Supplementary Material for:

Mammalian NUMT insertion is non-random

Junko Tsuji[1], Martin Frith[2], Kentaro Tomii[1,2] & Paul Horton[1,2,3]

1. Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8561, Japan.

2. Computational Biology Research Center, AIST, 2-4-7, Aomi, Koto-ku, Tokyo, 135-0064, Japan.

3. To whom correspondence should be addressed. E-mail: horton-p@aist.go.jp. Tel:+81-3-3599-8064. Fax: +81-3-3599-8081.

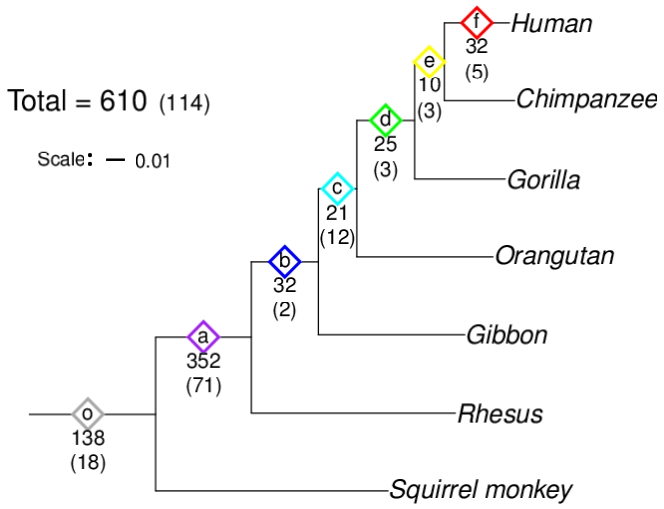

Figure S1: **NUMT phylogeny: insertion events during primate evolution.** The results of our NUMT age estimation is summarized. The number of NUMTs created at various stages of evolution along the path to humans is shown. For each age, the top figure given (*e.g.* 138) give the number of unique NUMTs and the bottom figure in parenthesis gives the number of other NUMTs The scale bar gives an indication of the number of nucleotide substitutions per site.

	P-values of NUMT flanks in different background	
oligomer	genomic background	comparable curvature background
TAT	2.0×10^{-15}	1.7×10^{-16}
TATA	1.4×10^{-11}	1.1×10^{-11}
TATAT	1.7×10^{-6}	8.3×10^{-12}
TATATA	4.2×10^{-4}	1.7×10^{-11}
TTTTAA	0.00139	0.01254

Table S1: **Oligomers enriched in human NUMT flanks compared to different backgrounds.** The statistical over-representation (by binomial test) of the presence of 3–6 mers in the 10bp flanks of NUMTs against two backgrounds is shown. Comparable curvature background denotes genomic regions which have a similar DNA curvature score distribution as NUMT flanks. For each length, the oligomer shown was the most significantly over-represented in NUMT flanks for both backgrounds. The last row shows P-values for the L1-EN consensus sequence.

	P-values of FAIRE(+/-) NUMT flanks			
oligomer	most frequent?	FAIRE$^+$ NUMT (freq.)	FAIRE$^-$ NUMT (freq.)	background freq.
TTT	+	3.1×10^{-21} (0.23)	6.2×10^{-2} (0.088)	0.077
TAT	-	1.7×10^{-19} (0.15)	2.4×10^{-19} (0.063)	0.041
TATA	+, -	1.4×10^{-11} (0.071)	1.5×10^{-11} (0.023)	0.013
TATAT	+, -	9.6×10^{-6} (0.034)	1.9×10^{-6} (0.010)	0.0046
TATATA	+, -	4.2×10^{-6} (0.031)	3.8×10^{-4} (0.0056)	0.0020
TTTTAA		0.2421 (0.014)	3.0×10^{-3} (0.0059)	0.0023

Table S2: **Oligomers enriched in human NUMT flanks in (non-)open chromatin regions in H1-hESC.** The statistical over-representation (by binomial test) of the presence of 3–6 mers in the 10bp flanks of 58 NUMT flanks in/near open chromatin and 1162 non-open chromatin NUMT flanks are shown. Open chromatin NUMTs flanks are defined as flanks which have annotated open chromatin within 10bp from their boundaries, as determined by FAIRE-seq with H1-hESC cell line (germ cell line). A '+' ('-') in the second column indicates the oligomer is the most frequent one of its length in the FAIRE$^+$ (FAIRE$^-$) NUMTs. Columns 3 and 4 give the P-value (binomial test) for FAIRE$^+$ and FAIRE$^-$ respectively. The parenthesized numbers in columns 3 and 4 (*e.g.* 0.23) give the overall frequency of the oligomer in FAIRE$^+$ and FAIRE$^-$ NUMTs flanks respectively (distinct from the frequency of at least one occurrence in each flank, used for the P-value computation). The final column lists the overall genome frequency.

Age	Number of NUMTs	Number of unique NUMTs
o	156	138
a	423	352
b	34	32
c	33	21
d	28	25
e	13	10
f	37	32

Table S3: **Number of human NUMTs from each phylogenetic age.** The number of human NUMTs for each phylogenetic age is shown. Unique NUMTs are those which show no evidence of duplication in the nuclear genome after their creation.

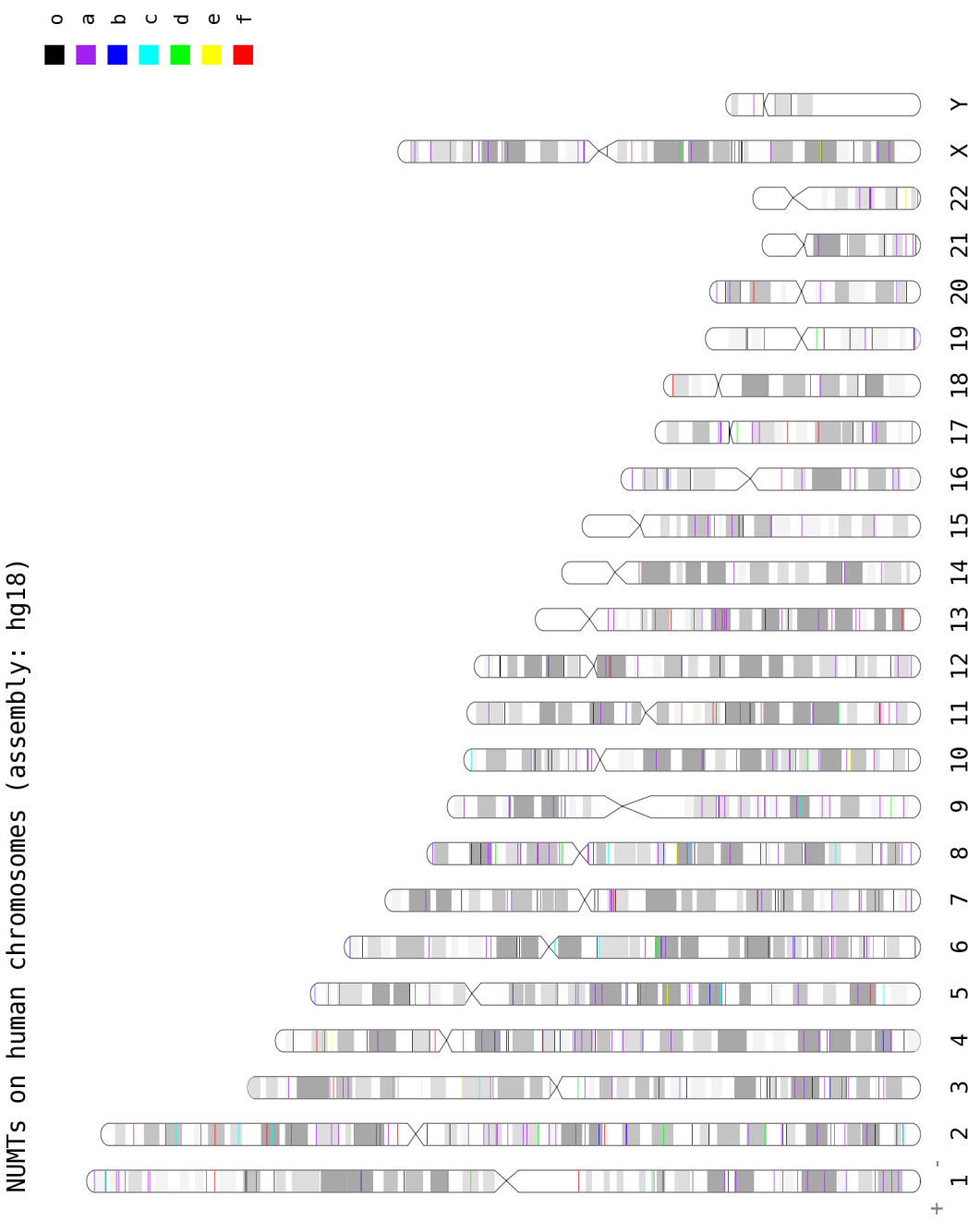

Figure S2: **Karyogram of non-duplicated NUMTs.** The chromosomal location of non-duplicated NUMTs is shown. The color of each NUMT indicates its age.

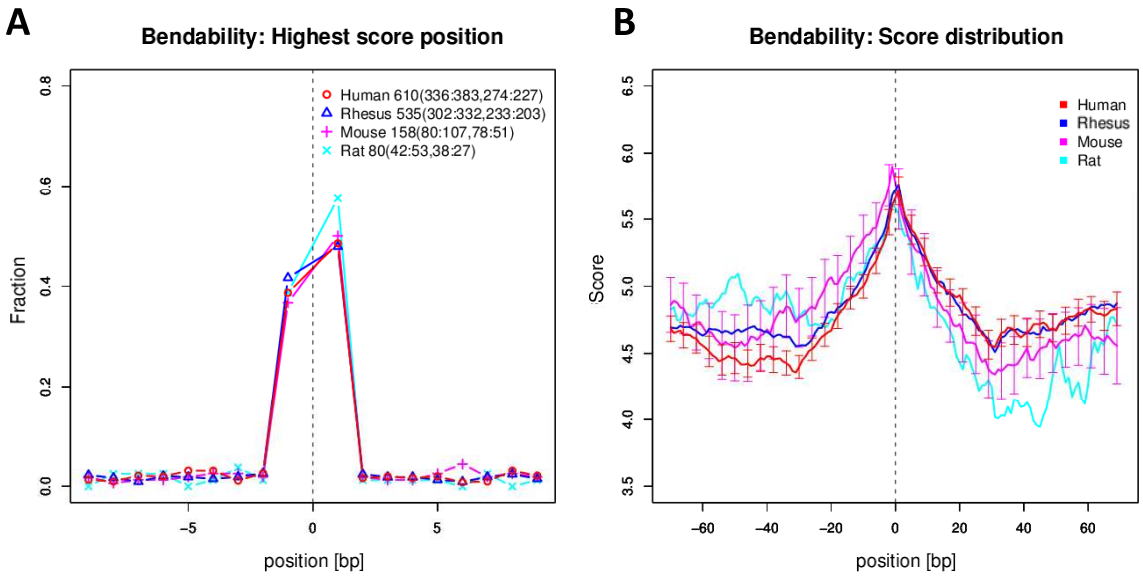

Figure S3: **Predicted DNA bendability in NUMT insertion sites. A**: The horizontal axis gives the distance from the inferred NUMT insertion site. The vertical axis gives the fraction of human NUMTs which attain a local maximum (within the 20bp window shown) in predicted DNA bendability. **B**: The score distribution of DNA bendability in concatenated NUMT flanks is shown. The vertical bars represent standard error. Clear peaks of bendability are observed at inferred NUMT insertion sites.

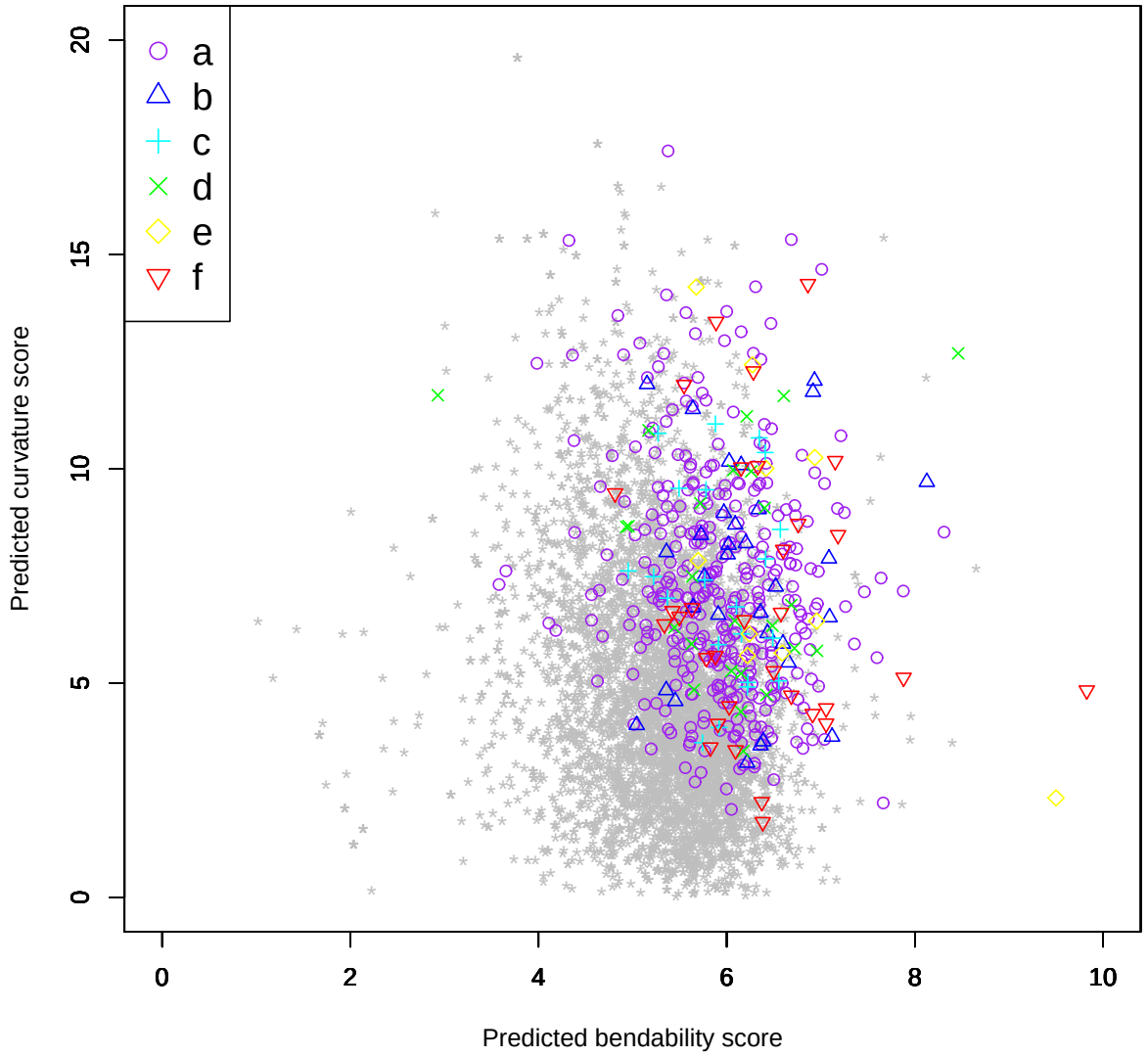

Figure S4: **Predicted DNA curvature and bendability in NUMT insertion sites.** A scatter plot of predicted DNA curvature and bendability at the inferred insertion site of human NUMTs is shown. Points representing NUMT insertion sites are colored by age (a:purple, b:blue, c:cyan, d:green, e:yellow, and f:red). Each NUMT is represented by a single point corresponding to one of the two bases bordering the inferred insertion site (more precisely, the "downstream" as determined by the arbitrary choice of mtDNA reference strand used to find NUMTs). The gray points represent 5,000 randomly chosen human genome loci.

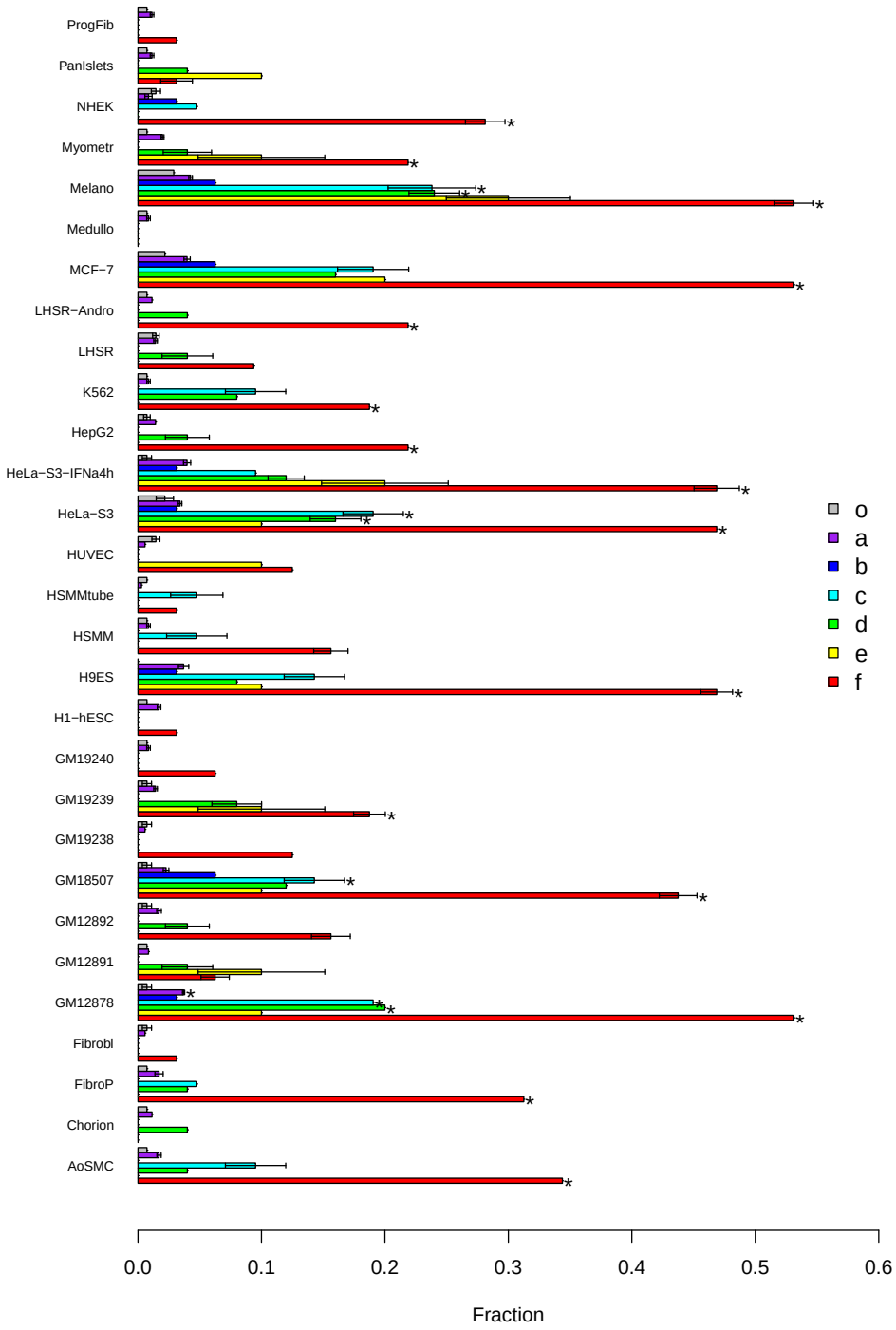

Figure S5: **Correlation between NUMTs in each age and open chromatin defined by DNase-seq.** The horizontal axis shows the fraction of NUMTs flank boundaries within 10bp of open chromatin regions defined by DNase-seq, and the horizontal bars represent the standard error of the mean fraction (when it is not identically zero) based on three replicate DNase-seq experiments. The bars marked with asterisks exhibit statistically significant correlation with open chromatin regions. The labels on the left hand are the cell line names.

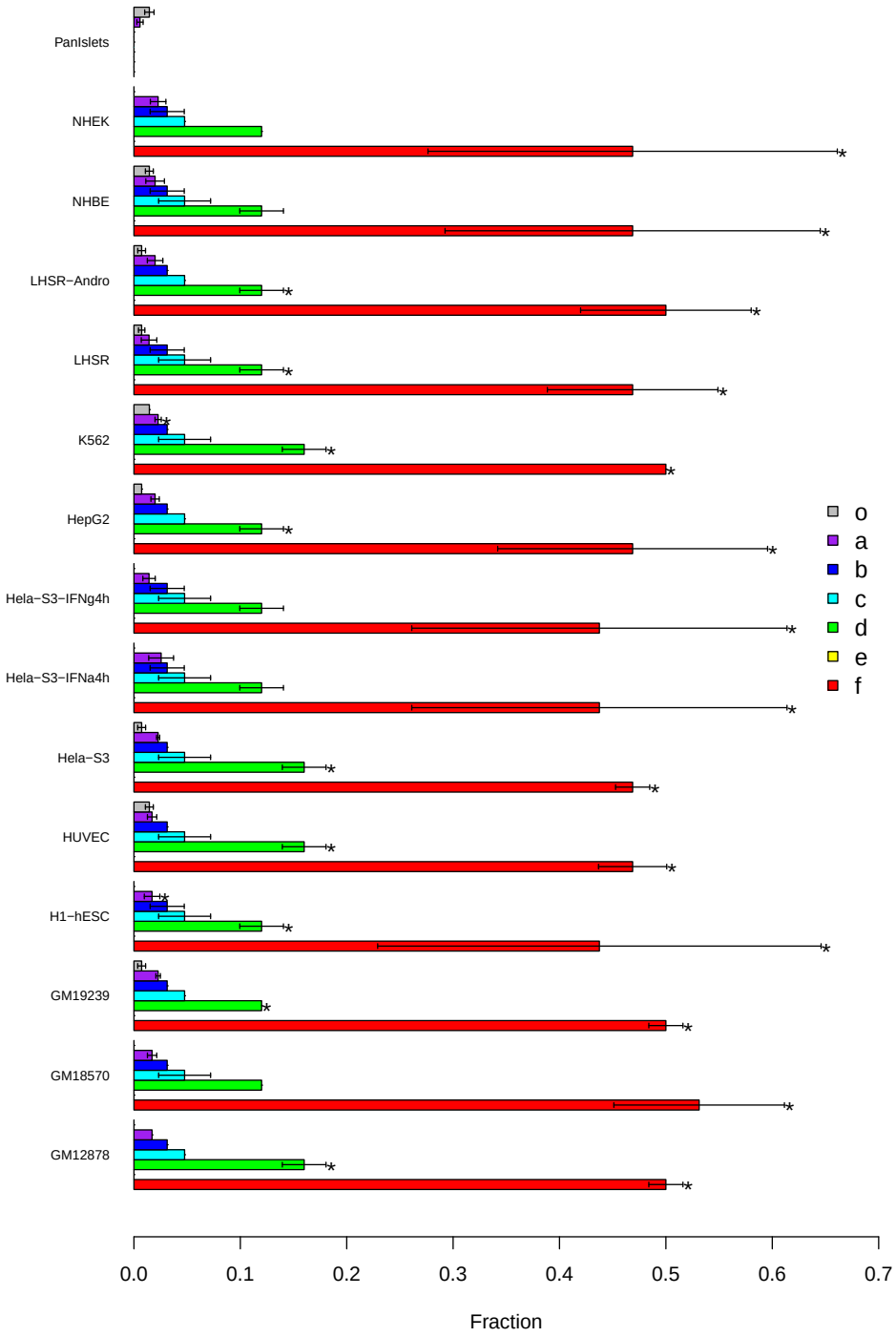

Figure S6: **Correlation between NUMTs in each age and open chromatin defined by FAIRE-seq.** The horizontal axis shows the fraction of NUMTs flank boundaries within 10bp of open chromatin regions defined by FAIRE-seq, and the horizontal bars represent the standard error of the mean fraction (when it is not identically zero) based on three replicate FAIRE-seq experiments. The bars marked with asterisks exhibit statistically significant correlation with open chromatin regions. The labels on the left hand are the cell line names.

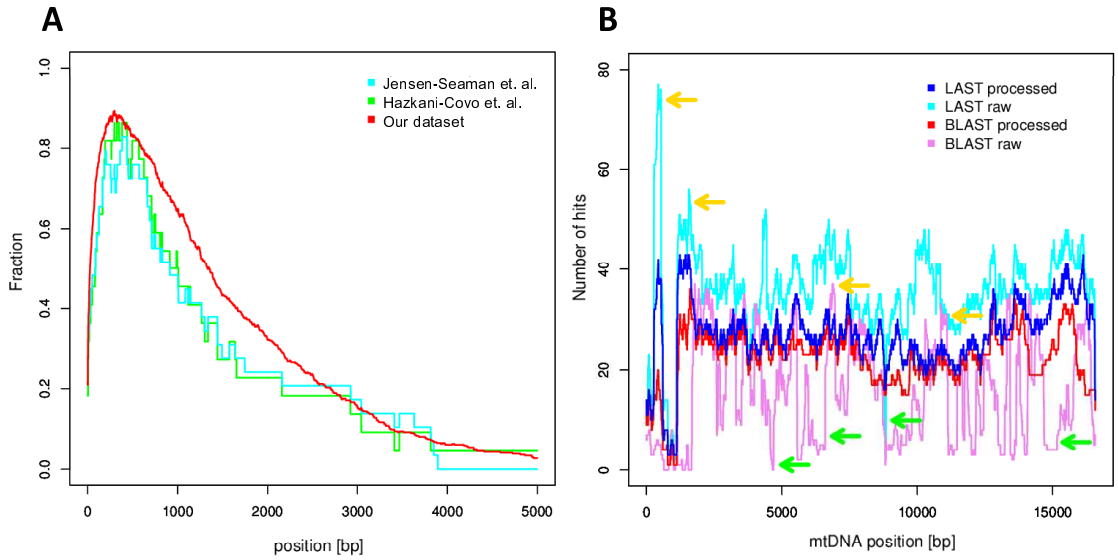

Figure S7: **Effect of dataset and computation details on conclusions. A**: The effect of different NUMT datasets and alignment score parameters on estimated NUMT flank retrotransposon density is shown. The horizontal axis gives the distance from the inferred insertion point, while the vertical axis shows the fraction of NUMT flanks which overlap with a `RepeatMasker` identified retrotransposon. Except that, to allow direct comparison with the analysis of Jensen-Seaman et al. [1], for each NUMT flank only the closest retrotransposon was considered (unlike Figure 1B in the main text). The red curve shows the density for the 610 NUMT flanks from our dataset. The blue and green curves represent density of `RepeatMasker` detectable retrotransposons in the flanks of the 37 and 32 NUMTs defined in the Jensen-Seaman et al. [1] and Hazkani-Covo et al. [2] datasets respectively. **B**: The effect of different alignment score parameters and our LAST (BLAST) hit post-processing on the distribution of mtDNA source regions for NUMTs is shown. The horizontal axis gives the mtDNA position and the vertical axis shows the number of NUMT progenitors found at a given position. The number of times each region of the mtDNA is the source of a NUMT is plotted for datasets prepared by various methodologies. The pink and cyan curves indicate raw BLAST (default parameters) hits and LAST hits respectively. The red and blue curves showed BLAST (red) and LAST (blue) hits after hit post-processing as described in methods. The yellow arrows highlight positions in which raw BLAST and raw LAST hits overestimate and the green arrows highlight where BLAST and LAST underestimate counts.

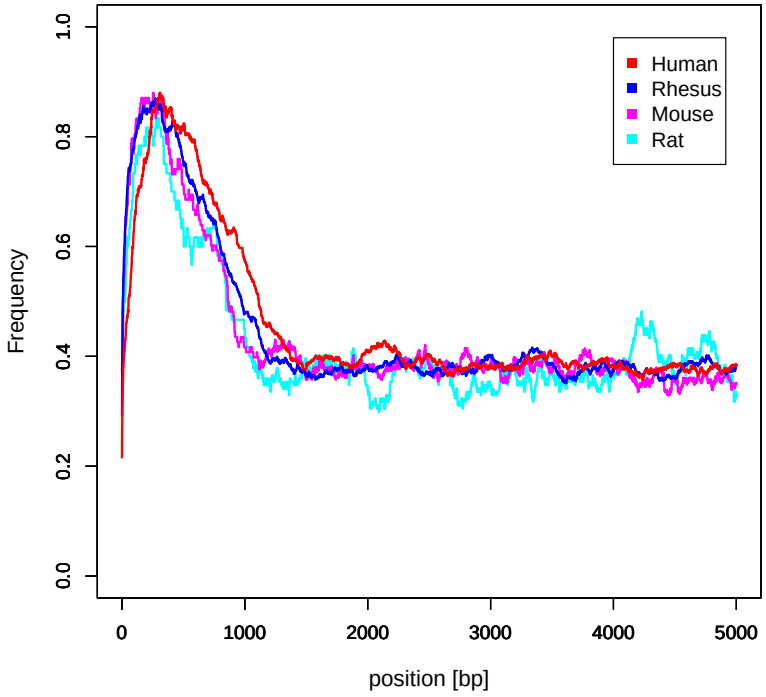

Figure S8: **NUMT-flanking retrotransposons detected from a local `RepeatMasker` run.** The fraction of bases found in retrotransposons is shown for NUMT flanks. This calculation was performed by running `RepeatMasker` locally on NUMT flanks, with the upstream and downstream flank of each NUMT concatenated together. The result is very similar to the results obtained by using the UCSC website `RepeatMasker` track "rmsk" directly (*cf.* Figure 1A in the main text).

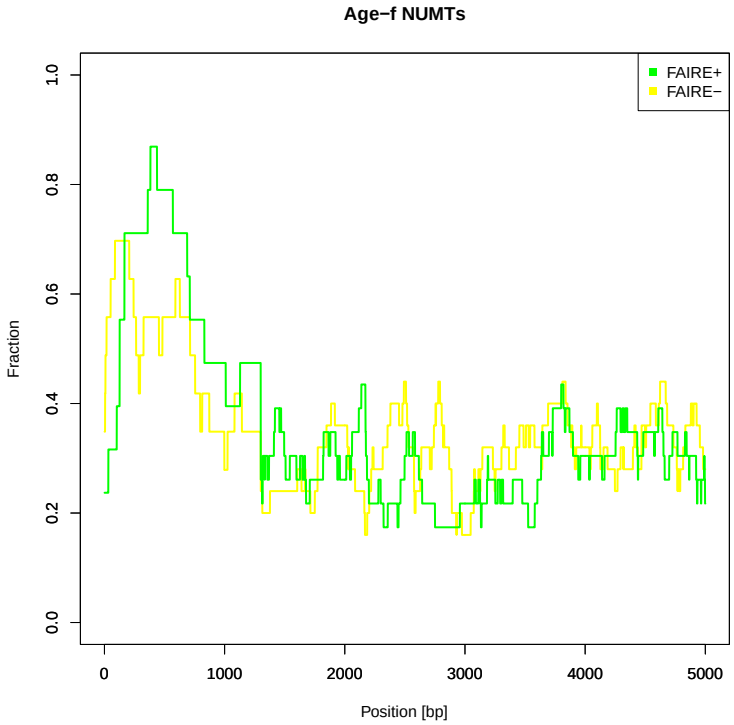

Figure S9: **Retrotransposon densities of NUMTs (not) in open chromatin regions defined by FAIRE-seq.** The fraction of bases found in retrotransposons is shown for the youngest age (f) NUMT flanks which are or are not near (10bp) open chromatin as defined by FAIRE-seq on H1-hESC cells. A broadly similar enrichment of retrotransposons is evident in the flanks of both classes of NUMTs.

Pitfalls to avoid when identifying NUMT insertion sites

Many previous studies have examined the features of NUMT integration sites and source mtDNA. Despite this fact, we report several previously unnoticed findings. We speculate the main reason for this discrepancy is different methodology used to detect NUMTs and define their boundaries.

Use of inappropriate scoring

As mentioned in the main text, we use alignment scoring parameters which are better suited for the range of evolutionary distance we are investigating.

Under and Over-counting original NUMT insertion events

Many older NUMTs in the nuclear genome have diverged from the original mitochondrial fragments by nuclear duplications, indels or retrotransposons. NUMT insertion events may be overestimated if no consideration is given to the possibility of post-insertion nuclear duplication (Figure S10A). Conversely, a single NUMT may be split into two smaller pieces, which may be missed and lead to under counting (Figure S10B) – especially if overly conservative scoring parameters or thresholds are employed. On the other hand, if both pieces are detected, but not chained together, large indels can also lead to over-counting. The circular nature of mtDNA is another factor which could lead to under counting of NUMT creation events (Figure S10C).

Certainly, some studies have addressed some of these issues, for example Mourier et al. [3] considered both the circular nature of mtDNA and chaining of BLAST hits with the DBA program. However, we believe that overall we have been the most careful and thorough in our NUMT defining methodology.

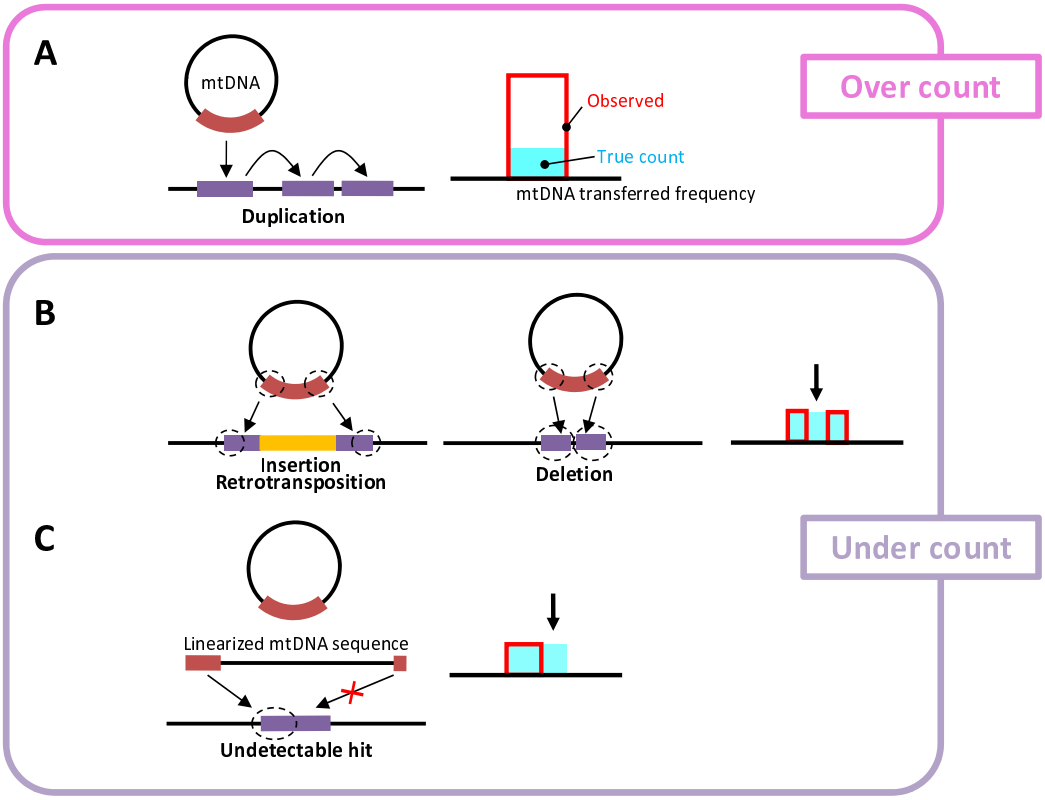

Figure S10: **Complications in counting original NUMTs.** Some ways in which original NUMTs might be miscounted when sufficient care is not taken during dataset preparation are shown. **A**: Post-insertion duplication of original NUMTs may cause over-counting. **B**: Insertions into NUMTs or (partial) deletion events may make the remaining NUMT fragments undetectable. **C**: Failure to take into account the circular nature of mtDNA may miss small NUMTs which cross the artificial boundary of the linearized sequence.

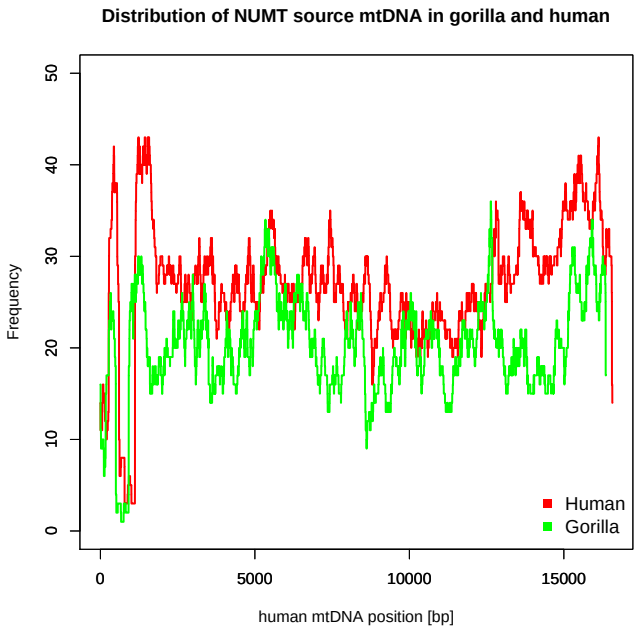

Figure S11: **Comparison of source mtDNA in gorilla and human.** A histogram of the frequency of forming NUMTs is plotted against the coordinates of the mitochondrial genome (mtDNA) for human and gorilla.

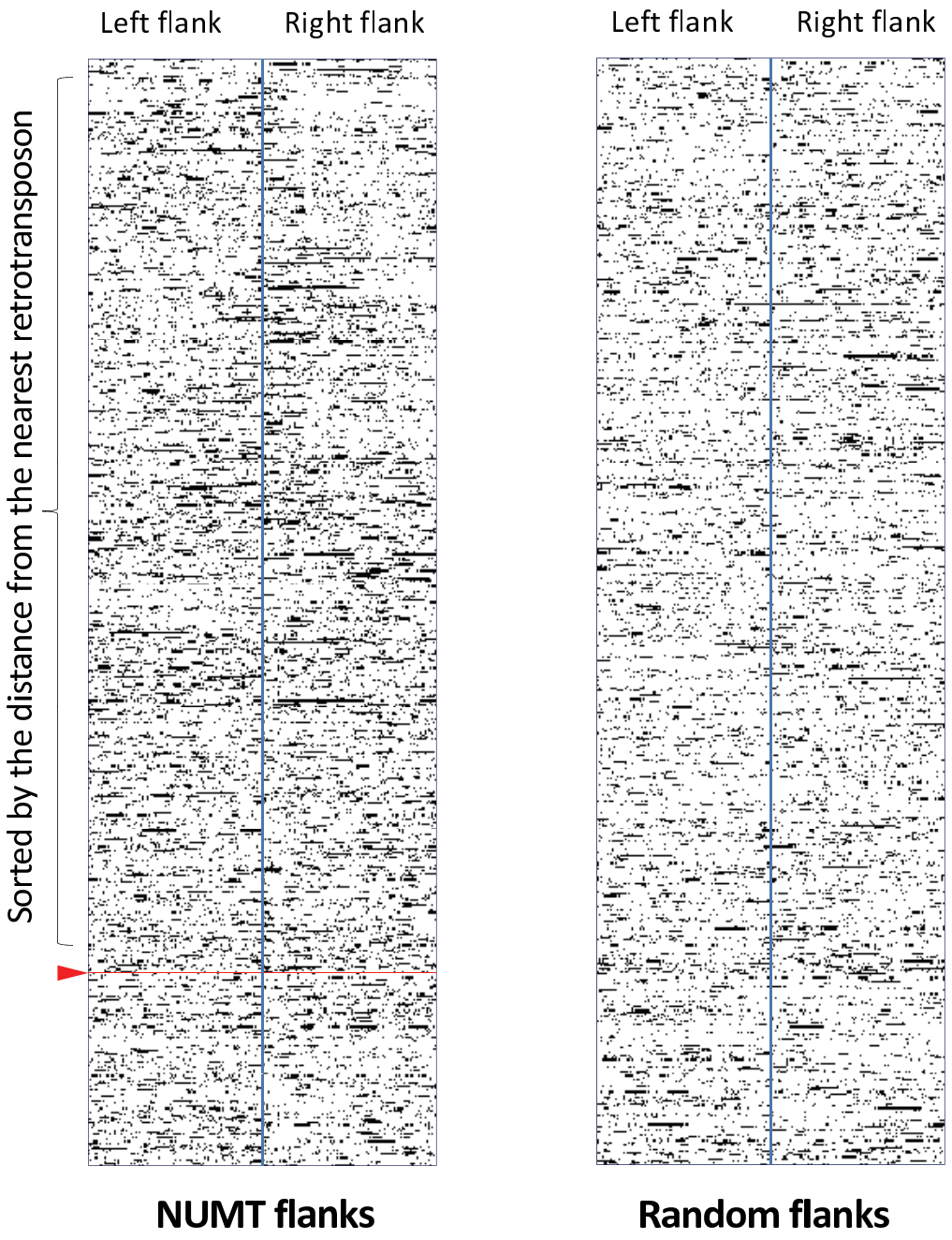

Figure S12: **Distribution of A+T rich oligomers in NUMT flanks.** The left pannel shows information for real NUMT flanks and the right for simulated, randomly inserting NUMTs. For each panel, each row represents two flanks from the same NUMT. The "upstream" direction is arbitrary (determined by the strands listed in the nuclear and mtDNA genomes we downloaded). The real NUMTs are sorted by distance to nearest retrotransposon up to 5000bp, with small distances at the top of the figure. A dot is plotted for each position in the center of a 5-mer consisting purely of A+T.

References

[1] Jensen-Seaman, M. I., Wildschutte, J. H., Soto-Calderon, I. D., and Anthony, N. M. (2009) *Journal of Molecular Evolution* **68**, 688–699.

[2] Hazkani-Covo, E., Zeller, R. M., and Martin, W. (2010) *PLoS Genetics* **6(2)**, e1000834.

[3] Mourier, T., Hansen, A. J., Willerslev, E., and Arctander, P. (2001) *Molecular Biology and Evolution* **18(9)**, 1833–1837.