, and SVA elements was only seen in hg38 and the simian RAG and was absent in the older RAGs".

We thank the reviewer's for pointing this out. We have provided a new Table S1 that reports the divergence levels of TEs at the family level, in which the above claim can be confirmed.

5. Fig 3e: In addition to the scatterplot, add a plot or a table of sum of gained bases. Can't directly draw this conclusion by looking at Fig. 3e, because only certain subfamilies are labeled.

We appreciate the reviewer's input; however, the requested information is already provided as Table S2 (Table S1 in the previous version). We have modified the paragraph to direct the reader properly.

> *Especially, the L2 and MIR subfamilies that emerged during this evolutionary time window significantly contribute to the new annotation, each gaining approximately an additional 10 Mb (Fig. 3D, Table S2). (lines 131-133)*

6. Line 92-93: " Together, these results confirm that precise reconstruction of TEs was achieved in the RAGs." is not supported by sufficient evidence. The findings of Fig 2 are consistent with our expectations, but do not support this claim.

Following the reviewer's advice, we have modified our statement as below. The point here is that if our method entails frequent false positives (calling non-TE-derived regions TEs), TEs annotated in the RAGs would not always have matching evolutionary ages as observed here.

> *Together, these results suggest that reconstruction and annotation of TEs were achieved in the RAGs without a high degree of false positives. (lines 94-95)*

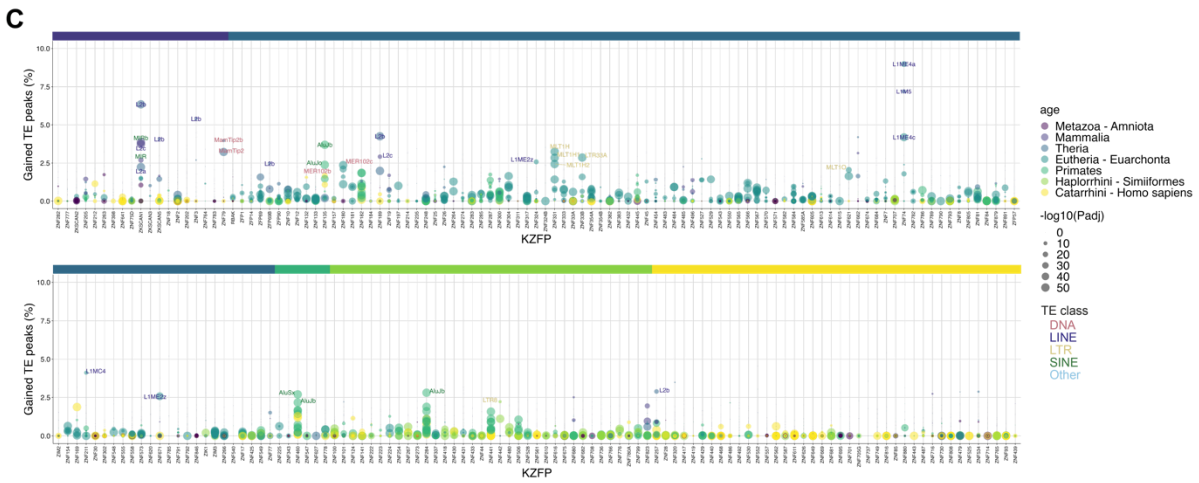
7. Fig. 4a: PLS, pELS, etc are not defined. More details need to be provided at least in Methods.

We have updated the figure legend of Fig. 4A to define the acronyms as suggested by the reviewer. Of note, these cCRE groups were defined and annotated in the previous study referenced in the main text[18].

> *(A) Percentage of cCREs overlapping the hg38 TEs and degTEs. The red numbers indicate cCREs that overlap with one or more degTEs. PLS, promoter-like signatures; pELS, proximal enhancer-like signatures; dELS, distal enhancer-like signatures.*

8. Why is there no ZKSCAN1 in Figure 4c, but only ZKSCAN2/3/5?

We apologise for dropping ZKSCAN1 from Fig. 4C. It was due to the space limitation, but it should have been included as it is used as an example in Fig. 4D. We have updated the Fig. 4C accordingly.



9. In the section on degTEs contributing to chimeric transcripts, the authors might want to perform GSEA or other gene set enrichment analysis with all the genes containing degTEs to further probe the effect of TcGTs on human development

Please see our response to the same issue raised in the point 5 from the reviewer 2.

Minor points:

1. Provide a README file to explain the files shared within "Enhanced hg38 transposable element annotation and associated RepeatMasker outputs of hg38 chromosomes and reconstructed ancestral genomes (RAGs)"

We apologise for the lack of clarity of the data provided on Zenodo. Following the reviewer's suggestion, we have added a description of the deposited files.

2. Fig 2c: Why are total numbers of bases of TEs born in Eutheria larger in Boreoeutheria, and even larger in Euarchontoglires, and so on?

We thank the reviewer for raising this point. Since precise transposition machineries for some ancient TEs have not been well understood, we can only provide our speculation. We presume that this is because either some of the TEs had been transposition-competent since their emergence for a while or they were increased through segmental duplications.

3. Fig 3a, explain the different rows in each track.

Following the reviewer's suggestion, we have updated the figure legend to clearly explain this point.

> *(A) (…) TE segments are visualised in multiple rows to indicate breaks of the segments.*

4. Fig 4b: add DBD in parentheses in the figure legends

Following the reviewer's suggestion, we have updated the figure legend for Fig. 4B.

> *(B) ENCODE TFBSs overlapping with one or more degTEs. The colours represent DNA-binding domains (DBD) of the TFs.*

5. Need a reference in line 161 "previously identified"

Following the reviewer's suggestion, we have added a reference to the corresponding line.

> *Moreover, we found that the binding sites of KZFPs that were previously identified as targeting evolutionarily old TE subfamilies overlap often with degTEs belonging to the same subfamilies[23] (Fig. 4C). (lines 191-193)*

6. Line 155: "ZMYM2 has been reported to be involved in TE silencing18,19, which may explain their frequent association with degTEs." Needs more explanation why.

Following the reviewer's suggestion, we have modified the corresponding sentence to clarify our claim.

> *ZMYM2 has been reported to be involved in TE silencing through its interaction with TE DNA[18,19], which may explain its frequent association with degTEs. (lines 181-183)*

7. Line 165: "which significantly binds L2 integrants" Provide more information on the significant binding, e.g. P value.

Following the reviewer's suggestion, we have added an adjusted *P*-value calculated with pyTEenrich (details in Methods) to the sentence.

> *To further probe the evolution of the KZFP-bound degTEs, we focused on a KZFP, ZKSCAN1, which significantly binds elements belonging to the L2 family (adjusted P < 2.2e-308). (lines 193-195)*

Referees' report, second round of review

Reviewer #1: My comments have been satisfactorily addressed, and I appreciate the additional effort to add new figures and analyses.

Reviewer #2: The authors have done a very nice job at addressing my comments and I think that the new manuscript is substantially improved.

Reviewer #3: All the reviewer comments were appropriately addressed.

Authors' response to the second round of review

NA