Cell Genomics, Volume *4*

# Supplemental information

# Ancestral genome reconstruction enhances transposable element annotation by identifying degenerate integrants

**Wayo Matsushima, Evarist Planet, and Didier Trono**

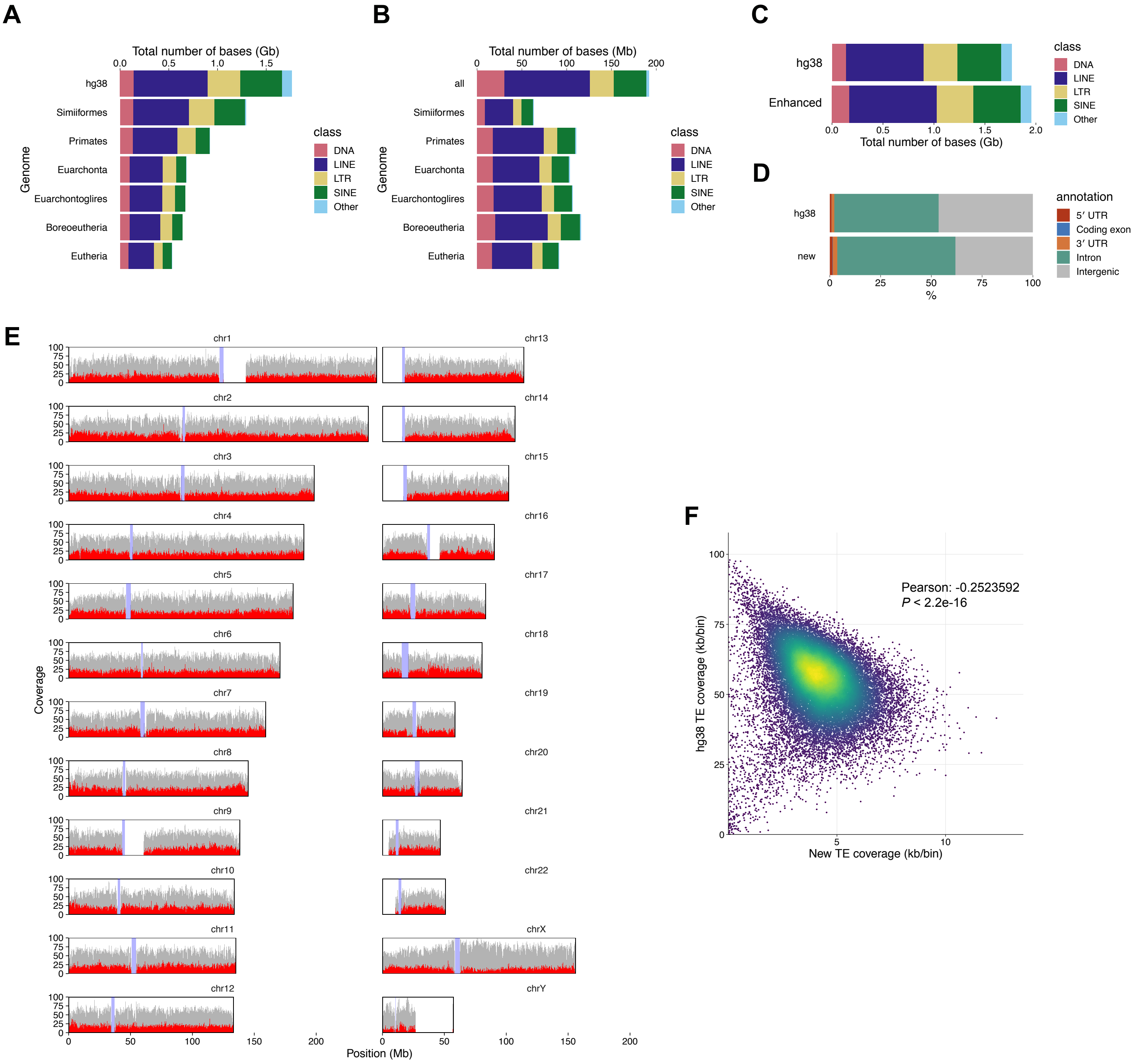

**Figure S1. Continued summary statistics of degTEs, Related to Figure 3**

(A) The number of TE bases of in the hg38 TEs and that lifted over to hg38 from each RAG. (B) Coverage of the lifted-over TEs from each RAG that do not overlap with the existing hg38 TE annotation. “all” represents the number combining the lifted-over annotations from all the RAGs. (C) TE class distributions in the hg38 and enhanced TE annotations. (D) Proportions of hg38 TEs and degTEs overlapping with distinct genomic annotations. (E) Genome-wide distribution of the hg38 TE (grey) and degTE (red) annotations. The y-axis represents a coverage (kb/bin) for 100-kb bins. The values for the degTEs were multiplied by four for a visualisation purpose. The blue shades represent the telomeric regions. (F) The same coverage data as (E) are shown in a scatter plot. Pearson’s correlation coefficient and *P*-value are shown.

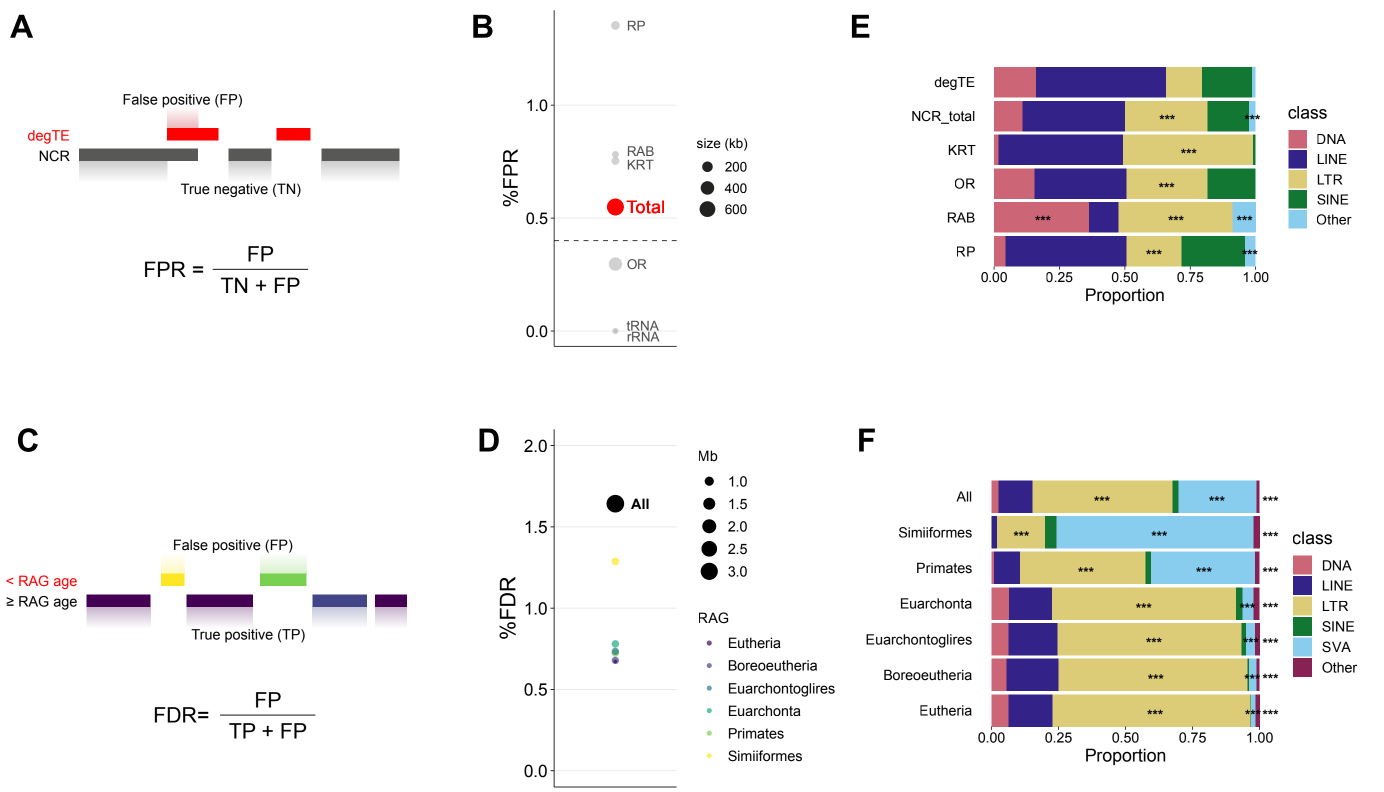

**Figure S2. Estimation of FPR and FDR of the method, Related to Figure 3**

(A) A schematic illustration of how FPR of the method was calculated using negative control regions (NCRs). (B) Six gene groups were used as NCRs, and FPRs obtained from each group as well as a total FPR are shown. RP, ribosomal protein; KRT, keratin; RAB, Rab GTPase; OR, olfactory receptor; tRNA, transfer RNA; rRNA, ribosomal RNA. (C) A schematic illustration of how FDR of the method was calculated based on the ages of TEs and the RAGs. (D) FDRs calculated for individual RAG and that when all the RAGs were used are shown. (E, F) TE class proportions contributing to false positives. TE classes that significantly more frequently overlap with NCRs than all the annotated degTEs (shown on the top) are highlighted with asterisks (Benjamini-Hochberg adjusted $P < 0.001$, one-sided Fisher's exact test) (E). TE classes that are significantly enriched in the degTEs that are younger than their RAGs compared to the proportions of all the discovered degTEs in each RAG are highlighted with asterisks (Benjamini-Hochberg adjusted $P < 0.001$, one-sided Fisher's exact test) (F).

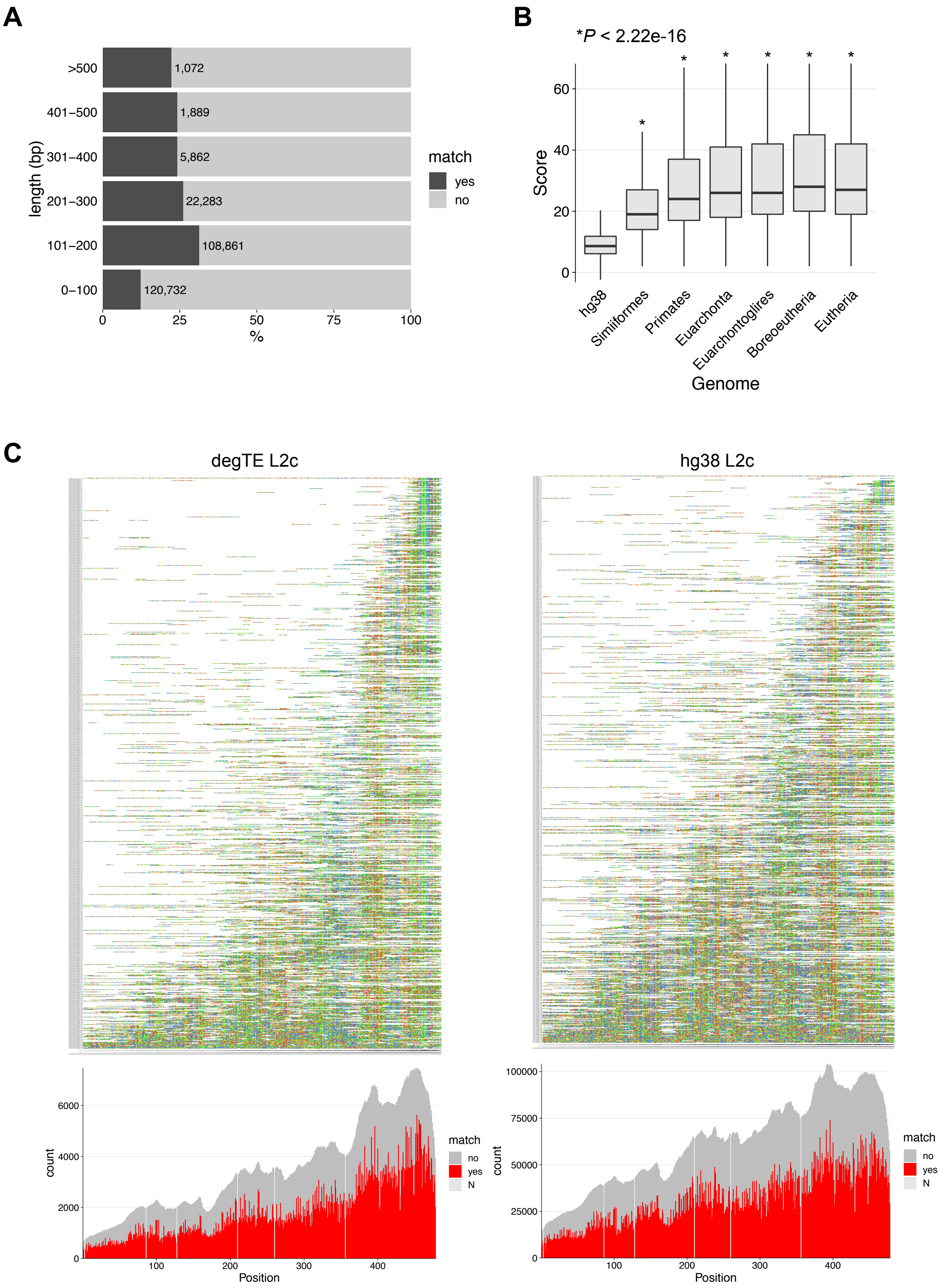

**Figure S3. Homology scores and discovery bias observed for degTEs, Related to Figure 3**
(A) degTEs corresponding to novel integrants were divided into multiple length bins, and percentages and the numbers of elements are highlighted for the ones exhibiting significant homology to the same TE subfamilies as the ones found in the corresponding regions in the RAGs. (B) Bit score distribution of degTEs and their corresponding sequences in the RAGs. *P*-values from Mann-Whitney *U* test comparing the distributions of degTEs and TEs in each RAG are shown. (C) L2c integrants aligned to the consensus. The left panes show L2c integrants in the Eutherian RAG that contribute degTEs, and the right panes show those found in hg38. The alignment plots were made based on randomly chosen 1,000 elements from each genome. Below, coverage plots over the L2c consensus is shown. For positions where the consensus base is N are shown in light grey.

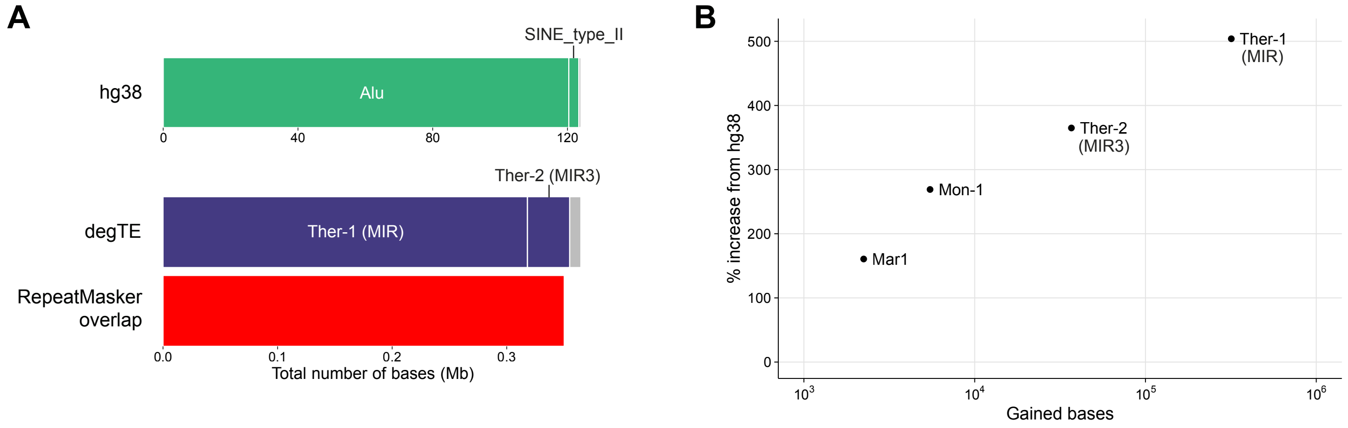

**Figure S4. Enhanced TE annotation achieved with SINEBase, Related to Figure 3**

(A) TEs discovered in hg38 and degTEs found by probing the Eutherian RAG. Only the top two TE subfamilies are labelled and the remaining families are shown in grey. Of the identified degTEs, those that were also annotated with the RepeatMasker-based method were shown at the bottom. (B) TE subfamilies that gained more than 1,000 bp are plotted to show the gained coverage and percent increase from hg38.

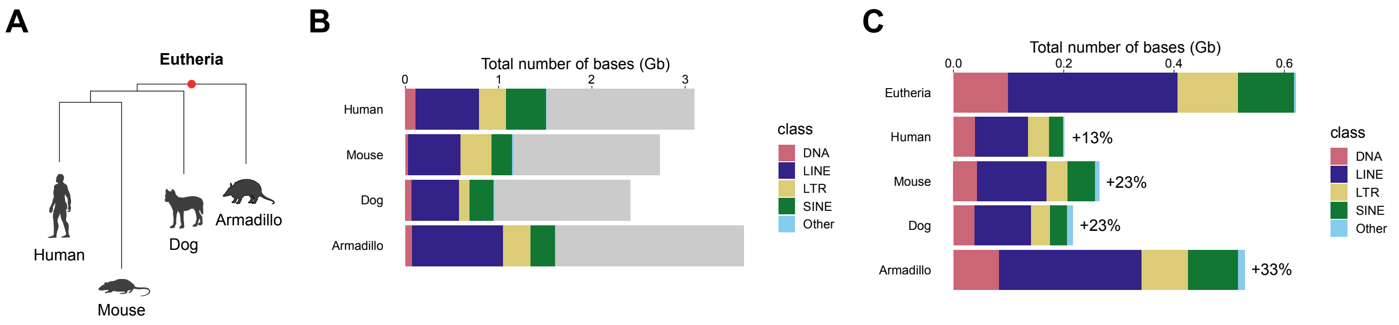

**Figure S5. Enhanced TE annotation of multiple Eutherian genomes, Related to Figure 3**
(A) A phylogenetic tree of the species for which enhanced TE annotation was performed using the Eutherian RAG (red point). (B) Proportions of TE classes annotated in each genome. Non-TE DNA is shown in grey. (C) TEs found in the Eutherian RAG and the coverages of it contributing as degTEs in each genome after liftover. The figures on the right represent the percent increase from the original TE annotation in coverage. Note that a different TE annotation was used for the human genome hence a different percent increase than the previous result.

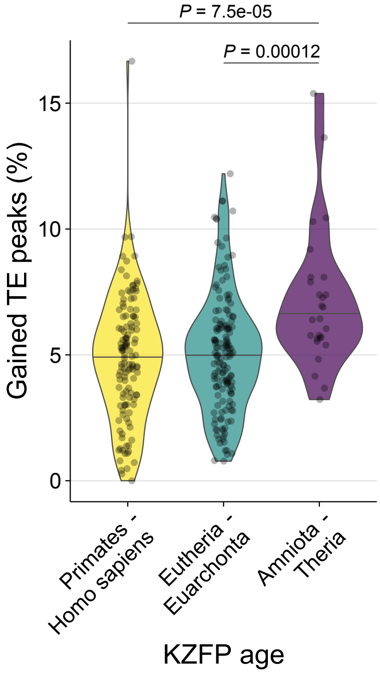

**Figure S6. Comparison of degTE overlap among KZFPs with different age groups, Related to Figure 4**
For each KZFP ChIP-seq data, the proportions of peaks newly associated with a TE through enhanced annotation relative to the total peak count are shown as a violin plot. *P*-values from Mann-Whitney *U* test comparing the distributions are shown on top.

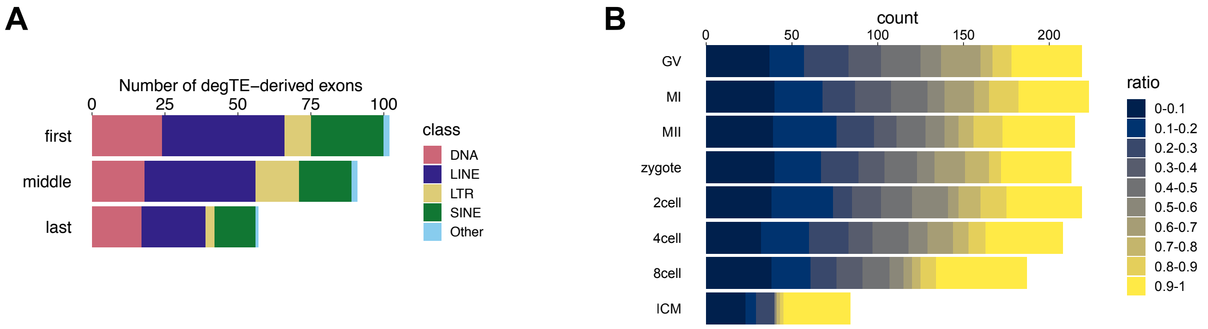

**Figure S7. Chimeric transcripts involving degTEs, Related to Figure 5**

(A) The numbers of degTE-overlapping exons that contribute to unannotated human embryonic transcripts with their positional information. (B) For TcGTs expressed in each stage of the human embryos, their relative expression in their associated genes is shown.

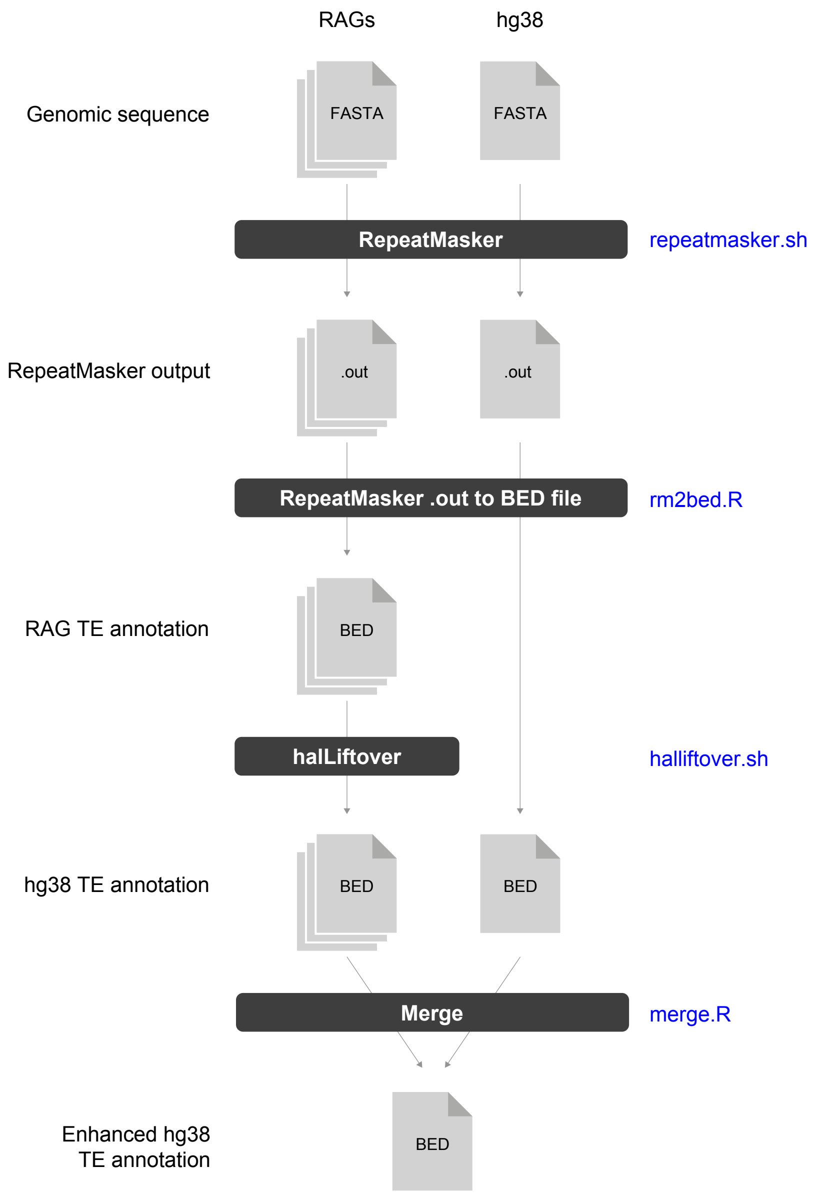

**Figure S8. Bioinformatic pipeline used to enhance hg38 TE annotation, Related to STAR Methods**
Schematic representation of the computational pipeline used to enhance the hg38 TE annotation in this study. The scripts used for each step are shown in blue and are available on Zenodo (doi:10.5281/zenodo.7716408).

| Node | Name |
|---|---|
| fullTreeAnc239 | Eutheria |
| fullTreeAnc238 | Boreoeutheria |
| fullTreeAnc115 | Euarchontoglires |
| fullTreeAnc114 | Euarchonta |
| fullTreeAnc110 | Primates |
| fullTreeAnc110point5 | Simiiformes |

**Table S1. Node names corresponding to the RAGs used in this study, Related to STAR Methods**