Supplementary Materials for

From telomere to telomere: The transcriptional and epigenetic state of human repeat elements

Savannah J. Hoyt *et al*.

Corresponding author: Rachel J. O’Neill, rachel.oneill@uconn.edu

Science **376**, eabk3112 (2022)
DOI: 10.1126/science.abk3112

The PDF file includes:

Materials and Methods
Figs. S1 to S53
References

Other Supplementary Material for this manuscript includes the following:

Tables S1 to S28
MDAR Reproducibility Checklist

Supplementary Material.

Tables:

Additional Files: Supplementary Tables.xlsx Tables S1-S4, S7-S12, S14-S21, S23, S25, S27-S28

1. Data Availability

Repeat Library for previously unknown Repeat Entries, UCSC assembly hub browser, RepeatMasterv2 Track CHM13v1.1, RepeatMaskerv2 Track GRCh38 + chrY, RepeatMaskerv2 Track HG002 chrX, RepeatMasterv2 Track Composites CHM13v1.1, RepeatMasterv2 Track previously unknown satellites and arrays CHM13v1.1, all scripts and codes used herein	(*121*)
Sequencing data and assemblies	BioProject PRJNA559484
Sequencing data, assemblies, and other supporting data on AWS	(*11*)
PRO-seq CHM13/RPE-1	PRJNA559484
PRO-seq HeLa	GSE179576
RNA-seq CHM13	PRJNA559484
CHM13v1.1 Meryl 21-mers and 51-mers	(*114*)

2. Repeat Annotations

Repeats model discovery with RepeatModeler & loci identification with RepeatMasker

To assess previously unannotated repetitive regions of the genome, a RepeatMasker4.1.2-p1 run was completed on the T2T-CHM13v1.1 assembly using the Dfam 3.3 library (*6*) with the following settings: sensitive setting (-s), using the species tag of human (-species human) and the NCBI BLAST-derived search engine RMBlast (*-e ncbi*): *$ RepeatMasker -s -species human -e ncbi.* These regions (RA1a) were then hard-masked, producing “hard masked genome 1”, HM1 (fig. S1). A RepeatModeler2.0.1 analysis was performed on the remaining (unmasked) regions. The output file, T2T-CHM13_Modeler_Repeats.fa, was run through LTR Harvest (*123*) (accessed from genometools/1.5.10) and Transposon PSI (v08222010) (*124*) to further refine previously unknown repeat calls. As the RepeatModeler2.0.1 algorithm implements a random sampling of the genome, the consensi generated from RepeatModeler2.0.1 (All_T2T-CHM13_repeats.fa) were used as a library for a secondary RepeatMasker run to collect all associated instances for each model generated on the T2T-CHM13 genome assembly (RA1b).

Annotation of tandem repeats and satellites

Tandem repeats and satellites were initially annotated in the above RepeatMasker run, based on a combination of alignments to satellite sequences in the RepeatMasker library and de novo repeat identification with Tandem Repeats Finder (TRF) v4.09 GUI version using standard parameters (*125*). This workflow left large sections of the genome unannotated, either because repeats within the sequence were too decayed to be recognized by TRF, or because the satellites were not contained in RepeatMasker’s database. We expanded annotation coverage of missing repetitive regions using ULTRA (*9*), an open-source tool that can annotate and provide statistically consistent scoring for very large repeat units (up to a repeat period of 4000), arbitrarily-long repetitive regions, and ancient repeats that have highly decayed repetitive signals. ULTRA v1.0 was run with the settings: *$ ultra -mi 2 -md 2 -p 4001 -mu 2 -ws 90000 -os 10000*, respectively impacting maximum number of insertions (-mi), deletions (-md), and repeat periodicity (-p), minimum number of repeat units (-mu), and window size and overlap (-ws and -os).

GAP identification

Gaps in the T2T-CHM13 repeat annotation were identified via bedtools v 2.29.0 (*89*) by subtracting RA1a and RA1b (fig. S1) from the whole T2T-CHM13 genome sequence. The resulting regions were then filtered for size (only gaps larger than 5Kbp were considered). These gaps were manually curated in a UCSC Genome Browser session to check for any feature annotation overlap. Tandemly repeated sequences for each gap were identified with a combination of TRF v4.09 (*125*) and ULTRA (*9*). Additionally, self-alignment plots generated in YASS (*126*) confirmed sequence repeat monomer length. Monomers and full tandem arrays of these regions were compared via alignment in MAFFT v7.471 (*107*) and Geneious v2019.1.3 to check for any possible overlap with either the region or previously annotated tandem repeats; only previously unknown repeats were kept and classified as "previously unknown array/monomer".

Manual curation of previously unknown repeat models

Following the production of RA1b, curation steps were implemented to refine previously unknown repeat models and produce RA1c (fig. S1A, S1B). Overlaps with CAT/gene annotations (*11*), segmental duplications (*65*), and tandem repeats found within GAPs and through ULTRA overlaps were manually curated. In the case of segmental duplications, if a previously unknown repeat was found only as a component of linked segmental duplications, it was not considered a repeat for classification. If, however, the repeat was found outside of linked segmental duplications, it was considered a repeat that had been captured by duplication events after its formation. Multiple sequence alignment (MSA) plots of the transposable element (TE) instances from RA1a aligned to the putative repeat consensi from RA1b were used to determine the divergence, and therefore the overall age of the TE family. Highly diverged, older sequences were set aside for later assessment (table S1), as these sequences may correspond to old fragments of known TEs. This hypothesis was reinforced following a cross_match analysis of the older consensi to the Dfam database consensi. In addition, repeats that failed to match a known repeat, even distantly, were assessed by evaluating 100 nt on the 5' and 3' flanking region of the instances contributing to the initial RepeatModeler consensi. RepeatMasker was used to assess the flanking regions to determine if the neighboring sequence matched consistently to known repeats and were saved for later evaluation as possible ancestral repeats (table S1). To confirm the set of additional repeat families (Tables 1 and 2, table S2) had not been previously defined in the Dfam database, we performed a cross_match analysis of the consensi corresponding to the repeat families to the curated consensi in the Dfam database. In addition, a cross_match self-comparison of the consensi was performed to assess possible array sequence structures or intra-library duplicates. Intra-library duplicates and matches to the Dfam database were removed. The identification of composite subunits was accomplished through assessment of genome-wide instances with Circa and pattern recognition in both the UCSC browser and RM output. BEDtools closest v2.29.0 (-k 2 -iu -D ref) and (-k 2 -id -D ref) was also used to assess the neighboring repeats and their frequency increasing the likelihood that they were part of a larger repeat, or composite. This curation led to the generation of a repeat library (RMv2) (final_repeats.fa) of previously unknown or unannotated repeats, satellites monomers, variants of previously known/classified satellites, subunits of composites and composite elements.

Compilation and polishing of Repeat Annotationv2

The models discovered as a result of the RepeatModeler2 analysis contained pieces of simple repeats and small pieces of previously defined TEs. As such, a RepeatMasker analysis performed by simply adding entries alongside previously annotated TE/repeat models in a library resulted in a large number of false positives. Therefore, a pipeline was developed (fig. S1C) to combine the additional entry annotations and the previously generated TE models in the Dfam database to produce a high confidence repeat masker annotation track for CHM13. A third RepeatMasker run was performed on HM1 using a library which included the Dfam database plus all additional entries resulting in Repeat Annotations 1c (RA1c). RA1c was then combined with RA1a (Dfam library only) and the resulting combined outputs were intersected with gap entries ("previously unknown array/monomer") and additional family entries. Additional family entries were filtered for elements with high confidence based on MSA plots and a SW score of 250. These combined efforts resulted in the production of a final T2T-CHM13v1.1 RepeatMaskerv2 track (RMv2) for the UCSC genome browser.

Following development of RMv2, unit-length and composite unit genomic instances were determined by performing a self-comparison via cross_match to determine the maximum SW score for a particular TE model. A score conservatively lower than the determined maximum was then used in the alignAndCallConsensus.pl program (-sc #) to align the TE instances to the consensus. The resulting MSA were used for the classification of composite repeats and subunits therein (*6*, *127*, *128*) .

Reverse liftOver analysis and repeat fasta comparison

Liftover chains were generated from LASTZ sequence alignments between GRCh38 and T2T-CHM13. It is important to note that while liftOver will accurately give the coordinates between the alignments of two assemblies, there could be problems with misalignments with large sequences, a persistent source of error in any alignment. Such a misalignment would result in the correct coordinate given as part of the LiftOver output, but not the same sequence contained within. This scenario is true for this study as GRCh38 contains gaps and estimated sequence sizes for centromeres and telomeres, leading to misalignments when compared to the more complete T2T-CHM13 assembly. In addition, different matrices automatically calculated based on GC content as part of the RepeatMasker program might be used to identify repeats in GRCh38 vs. T2T-CHM13, leading to different repeat boundaries and/or subtle changes in repeat annotations for aligned regions between the assemblies (and see (*19*)).

A bed file was generated from the T2T-CHM13v1.1 RM2 output and a reverse liftOver (*129*) performed to the GRCh38 genome assembly. The IntersectBED tool (*89*) was used with default parameters to compare the unlifted T2T-CHM13v1.1 coordinates with regions lacking synteny to GRCh38 and separate these coordinates into one of two categories: syntenic or non-syntenic (table S6-S7). The IntersectBed tool (*89*) was also used with both strict (-f 0.9 -r) and permissive (-f 0.5 -r) parameters to compare the lifted GRCh38 coordinates to the GRCh38 RM-comp output. The results of the intersection were parsed based on the TE annotation match. The possible categories of the intersection analysis included: full match, class match, family match, no match, set asides. A full match required, under either the strict or the loosened parameters detailed above, that both the family (e.g., *Alu*Sx) and class (e.g., SINE/*Alu*) of the T2T-CHM13 and GRCh38 RM-comp loci were identical. In the event the family differed, the intersected locus was labeled as a class match. If the family matched, but the subfamily differed it was labeled as a family match whereas if the class, family, and subfamily differed, it was subsequently labeled as a no match. Loci lacking a repeat match altogether following intersection analysis were set aside for a further detailed direct fasta comparison as described below. Special attention was paid to intersection loci in which the locus was identified as an SVA element, as these elements contain, and are frequently mislabeled, *Alu* elements. Exceptions were made in the level of match if an SVA in the T2T-CHM13 output matched to an *Alu* in the GRCh38 RM output and vice versa.

A parallel and complementary analysis comparing the set aside T2T-CHM13v1.1 loci fasta sequences was completed. A similarity score was assigned to each repeat based on crossmatch output as a percentage of the maximum score. Sequences with a score of greater than 90% and/or shorter than 50 bp were the threshold for concordant similarity or insufficient information for comparison, respectively. These were labeled as highly diverged and/or short loci. All other sequences were considered as potential polymorphic loci. The term *polymorphic* is used here to describe the genetic variation occurring between individuals in a population, such that each individual may contain a different repertoire of TE insertions.

Composite Elements

We defined a composite element as a repeating unit consisting of three or more repeated sequences, including TEs, simple repeats, composite subunits, and/or satellites, that is found as a tandem array in at least one location in the genome. A composite subunit is a previously unknown repeat annotation that is most often found within a composite. Note that a composite subunit repeat may be found outside of the composite, but it is not common. Segmental duplications (SDs) were called for T2T-

CHM13 using a 1Kbp cutoff (*65*); while the location of some composite elements within a family are present as a single copy and thus are likely SDs derived by non-allelic homologous recombination (NAHR) (*90*), a composite family is distinguished by the presence of composite elements in an array in at least one location, thus falling into a “megasatellite” classification (*130*).

Most composites are found in a tandem array only on a single chromosome (figs. S6A-F and S7B-G), and in eight cases each core unit contains protein-coding annotations (fig. S7), indicating that unequal crossing over events and concerted evolution among composite units contribute to the expansion or contraction of gene families within humans. Several of these composites were annotated as staggered segmental duplications encompassing only the tandem array (e.g. fig. S8). One composite, 5SRNA_Comp, consists of a portion of the 5SRNA, an *Alu*Y and two subunit repeats as an array of 128 repeating units with high sequence similarity on Chromosome 1 (fig. S9A,B). Monomers of 5SRNA_Comp are located at 49 locations across 13 chromosomes, (fig. S9C) and lack the *Alu*Y; rather they carry an LTR2 (fig. S9D). Thus, the distribution of monomers is likely the result of segmental duplication events through non-allelic homologous recombination (NAHR). In contrast, the *Alu*Y insertion (resulting in deletion of the LTR2) preceded the expansion of this composite into a high copy number array. Alternatively, TE-free copies (lacking either TE) expanded slightly and then two separate copies each experienced a de novo TE insertion (one with *Alu*Y and one with LTR2). Since the LTR2 copies only have a single LTR and no internal sequence, it is possible that after a full-length LTR2 insertion a NAHR event resulted in the near-complete loss of this sequence, leaving the solo LTR2 behind. The copies with *Alu*Y then expanded to form the array and the LTR2 copies were involved in more segmental duplications (as is common with Chromosome 9 and the acrocentric chromosomes).

Two composites are found arrayed across several chromosomes. The ACRO_Comp (*131*, *132*) is a unit found across 12 chromosomes (fig. S10A), including as tandemly arrayed sequences across the five acrocentric chromosomes (Chromosomes 13, 14, 15, 21, 22) with high sequence identity across composite units (fig. S10B). The LSAU-BETA composite is found across 16 chromosomes and in both tandem arrays and as single monomers (fig. S11A, B). The LSAU-BETA composite has a variant form (LSAU-5403) in CHM13 (fig. S11B) and includes subunit repeats consisting of D4Z4 (*133*) and LSAU ((*134*), overlaps with the *DUX4* genes and microRNA genes (*MIR8078*) and has been implicated in facioscapulohumeral muscular dystrophy (FSHD)(*133*, *135*). Complete reference sequence spanning these complex arrays afforded the opportunity to assess intra-array variability. We find that LSAU composites found in centromere transition regions share lower identity within an array (80-95%) than LSAU composites found within interstitial arrays (near 100% identity), illustrating the utility of the

CHM13v1.0 reference for future studies of the evolutionary trajectories of repeat arrays contextualized to chromosome location.

We annotated a highly complex composite, TELO_Comp, that consists of multiple arrays and other composites (Fig. 1E (top) and fig. S12). TELO-Composites are found on 10 chromosomes (fig. S12A) at interstitial, pericentromeric and subtelomeric loci. The canonical TELO_Comp consists of three 3Kbp composites (TELO-A, -B, -C subunits), each containing multiple TEs, downstream of a variable length array of a 49bp satellite repeat unit, ajax, bounded by a duplicated sequence, teucer (Fig. 1E and fig. S12B). CHM13 annotations for TELO_Comp TELO-A subunit were extracted from the genome as fasta sequences via bedtools (*89*) (table S8). Sequences were aligned with MUSCLE (*136*). The evolutionary history was inferred by using the Maximum Likelihood method and General Time Reversible model (*109*). The tree with the highest log likelihood (-8364.23) is shown in fig. S12. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying the Maximum Parsimony method. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.2836)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This analysis involved 24 nucleotide sequences with a total of 2845 positions in the final dataset. Evolutionary analyses were conducted in MEGA X (*137*, *138*).

Across 24 loci, all TELO_Comp elements contain a TELO-A subunit downstream of the ajax satellite array (fig. S12C). Among the subtelomeric elements (fig. S12C, blue arrows), all contain TELO-B and TELO-C subunits upstream of a shared subunit repeat found across all TELO-Comp elements (10479). In depth analysis of the overall structure of the subunits across all loci, and phylogenetic analyses of the TELO-A subunit (table S8), indicates that subtelomeric units are a monophyletic group of recent origin, while interstitial and pericentromeric units are polyphyletic. Elements within this latter group lack a TELO-B subunit; between the TELO-A and TELO-C subunits, all of these TELO_Comp elements contain a second 10479 subunit repeat, with the exception of three elements (Chromosomes 1, 10, 11) that also lack a TELO-C subunit. While high bootstrap support for the clustering of subtelomeric elements indicates recent derivation, likely by segmental duplication events (fig. S13A and table S9), location-specific repeat diversification in subunit content and structure as well as ajax repeat copy numbers, which retain high sequence identity (fig. S13C) is observed. Moreover, each subtelomeric unit contains the ajax array proximal to the telomere, indicating that inverted orientations are favored at subtelomeric loci.

On Chromosome 7, three TELO_Comp loci contain additional tandem arrays of the TELO-Comp subunit consisting of ajax and teucer sequences, with variable ajax repeat numbers and variable tandem arrays of the teucer-ajax subunit (Fig. 1E, fig. S13 and table S9). Further phylogenetic analyses for both ajax and teucer sequences reveal subtelomeric arrays evolve under neutral evolution while pericentromeric arrays evolve under concerted evolution (figs. S14-S15 and tables S10-S11). Moreover, phylogenetic analyses of ajax-teucer monomers from the Chromosome 7:56525533 locus indicate that both elements form a composite that evolves as a single unit, suggesting a higher-order repeat, or super-repeat, structure across that locus (figs. S14-S15). Collectively, the inclusion of composite elements in the annotation tracks for T2T-CHM13, afforded by the polished and contiguous reference assembly, provide the research community with a set of guideposts around which to pinpoint potentially pathogenic variants.

Dot Plot Analyses

To generate pairwise sequence identity dot-plots we used the software package StainedGlass (*139*). The input for this program is sequence fragmented into windows (1Kbp) after which all possible pairwise alignments between the fragments are calculated using minimap2 (*140*). The color used in the dot-plot was then determined by the sequence identity of the alignment which was calculated as:

$$ ID = 100 \biggl(\frac{M}{M+X+I+D} \biggr) $$

where ID was the percent sequence identity, M the number of matches, X the number of mismatches, I the number of insertion events, and D the number of deletion events. When there were multiple alignments between the same two sequence fragments all alignments other than the one with the most matches were filtered out regardless of their sequence identity. The resulting matrix of percent identity scores was then visualized using ggplot and geom_tile. All code and documentation are available at (*141*).

Methylation Metaplots

Nanopore CpG methylation data for T2T-CHM13 and HG002 was processed according to the methods outlined in (*21*). Genomic coordinates were normalized by the repeat start and end for each repeat type and CpG methylation frequency was calculated by fraction of methylated reads to total coverage within bins in T2T-CHM13 or HG002 with the BSGenome Bioconductor package (*21*, *91*). Multiples of three bins were further smoothed with the “rollmean” function from the R package Zoo (https://cran.r-project.org/web/packages/zoo/ index.html). Points represent the median smoothed methylation per bin and shaded ribbons represent the 25th and 75th percentiles.

Identification of full-length and truncated TEs; TE aging

The active families in the human genome for SINEs, LINEs,and retroposons are *Alu*Y, L1Hs, and SVA_E/F, respectively; the recently active family in the human genome for ERVs is HERV-K (*93*). Elements belonging to each family were extracted from the compiled CHM13 and GRCh38 RepeatMasker outputs. Full-length *Alu* belonging to the *Alu*Y family were defined as having a 3' start no shorter than 4 nucleotides (nt), and a 5' end position equal to or greater than 267 nt. A parallel *Alu*Y categorization was conducted based strictly on divergence (< and > 2%) to represent the youngest group of *Alu*Y elements and therefore more likely to be lineage-specific (fig. S18). Full-length L1Hs sequences were defined as having a length greater than or equal to 6000 nt. SVA_E and SVA_F elements were defined as full-length if the 3' start was no shorter than 50 nt with an end position greater than or equal to 1336 nt, to allow for length variability in the variable number tandem repeat (VNTR) region. ERV elements were subdivided into five categories based on sequence length and structure (based on presence of flanking 5' and 3' long terminal repeats (LTRs) and internal coding sequence). These categories are as follows: > 7500 bp elements with both 5' and 3' long-terminal repeats (LTR) (GT/LTR+), > 7500 bp elements with only one LTR (GT/LTR-), < 7500 bp elements with both 5' and 3' LTRs (LT/LTR+), < 7500 bp elements with only one LTR (LT/LTR-) and < 7500 bp elements with a complex combination of LTR and internal sequences (LT_complex). Full-length element and ERV structural category counts and locations can be found in table S12. These full-length sequences were subsequently cross-referenced with PRO-seq data, with the exception of the LT_complex category to determine transcriptional activity.

All classes of TEs (excluding DNA transposons) were grouped into relative age groups based on divergence and phylogenetic distribution (*6*, *88*, *94*–*99*), according to table S13. LINEs, SINEs and retroposons were grouped by subfamily, while LTRs were grouped by family.

3. TE-based Precision Run-on Sequencing (PRO-seq) analyses

Cell permeabilization for CHM13 and RPE-1

For each replicate, adherent cells were washed 2x in cold 1x PBS before adding 5mL of buffer P (10mM Tris-Cl pH 8.0, 10mM KCl, 250mM sucrose, 5mM $MgCl_2$, 1mM EDTA, 0.05% Tween-20, 0.5mM DTT, 10% Glycerol). Cells were scraped, collected and 10uL was removed for cell counting and the remainder was centrifuged at 1000xg for 5 min. 1mL of buffer W (10mM Tris-Cl pH 8.0, 10mM KCl, 250mM sucrose, 5mM MgCl2, 1mM EDTA, 0.5mM DTT, 10% glycerol) was used to gently resuspend cell pellets, before adding an additional 9mL of buffer P, inverting and centrifugation at 1000xg for 5 min. 500uL of buffer F (50mM Tris-CL pH 8.0, 40% glycerol, 5mM $MgCl_2$, 0.1 mM EDTA, 0.5mM DTT) plus 0.5uL of RNase-inhibitor (SuperAse from Ambion) was used to resuspend the cell pellets, followed by another 500uL to wash the tubes; both were pooled together (1mL total), transferred to a 1.5mL tube, and centrifuged at 1000xg for 5 min. Finally, permeabilized cells were resuspended in 57µL of buffer F with 1µL of RNase-inhibitor added before snap-freezing in liquid nitrogen and storage at -80°C.

Illumina Library Preparation for CHM13 (and RPE-1, see Note S4)

PRO-seq libraries (from replicates) were prepared as previously described (*22*) with minor modifications. Approximately 2 x 10^6 permeabilized cells were mixed with permeabilized *Drosophila* S2 nuclei in all 4-biotin-NTP run-ons (5 x 10^4 *Dm* nuclei in each replicate), and run-on RNA was extracted with Norgen columns and eluted in 50uL H20. Base-hydrolysis included incubating in 25µL cold 1N NaOH for 10min on ice, followed by the addition of 125uL cold 1M TrisCl pH 6.8, a gentle vortex, and brief spin down before enrichment with streptavidin-beads. Following 3'-ligation and the second bead binding, both end-repair reactions and the 5'-ligation were all performed with nascent RNAs still bound to the beads (*100*). These on-bead reactions were performed in a total volume of 20µL with constant rotation before elution from the beads and the subsequent reverse transcription and PCR steps. Test amplifications were performed on 5% of the library and samples were amplified to the ideal number of cycles for final preparation. Following final amplification, libraries were PAGE-purified to remove adapter-dimers and select molecules between 140-650bp in size. Libraries were then sequenced on an Illumina NextSeq 550, producing single-end 75bp reads.

Pre-processing and mapping of CHM13 (and RPE-1, see Note S4) PRO-seq data

Raw fastq files were first quality trimmed (Phred score >=20) and adapter sequences removed using cutadapt (*101*). Reads below 20nt were removed and remaining reads were reverse complemented using the fastx-toolkit (*102*). *Drosophila* spike-in reads were removed by aligning to the Dm6 genome with bowtie2 (*103*) using "--very-sensitive" options. Remaining reads were then aligned to T2T-CHM13 with bowtie2 (*103*) using "-k-100" and default (end-to-end alignment, "best match") options. Sorted bam files were converted to bed files with BEDtools (v2.29.0) (*89*), which were subjected to one or more of the following: 1) single copy k-mer filtering, 2) normalization with non-mitochondrial alignments to obtain counts in Reads-per-million Mapped (RPMM) wherein the raw counts N were normalized using the following equation: RPMM=[N/million_non-mito_alignments]], 3) conversion into BigWig files (BEDtools (v2.29.0) (*89*), GenomeBrowser/20180626) for data visualization.

Heatmaps and composite profiles

Count matrices for heatmaps and composite profiles were generated using deepTools2 (*106*) using Bowtie2 default (end-to-end alignment) “best match” data (Fig. 2, 3 C-D, fig. S18, 21, 25 B, C). Repeat element groups with a large number of regions were randomly subset to a maximum of 5,000 regions. Scaling and anchoring elements for heatmap and composite profiles is as follows. Scaling and anchoring elements for heatmap and composite profiles is as follows. SST1 elements and HERV-Ks (all structural groups) were scaled. Scaled data in bigwig format was binned using 10bp windows and repeat elements were scaled to an equal size of 1Kbp with the flanking 100bp included in the matrices. All other repeat elements were anchored to the 3’ end (as the truncated elements are most likely to be 5’ truncated), sorted by region length, and shown a certain distance into (or towards the 5’ end of) the element based on the expected full-length of each respective element: *Alu* (0.5kb), L1 (7kb), SVA (4kb). Bins summarized the underlying bigwig data by taking the maximum value and composite profiles were created by averaging each bin across all regions in the group. Standard error is calculated for each group of elements and is shown as grey shading around the composite profiles.

For method comparison, Bowtie2 k-100, Bowtie2 k-100 filtered for single copy 21-mers by map location and Bowtie2 k-100 filtered for single copy 21-mers by map location and read content were implemented as below (fig S18, S21, S25 B, C). Highlighted within each comparison is the Bowtie2 default “best match” data. Included for each comparison are heatmaps of single copy 21-mer k-mers for each element.

For method comparison, Bowtie2 k-100, Bowtie2 k-100 filtered for single copy 21-mers by map location and Bowtie2 k-100 filtered for single copy 21-mers by map location and read content were implemented as below (fig S18, S21, S25 B, C). Highlighted within each comparison is the Bowtie2 default “best match” data. Included for each comparison are heatmaps of single copy 21-mer k-mers for each element. The single copy 21-mer plots illustrate the regions of each TE that lack sequence specificity and are therefore, most prone to read loss through either k-mer filtering method (BT2 k-100 with single copy k-mer filtering, BT2 k-100 dual k-mer filtering).

Methylation heatmaps for HERV-K were generated in R ggplot2 by normalizing repeat size by start and end position and using geom_tile() to plot CpG methylation frequency at each position (*21*). For all other elements methylation heatmaps were made by aligning the repeat elements at the 3’ and using geom_tile() to plot CpG methylation frequency at each position (*21*) .

Statistical analyses and data visualization - repeat models

BEDtools (v2.29.0) (*89*) map was used to calculate average methylation (*-o mean*) and CpG density (*-o count*) across all repeats in RepeatMaskerV2 (RMv2) and incorporated into the 3D graphs and parallel plots, made using JMP® (*142*). This method was also used to calculate average methylation for SST1.

Genomic data was visualized for presentation using RIdeogram (v0.2.2) (*104*) and Circos (v0.69-6) (*105*). Circa (v1.2.2) was used to generate segmental duplication ribbon plots (https://omgenomics.com/circa/). JMP® (*142*) was used to make 3D graphs and parallel plots (note: *Alu* and L1 (except L1Hs) were subsampled to 2% for these plots due to high copy number). Genome browser tracks and CenSat annotations for T2T-CHM13 are as described in (*11*, *12*, *21*, *65*). Microsoft Power BI Pro (version 2.98.683.0) was used to create ribbon plots.

SST1 Phylogenetic Analyses

T2T-CHM13 annotations for SST1 were combined with known repeat locations and extracted from the genome as sequences via bedtools (*89*). The resulting elements were annotated with chromosome, coordinates, full-length, intersection of centromere, telomere or interstitial chromosomal locations, and average methylation of the element (tables S14-17). Sequences (table S15) were aligned with MAFFT (*107*). The evolutionary history was inferred by using the RAxML-NG method (*108*) and the GTR+G (general time reversible model with a gamma distribution of rate variation among sites) model (*109*) as matched by jModelTest (*110*). The consensus tree shown in Fig. 3A was generated from the resulting 100 bootstrap replicates.

SST1 PRO-seq data analyses

SST1 PRO-seq overlap repeat grouping cutoffs (repeats with ≥ 15) were determined by plotting the distribution of read overlaps across all SST1 repeats (fig. S23, table S14 and table S16). An unpaired *t* test (table S17) was performed to quantify the statistical significance of differences among SST1 repeats with high v. low read overlap by repeat length, percent divergence, percent insertions, and percent deletions as identified by RepeatMasker and average methylation as determined by (*21*) as described below (*18*).

BEDtools (v2.29.0) (*89*) was used to intersect SST1 and L1Hs repeats with genomic locations (including centromere satellite annotations (*12*)), methylation (*21*), and transcriptional data (*18*); these data were used to generate repeat groupings (e.g., overlapping a specified satellite annotation; <0.5 average methylation/ ≥0.5 average methylation, etc.). Percent divergence from the consensus for repeats was taken from the RepeatMasker output. Violin plots were generated via GraphPad Prism software (v9.1.1).

4. Centromere Transcription (PRO-seq and RNA-seq) Analyses

TE Embeds within cenSAT annotations

As per the RMv2-alpha file generation above, the RepeatMasker AnnotationV2 was intersected (BEDtools v2.29.0) with all cenSAT annotations to identify and label those repeats overlapping any of the major satellite groups (e.g., alpha, beta, HSAT). A minimum of 1bp overlap was used to assess whether a TE was embedded and/or at the edge of one of these satellite regions.

Mitotic Synchronization and Release for HeLa time course

Given the low rate of cell division and synchronization challenges presented by CHM13 cells, HeLa-S3 cells were used as a proxy, noting the caveat that this cell line carries high levels of karyotypic instability (*111*). HeLa-S3 cells at 25-30% confluency were treated with 2mM thymidine for 24 hours, released in fresh medium for 3 hours, then treated with 100ng/mL nocodazole for 12 hours (*112*). Mitotic cells were collected by shake-off, centrifuged and washed in 1x PBS, and then either grown on 15cm dishes in fresh medium for the corresponding time or immediately permeabilized (mitotic sample).

Cell Cycle Analysis for HeLa time course

Prior to cellular permeabilization, 10% of each sample was removed and fixed in 75% cold (-20°C) ethanol. Cells were then stained with propidium iodide and DNA content was analyzed using a BD FACS Aria II. FCS files were read into R for downstream analyses using the flowCore package (Fig. 4A, left). Separately, mitotic HeLa cells were also stained with DAPI and manually analyzed by microscopy in order to differentiate cells in G2 from those properly arrested in prometaphase by level of DNA condensation.

Cell Permeabilization for HeLa time course

For each replicate time point, both floating cells in the growth medium and cells removed by scraping in 1x PBS were collected, pooled, and centrifuged at 1000xg for 5 min. Cells were resuspended in 1x PBS and 10% was removed for FACS analysis before completing the wash. 1mL of buffer P (10mM Tris-Cl pH 8.0, 10mM KCl, 250mM sucrose, 5mM $MgCl_2$, 1mM EDTA, 0.05% Tween-20, 0.5mM DTT, 10% Glycerol) was used to gently resuspend cell pellets, before adding an additional 9mL of buffer P and incubating on a shaker for 5min. Permeabilization was assessed with trypan blue and samples not initially permeabilized were again incubated on a shaker with buffer P containing 0.05% NP-40 for 5min. Permeabilized cells were then centrifuged at 1000xg for 5 min before resuspension in 1mL buffer F (50mM Tris-CL pH 8.0, 40% glycerol, 5mM $MgCl_2$, 0.1 mM EDTA, 0.5mM DTT), transferred to a 1.5mL tube, and centrifuged at 1000xg for 5 min. Finally, permeabilized cells were resuspended in 55µL of buffer F and 1µL of RNase-inhibitor was added before snap-freezing in liquid nitrogen and storage at -80°C.

Library Preparation for HeLa time course

PRO-seq libraries were prepared as previously described (*22*) with minor modifications. 0.9-4.5 x 10^6 permeabilized cells were mixed with permeabilized *Drosophila* S2 nuclei in all 4-biotin-NTP run-ons (1 x 10^6 *Dm* nuclei in each A replicate and 5 x 10^4 in each B replicate). The rest of the library preparation is the same as in CHM13/RPE PROseq above and libraries were also sequenced on an Illumina NextSeq 550, producing single-end 75bp reads.

Pre-processing and mapping HeLa time course data

All data was pre-processed, mapped and post-processed the same way as CHM13 and RPE-1 (see below) with the following few exceptions: 1) alignments to the *D. melanogaster* genome included the "-k 1" option, 2) alignments to CHM13v1.0 included the "--very-sensitive" option and were only done with "-k 100", and 3) normalization was done using a combination of *D. melanogaster* spike-ins (to most accurately compare transcription levels across timepoints), and non-mitochondrial alignments (to uniformly rescale the counts across timepoints so as to obtain final values on a Read-Per-Million-Mapped scale comparable to that of the CHM13 data). Starting from the raw number of reads overlapping a given repeat N, the following equation was used to obtain the normalized counts:
*[N/Dmel_norm factor]/(median_across_timepoints[*million_non-mito_ alignments*/Dmel_norm_factor]*)*]*

H9 ChRO-seq data availability and pre-processing

External ChRO-seq data (GSE142316) for four different developmental stages (ES, DE, duodenum, ileum) of H9 cells was used for comparison to the CHM13 cell expression data (GEO GSE142316) initially reported in (*113*). H9 ChRO-seq data was pre-processed using the proseq2.0 pipeline as laid out in (*143*).The script proseq2.0.bsh was used with parameters -SE -G --UMI1=6 --UMI2=6 --Force_deduplicate=FALSE.This script generated adapter-trimmed and deduplicated fastq files which were used as input to Bowtie2 and CASK for repeat composition analysis.

Pre-processing, mapping and post-processing of RNA-seq data (CHM13 and HG002)

Data from CHM13 paired-end native RNA-seq using oligoDT (*12*) was processed with the same workflow as the CHM13 PRO-seq data, with the following modifications: reads below 100nt were removed, no reverse complement was required (as this is PRO-seq specific), *Drosophila* spike-ins were not included and therefore, did not need to be removed, and properly paired reads were filtered for with the SAM flag F1548. For the CASK analysis, only mate1 of each replicate was used. External paired-end RNA-seq data (ribodepleted) for HG002 (GM24385) was used for comparison to the CHM13 cell expression data on Chromosome X (SRA: SRR13086640). HG002 data was pre-processed with the same workflow as the CHM13 RNA-seq, but then mapped to a combined assembly of T2T-CHM13 autosomes, HG002 chrX, and GRCh38 chrY with Bowtie2 using the default option and normalized with non-mitochondrial reads.

Repeat transcript quantification approaches

As a complement to the comprehensive TE (herein) and centromere satellite repeat annotations (*12*), we implemented a three-pronged approach to defining the transcriptional landscape of CHM13 centromeres (fig. S27). In a *mapping-dependent* approach compared to Bowtie2 default "best match, we mapped PRO-seq and RNA-seq data using Bowtie2 k-100 and intersected reads with single copy k-mers based on the T2T-CHM13 assembly and whole genome shotgun (WGS) PCR-free reads (*11*, *114*). As a complement, we implemented an original approach, *mapping*-i*ndependent* sequence classification (CASK, fig. S28), which utilized repeat annotations from CHM13 to form a database of k-mers capable of discerning specific repeat types or a refined group of repeats (i.e., ambivalence group). Unmapped PRO-seq and RNA-seq reads were annotated using CASK and the CHM13-dependent k-mer database. Finally, in a *genome-independent* approach, PRO-seq and RNA-seq reads were processed through RepeatMasker using the human Dfam 3.3 library (i.e., not specific to T2T-CHM13) (fig. S27). Simultaneously, RepeatMaskerV2 (RMv2) was intersected (BEDtools v2.29.0) with cenSAT annotations for alpha-satellite only to identify and label those repeats overlapping alpha satellite designated HOR, dHOR, MON, and "none of the above" regions (requiring a minimum of 1bp overlap). This dataset was defined as the alpha-satellite specific RepeatMaskerV2 annotations (RMv2-alpha).

Mapping dependent PRO-seq analyses

For each PRO-seq dataset, Bowtie2 default "best match" reports a single alignment for each read, thus providing *locus level* transcriptional profiles. Unfiltered Bowtie2 k-100 mapped PRO-seq (two independent libraries) and RNA-seq data (two independent libraries) reports up to 100 mapped loci for each read, thus providing over-fitted transcriptional profiles (fig. S27 and figs. S29-S31, S34, S36-S37). The benefit of having a complete, high-quality long-read assembly such as T2T-CHM13, allows for the generation of genome-wide single copy k-mers spanning even the most repetitive regions of the genome. These single copy k-mers (generated through Meryl (*144*)) were based on the T2T-CHM13 assembly itself and whole genome shotgun (WGS) PCR-free reads (*11*) for increased confidence as single copy. Multiple tiers of filtering were applied to the Bowtie2 k-100 mapped PRO-seq (two independent libraries) and RNA-seq data (two independent libraries) using these single copy k-mers as follows (fig. S27 and figs. S29-S31).

The first tier of filtering involved bed files of the mapped reads filtered through these single copy k-mers using overlapSelect with the setting “-overlapBases=XXbp”, where XX is equal to the length of the single copy k-mer being overlapped (21bp or 51bp). This required that a minimum of the entire length of the single copy k-mer must overlap a given read in order for that read to be retained and provided a lower bound *locus-level filtering*.

In parallel, each alignment was filtered with single copy k-mers to ensure only one alignment survives using the code from (*114*) and as performed for the long-read marker assisted applied therein, providing *read-level filtering*. In brief, 21-mers were collected from the Illumina PCR-free WGS reads to build a k-mer database with Meryl (*144*). We chose k=21 following (*145*) to allow a maximum k-mer collision rate of 0.005, which is close to the Illumina sequencing error rate, from the given 3.2Gb genome size. Subsequently, the k-mer database was filtered by frequency greater than 42 and less than 133 to obtain globally single-copy k-mers in the genome. These k-mers were intersected with single copy k-mers in the assembly to build the marker set, to ensure the markers are globally unique in the genome and found only once in the assembly. The markers are looked up in the aligned read sequences, and only one alignment with the most markers gets chosen. If the number of markers ties among multiple alignments, only one with the best alignment score gets chosen.

In a third tier of filtering, both *read-level* and *locus-level* filtering are applied. Following read filtering, the filtered bam is converted to a bed file and filtered through the single copy k-mers using overlapSelect using the single copy 21-mers to avoid overfiltering. This dual filtering represents the strictest filter and results in the removal of any read that was retained due to a bp difference in the read itself, thus providing read- and locus-level lower-bounds for mapped reads (fig. S27).

The resulting bed files from each of the mapping methods (k-100, default, and k-mer filtered) were used for counting reads overlapping repeats or alpha-satellites and for bigwig generation for visualization, as described above.

Mapping independent analyses: CASK (Classification of Ambivalent Sequences using k-mers)

We used CASK (Classification of Ambivalent Sequences using K-mers), a mapping-independent method to identify reads originating from repeat elements using their k-mer composition (fig. S28A). Briefly, the genomic location of all repeats and their type annotation (e.g., L1H, L1P, *Alu*Y, etc.) were extracted from the T2T-CHM13 RepeatMaskerv2 annotations (RMv2, fig. S1). For each repeat type, the genomic sequences of all the instances of that repeat were extracted into a type-specific fasta file. These fasta files were input into KMC (*146*) to generate for each repeat type a type-specific k-mer database, consisting of all the k-mers (k=25) found across all instances of that repeat. Each type-specific k-mer database was then filtered to remove k-mers also present in parts of the genome that did not overlap with any repeat elements (these k-mers have limited usefulness for the purpose of identifying reads originating from repeats and could lead to false identification of repeats). Note that while each k-mer is represented only once in any given type-specific database, many k-mers are not single copy genome-wide and may be found multiple times within the same or different instances of the repeat. Additionally, many k-mers are not exclusive to a given repeat type and may be shared across different repeat types and the corresponding type-specific k-mer databases (e.g., a k-mer found in L1Hs may also be found in L1P, etc.). For each k-mer, we defined its ambivalence group as the set of all repeat types within which this k-mer was found. Using a custom pipeline, the type-specific k-mer databases were combined into a single "annotated k-mer database" listing all the k-mers found across repeats and their corresponding ambivalence group.

Starting from the trimmed and deduplicated fastqs (PRO-seq, RNA-seq, ChRO-seq), sequencing reads containing one or more k-mers matching a k-mer in the annotated k-mer database were extracted using BBduk. For each read, we then computed the intersection of the ambivalence groups of all the matching k-mers within the read. This intersection can be interpreted as a consensus repeat assignment from all of the k-mers in the read. If the intersection contained a single repeat-type (e.g., L1P), the read was assigned to that repeat. If the intersection contained multiple repeat types, the read was annotated as ambivalent, and the possible set of repeat types for this read was recorded (e.g., an ambivalent read could receive an assignment {L1H or L1P}). Such ambivalent reads were used to compute *upper bounds* on the number of reads originating from a given repeat type (e.g., fig. S28). CASK data shown without error bars ignore reads with ambivalent assignments and thus represent *lower bound* estimates of the repeat expression. Finally, although this scenario was rare (fig. S28), if the intersection was empty (e.g., if one k-mer in the read was specific to L1P and another k-mer was specific to L1Hs), the read was annotated as containing "conflicting k-mers" and discarded from further analysis.

Mapping independent analyses: RepeatMasking of PROseq and RNAseq reads

RepeatMasker (v4.1.2-p1) was run on the trimmed reads of the individual replicates of the CHM13 PRO-seq and RNA-seq datasets (fig. S30), as well as all other PROseq datasets included in this study (RPE, H9, and HeLa, fig. S32) using a library consisting of the Dfam 3.3 database plus the additional entries discovered as part of the TE analysis of the T2T-CHM13 genome assembly (RMv2, fig. S1). The resulting RepeatMasker output files were then summarized using RM_summarizer.pl (*43*)(perl v5.30.1) to obtain the number of reads containing each repeat type. For the paired-end RNAseq datasets, mate1 and mate2 were run individually and the counts were summed. For both RNA-seq and PRO-seq, the relative abundance of each repeat was similar across replicates, and thus counts from both replicates were summed (figs. S30B and S32). A small percentage of reads that were retained had >1 repeat designation across the length of the read (ranging from 0.21-1.19% of all PRO-seq reads, 1.32% of all RNA-seq reads) (and see note below: *Approaches to avoid mappability artifacts and Interpretation of transcription level estimates*).

Detecting Repeat Transcription - Method Comparisons

BEDtools (*89*) coverage was used to obtain counts of reads overlapping repeats defined in RMv2 across all mapping methods (see above), requiring at least 50% of the read (using "-counts -F 0.5") to overlap the repeat element. This method was also used to determine how many *repeats* had reads overlapping (fig. S29-S31, table S18). Alpha-satellite specific RepeatMaskerV2 annotations (RMv2-alpha) were used to obtain counts of reads that overlap repeats in the same manner as above (BEDtools coverage -counts -F 0.5). Since 50% of the length of a pre-processed PRO-seq read is ~25-30bp, and the CASK k-mer length is 25bp, these parameters are roughly equivalent for this comparative analysis. The relative abundance of each repeat was similar across replicates, and thus counts from both replicates were summed.

To compare all approaches, we first used different Bowtie2 mapping parameters (default "best match", k-100 and k-100 filtered for single copy k-mers in T2T-CHM13 using locus-level, read-level and dual locus- and read-level filtering) on PRO-seq and RNA-seq datasets (fig. S29). For PRO-seq datasets, default parameters that report only the best mapped location ("Bowtie2 default") and mapping parameters that support multi-mappers combined with intersecting those reads to only report those that overlap with a single copy 51mer ("Bowtie2 k-100 locus-filt 51mer"), show largely concordant repeat calls. In contrast, RNA-seq datasets show largely concordant repeat calls for the "Bowtie2 k-100", "Bowtie2 k-100 locus-filt 51mer", and" Bowtie2 k-100 locus-filt 21mer" with a larger difference in calls compared to the Bowtie2 default parameters. Thus, different mapping approaches impact data interpretation for different types of transcription datasets. Applying a stricter level of filtering (read-level filtering) to either PRO-seq or RNA-seq dataset results in a reduction of the overall repeats transcribed (fig. S29) and in the raw read counts (fig. S31). Overall, Bowtie2 default parameters report fewer repeat calls than Bowtie2 k-100 locus-filt (51mer or 21mer), but still more than Bowtie2 read-level filtered or even read- and locus-level dual filtered (a combination of the two filtration methods, and thus the most stringent of all mapping methods).

When comparing across all repeat transcript annotation methods for T2T-CHM13 (fig. S27), we find that CASK, RM, Bowtie2 default, and Bowtie2 k-100 21-mer filtration annotations for PRO-seq data were largely concordant in the relative abundance of each repeats class, while repeat annotations for RNA-seq data from T2T-CHM13 were more variable across methods (fig. S30).

Given that PRO-seq captures nascent transcription while RNA-seq cannot distinguish newly synthesized transcripts from stable and accumulating transcripts, combined with variable repeat calls across methods, indicates that PRO-seq provides a more robust representation of active transcription. Of note, while SINEs were the predominantly transcribed repeat across all datasets irrespective of the method employed (Bowtie2, CASK, RM), allowing multi-mappers Bowtie2 k-100 or Bowtie2 k-100 single copy k-mer locus-filtered, resulted in SINE read counts increasing, likely due to their high abundance in the human genome. Likewise, removal of multi-mappers and reads without single copy k-mers reduced the SINE read counts even more, further supporting this rationale. In contrast to SINEs, satellites were among the lowest transcribed repeats regardless of method employed (Bowtie2, CASK, RM), even when over-fitted with Bowtie2 k-100 (figs. S34 and S36).

Approaches to avoid mappability artifacts and interpretation of transcription level estimates across centromeres

Transcription over genes and other non-repetitive elements of the genome is typically analyzed through sequence alignment, followed by removal of multimappers, and then quantified with estimators such as FPM (Fragment per Million) or TPM (Transcript per Million).
This approach of excluding multi-mappers altogether is highly inaccurate when examining signals over repeat elements as it results in an underrepresentation of the repeats with high copy number and low sequence divergence. Alternatively, one may allow multi-mapping reads, and then choose to either keep all alignment candidates or pick one at random, but this results in, respectively, overreporting or inaccurately reporting the transcription levels.

To acknowledge these challenges and quantify transcription from repeat elements in a way that does not suffer from these biases, we combined different quantification approaches guided by the following two principles:

1) While it may be impossible to exactly map a read coming from highly repetitive regions, one might still be able to determine which repeat family the read originates from. By quantifying transcription at the class-level (aggregating all loci from the same repeat class), rather than at the level of individual loci, one loses spatial resolution but gains specificity and accuracy in the call.

2) While it may be impossible to obtain an accurate expression level for a specific instance of a repeat element, one can obtain lower and upper bounds for the expression level at that element. Upper-bounds can then be used to demonstrate that the lack of signal over a particular element is real, and not an artifact from low mappability. Likewise lower-bounds can be used to demonstrate that the signal at a particular repeat is real, and not an artifact from misalignment.

This section provides information on each of the methods we used, including how their quantitative output should be interpreted and their possible caveats.

1) **CASK**

- What quantitative output does it provide? **Lower and upper bounds of transcription, coarse-grained at the level of repeat classes or repeat types.**
- How does it work? We developed CASK specifically to circumvent mappability-related issues by detecting reads from specific repeat classes using their k-mers signature, rather than through alignment to the genome (see algorithm details in fig S28). CASK is a mapping-independent method. Intuitively, CASK scans each read for the presence of k-mers matching those present in the various repeat classes. Importantly, CASK searches for k-mers in the full genome assembly rather than in consensus sequences for each repeat class. Thus, because we have a complete telomere-to-telomere assembly, this guarantees CASK cannot miss reads (aside from rare edge cases discussed below). Importantly, CASK does not rely only on single copy k-mers (those appearing a single time in the genome), but rather uses all the k-mers found across all instances of each repeat class across the genome (including those with high copy number in the genome). Thus, even reads originating from regions with low mappability will be assessed.
- Potential reads missed or misannotated by CASK? Let's consider a read coming from a repeat with very high genome representation at exact or near exact match. To simplify this discussion, let's imagine the true repeat of origin is a HOR, but the same reasoning can be made with other repeats. One can be concerned with the following scenario:

- Scenario 1) the read contains a sequencing error within a HOR k-mer which pushes it outside of the HOR k-mers database. While this can occur, the error rate with short-reads sequencing is low (1/1000 bp for an average Q30). Additionally because the PRO-seq reads are ~75bp in length, a single bp substitution error will affect ~1/3 of the k-mers in the read (we use k-mers of length 25). Thus in order for a HOR read to be missed due to sequencing errors, it would need at least 3 sequencing errors, the probability of which is vanishingly small.
- Scenario 2) None of the k-mers in the read are specific to HORs. For example, consider a hypothetical situation where all the k-mers in the reads are present in the HOR k-mer database, but also in the L1 k-mer database. As shown in the example Read #3 of fig. S28C, CASK can still annotate these reads as "HOR or L1", to indicate that we are not sure whether this is a HOR or L1 read. We call this situation an "ambivalent assignment". One can then choose to either discard those ambivalent assignments, or include them in the final tally for both HOR and L1 counts. By choosing the latter, we can obtain an upper bound on the number of reads from each repeat family. These upper bounds are displayed in fig. S31. Note that the upper bound is typically very close to the base count (which excludes ambivalent assignments), so this scenario turns out to be relatively rare.
- Scenario 3) The read spans across 2 repeat classes or contains 2 or more k-mers that are never found in the same repeat family (which would happen when the read spans across 2 repeat classes or contains specific errors). This occurs only in a small percentage of reads (ranging between 1.7% and 2% of all reads containing k-mers from repeat families in our RPE-1 PRO-seq dataset, or between 0.6 and 0.8% of all sequenced reads, and is thus unlikely to affect the overall composition of the transcribed repeats
- Scenario 4) The read comes from a HOR which is missing in the genome assembly (in which case some k-mers might be missing in the HOR k-mer database). We cannot rule out this scenario for the HeLa, RPE-1, and H9 datasets and for the embryonic cells development time course, but we are guaranteed that it does not occur in CHM13 for which we have a T2T-level, gap-free assembly.

2) **Repeatmasker**

- <u>What quantitative output does it provide?</u> **Estimates of transcription, coarse-grained at the level of repeat classes, families, or subfamilies.**

- How does it work? Repeatmasker uses consensus sequences for each repeat subfamily, rather than the whole genome sequence. We thus used Repeat Masker as an orthogonal, assembly independent alternative to CASK. Both methods concord in their estimates of transcription at the level of repeat classes. In particular, both CASK and Repeatmasker reveal low levels of transcription from satellites.
- Potential reads missed or misannotated by Repeatmasker? Since Repeatmasker compares the consensus sequences of specific repeat subfamilies to each read and identifies matches, there is the possibility that a single read could be masked as >1 repeat subfamily. This is rare (stated in the supp. methods as <1.4% of reads), but does occur and therefore, this could result in a slight overestimation of repeat counts.

3) **Bowtie2 with near-exhaustive search**

- What quantitative output does it provide? **locus-level, upper-bounds of transcription**.
- How does it work? Reads are aligned to the T2T-CHM13 assembly with Bowtie2 allowing up to 100 alignments per read (-k-100). All valid alignments are kept, yielding a list of candidate loci for each transcript. This method is not meant to estimate true transcription levels at individual loci, because it inflates read counts for loci with high copy number and low sequence divergence. However, for most repeat classes, the number of valid Bowtie2 alignments is less than 100. Thus for these repeat classes, the levels reported by this method are, by definition, upper-bounds of the true transcription levels at specific loci. Indeed, if a read comes from locus L, one out of all the valid alignments must map to L.
- Potential reads missed or misannotated? Reads from highly abundant repeats, such as *Alu*s, may have over 100 valid alignments, yet only up to 100 alignments are retained due to computational complexity limits.

4) **Bowtie2 with near-exhaustive search and single copy k-mer locus filtering.**

- What quantitative output does it provide? **locus-level, lower-bounds of transcription**.
- How does it work? K-mers that are single copy in PCR-free WGS reads and single copy in the T2T-CHM13 assembly are used to filter reads following the mapping method in 3). This filtering is based on *location* and requires a given read alignment to overlap an entire single copy k-mer in the assembly in order to be retained. With this method, a larger single copy k-mer is ideal (e.g. 51bp) since they are more abundant in the genome than a smaller k-mer (e.g. 21bp) so more reads are retained.
- Potential reads missed or misannotated? There is a possibility that a given read with >1 alignment is retained at both loci as long as both loci are considered distinct. For example: this can occur if a given read comes from a paralog with a SNP, it can still

align and be retained at both loci since this method does not require the read to have a single copy k-mer itself. Therefore, while this method is considered to be a low-bounds of transcription, it still runs the risk of having inflated transcription levels.

5) Bowtie2 with near-exhaustive search and single copy k-mer read filtering.

- What quantitative output does it provide? **read-level, lower-bounds of transcription**.
- How does it work? The same k-mers that are used in 4) are used here, except the application is different. This filtering method is based on the presence of single copy k-mers in the reads themselves. When a near-exhaustive search of alignments is used and a given read has >1 alignment, the alignment with the most single copy k-mers is retained. In the case of a tie, the first alignment is selected. This ensures that there is only 1 alignment per read. With this method, a smaller single copy k-mer is ideal (e.g. 21bp) as a larger single copy k-mer (e.g. 51bp) would not tolerate any SNP or error base and would result in over-filtering the reads.
- Potential reads missed or misannotated? While this method is stringent and is considered to represent a lower-bounds of transcription compared to 4), there is still the possibility that some reads passed through this filtering step when they shouldn't have. This could occur if a read has a basepair difference (one explanation being transcriptional slippage) compared to the original sequence it was transcribed from. As a result of this difference, this read could now contain an single copy k-mer during this filtering process and be retained, when it should have been discarded.

6) Bowtie2 with near-exhaustive search and single copy k-mer read filtering followed by single copy k-mer locus filtering.

- What quantitative output does it provide? **Read and locus-level, lower-bounds of transcription**.
- How does it work? This method takes the filtered reads from 5) and applies the locus-based filtering method from 4) to provide the most stringent filtering method in this study, and therefore, the lowest-bounds of transcription.
- Potential reads missed or misannotated? While the filtering method in 5) is highly effective, this combination of single copy k-mer filtering methods, read and locus-based, results in the removal of any reads that passed through 5) due to a basepair difference when they shouldn't have because of an actual lack of single copy k-mers. This is because when these reads are then required to overlap a single copy k-mer locus in the assembly, it won't match the original sequence it was transcribed from and will not be retained.

7) **Bowtie2 with default, "Best match" parameters.**

- What quantitative output does it provide? **Locus level wherein each read is mapped to only a single, best match location.**
- How does it work? This method (no flag) reports a single, “best match” end-to-end alignment for each read, but if a read could map with 100% accuracy to two locations, it assigns a single location randomly.
- Potential reads missed or misannotated? By selecting only a single match location, the likelihood of diluting signal among sequences of high identity (i.e. repeats, particularly young repeats) by random assignment (i.e. promoters of different elements within the same class) increases.

5. WaluSat Phylogenetic Analyses

WaluSat+AluSx

WaluSat: The evolutionary history was inferred by using the Maximum Likelihood method and General Time Reversible model (*109*). The tree with the highest log likelihood (-2227.42) is shown in Fig. 5A. Initial tree(s) for the heuristic search were obtained automatically by applying the Maximum Parsimony method. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 2.0168)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This analysis involved 1057 nucleotide sequences. There were a total of 73 positions in the final dataset. Evolutionary analyses were conducted in MEGA X (*137*).

AluSx (with WaluSat): The evolutionary history of *Alu*Sx-like elements was inferred by using the Maximum Likelihood method and T93 model. The tree with the highest log likelihood (-2155.25) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial trees for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (+G, parameter = 0.8559). This analysis involved 71 nucleotide sequences. (partial deletion option). There were a total of 289 positions in the final dataset. Evolutionary analyses were conducted in MEGA X (*137*).

Dotplots were generated by comparison of 1.5 kb sequences flanking both 5’end and 3’ end regions of WaluSat insertions with FlexiDot (*147*) with the following parameters: -i tgut_sat_dim.fasta -p 2 -D y -f 0 -t y -k 10 -w y -r y -x y -y 0 -o PREFIX.

TSDs were identified using self-alignments implemented in Geneious and manual curation.

G-quadraplex (G4) analysis

G-quadruplex analysis was done with the GUI version of G4Hunter (*115*). *In silico* tandem array of the 64 nt WaluSat sequence was constructed in fasta. Coordinates for the Chromosome 14 WaluSat array are indicated in Fig. 5A.

6. TE transduction Analyses

Validating and categorizing putative TE transductions

DNA transduction events were analyzed using the modified TSDfinder tool (*67*) and full-length elements for L1 and SVA. In order to identify 3’ transductions mediated by L1 elements, we ran the TSDfinder with the default parameters (FIVE_PR_FLANK = 100, THREE_PR_FLANK = 3000) and the LINE/L1 families present in the RepeatMaskerv2 output developed in this study. To identify SVA-driven 5’ and 3’ transductions, TSDfinder was run using SVA elements with the following parameters: FIVE_PR_FLANK = 3000, THREE_PR_FLANK = 3000. Since TSDfinder originally was designed to find only 3’ transductions, we used the following commands to identify 5’ transductions based on the assumption there must be a 20 nucleotide distance between the 5' TSD and the start of the SVA:

```
paste <(awk -F "\t" '{OFS="\t"; print $1,$2,$3,$5,$8,$10,$11}' PRE_INSERTION_LOCUSChr* | sort -V
-k1,1) <(grep -w -f <(awk -F "\t" '{OFS="\t"; print $1,$2,$3,$5,$8,$10,$11}'
PRE_INSERTION_LOCUSChr* | sort -V -k1,1 | cut -f1) <(cat MERGEChr* | sort -V -k1,1)) | cut -
f1,2,4,5,6,7,10,16 | awk '{if ($3 ~ /\[/) print length($3)-4 "\t" $0; else print length($3) "\t" $0}' | awk '{if
($8 ~ /\(/ ) print "-" "\t" $0; else print "+" "\t" $0}'  > SVAsWithTSD.tsv
```

```
awk '{if ($3 == "-") print ($5-$6)-2 "\t" $0; else print $4+$6 "\t" $0}' SVAsWithTSD.tsv | awk '{if ($4 ~ /+/
&& ($8-$1) > 20) print $0; else if ($4 ~ /-/ && ($1-$9) > 20) print $0}' > SVA_5pr_transductions.tsv
```

Those events labeled as “3prTS” (3prTS stands for 3’ transduction) were subtracted and subjected to the validation process as described below. To validate and categorize each putative transduction discovered by TSDfinder, we defined a set of certain thresholds and four different confidence levels as follows (summarized in fig. S47).

Level 0: The lowest confidence level contains putative transductions discovered by running TSDfinder. In this step, only those putative transductions whose both TSD and a poly(A) tail that consisted of homopolymer stretches of A’s were removed as they likely are artifacts. This level consisted of 23,602 events.

Level 1: Here, we filtered out those transductions that were located in the segmental duplication regions of CHM13 as per (*65*). The first tier of filtering reduced the transduction events to 21,996.

Level 2: In this step, the progenitor of each transduction event is identified. Initially, we extracted 3 kbp downstream from full-length LINE/L1 (minimum length 5900 bp) and Retroposon/SVA elements (as defined in RepeatMasker annotation). If any of these 3 kbp segments were overlapping, we kept the closest one to the 3’ end of the corresponding chromosome. Next, we removed those putative transduction progenitors in the segmental duplication loci and made a “blastable" database using the following command from BLAST suite (v2.11.0+) (*148*):

```
makeblastdb -in input.fa -dbtype nucl -parse_seqids
```

In the case of SVA 5’ transductions, 3Kbp upstream from the 5’ end of all Retroposon/SVA elements were extracted to create a cognate database (if any of these 3Kbp segments were overlapping, the closest one to the 5’ end of the corresponding chromosome was kept). In this step, we also removed those 3Kbp segments overlapping with segmental duplicates. Subsequently, the sequences of putative transductions were collected and masked with RepeatMasker (v4.1.0) (*88*) (*-q -species human -xsmall*), and aligned to their related databases using BLASTN (v2.11.0+) (*148*) (*-evalue 0.05 -max_target_seqs 5 -perc_identity 90*). The results of each BLAST search were analyzed to find the progenitor of each transduction event. A progenitor was considered to be the source of a transduced sequence only if all of the following criteria were met: 1) the identity between two sequences (i.e., a query and a subject) was equal or greater than 90%, 2) hit and subject had the same orientation, 3) at least 30% of the length of putative transduction was included in the alignment, 4) the start coordinates for a pair of corresponding query and subject were within 20 nucleotides of each other. Additionally, we checked whether or not the remaining transduced segments carried another non-LTR (complete or partial) in their sequences. In many cases, the entire transduced DNA or its terminal fraction overlapped with another non-LTR element, rendering it difficult to confidently attribute a poly-A tail followed immediately by 3' TSD (one of the transduction signatures) to a real transduction event because they may occur coincidentally with the inserted sequence. Hence, we kept only those transduction events whose entire sequence was either depleted of non-LTR elements or had another non-LTR in its middle section flanked by a unique DNA sequence. For this, we used bedtools intersect command (*89*) as follows:

```
bedtools intersect -wao -a transduced_segments_passed_blast_criteria.bed -b non-LTRs.chm13.bed
-f 1.0 -s | awk '$NF == 0 {print $0}' | cut -f1-6 > TEMP.bed
```

```
awk '{if ($6 == "-") print $1 "\t" $2 "\t" $2+50 "\t" $4 "\t" $5 "\t" $6; else print $1 "\t" $3-50 "\t" $3 "\t" $4
"\t" $5 "\t" $6}’ TEMP.bed > TEMP_50bp.bed
```

```
cat <(bedtools intersect -wao -a TEMP_50bp.bed -b non-LTRs.chm13.bed -s | awk '$NF == 0 {print
$0}' | cut -f1-6)  <(bedtools intersect -wao -a passed_blast_criteria.bed -b non-LTRs.chm13.bed -s |
awk '$NF == 0 {print $0}' | cut -f1-6) > transductions.level2.bed
```

The transductions that met all criteria were classified as level 2 confidence DNA transduction events. The second tier of filtering reduced the transduction events identified in CHM13 to 129 events.

Level 3: As a final validation, we checked whether or not pairs of identified offspring-progenitor were annotated as the same family type. Accordingly, we classified a transduction event as level 3 *only* if it was transduced from a source retroelement of the same family type. The last tier of filtering dropped the transduction events identified in CHM13 to 81.

In total, 60 L1s were sources of transduced DNA, among which *L1PA2* with the size of 6,029 bp located on chromosome 2 (coordinates: 83769403-83775432) seemed to be the most prolific with three offspring, while the remaining progenitors each generated one or two offspring. In the case of SVA 3’ transductions, four elements were verified as the sources of transductions among which SVA_F with the size of 2,036 bp (chr2:47437574-47439610) was the most productive with two offspring. We found that nine SVAs appeared to be sources of 5’ transductions. One SVA locus with two offspring transduced genetic material via 5’ transduction: chr1:37640437-37642497 (SVA_F).

Functional annotation of TE transduction events

To assess the potential impact of TE transduction events on protein-coding gene evolution via exon shuffling, we compared each transduced sequence with the human proteome. To investigate whether Level 3 transductions carry protein-coding sequences, we conducted a BLASTX (*148*) analysis. First, we downloaded the human proteome (https://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/human.1.protein.faa.gz, on 15.05.2021) and created a BLAST database using the following command:

```
makeblastdb -in human_proteome.fa -dbtype prot -parse_seqids.
```

The sequences of Level 3 transductions were aligned to the human proteome database using blastx:

```
blastx -query <input.fa> -db human_proteome.fa —task blastx-fast -evalue 0.00001 -outfmt '6 qseqid
qlen sseqid slen length pident nident mismatch gaps qstart qend sstart send qseq sseq sstrand qcovs
qcovhsp qcovus evalue score’ -max_target_seqs 5 -soft_masking true
```

BLASTX analysis showed that none of the transductions in the level 3 dataset were similar to a CDS sequence, possibly due to our stringent filtering which reduced the overall transduction calls, thus reducing the likelihood of finding rare CDS overlap.

Comparative analysis of TE transductions between T2T-CHM13 and GRCh38

To identify whether any of the level 3 transductions are specific to T2T-CHM13 or are shared with GRCh38, we first extracted the sequence of each transduction plus 1Kb downstream (for 5' SVA, 1Kb upstream was extracted). Next, we mapped these sequences to the GRCh38 genome using BLAST using following commands:

```
makeblastdb -in primary_chrs_GRCh38.fa -dbtype nucl -parse_seqids
```

```
blastn -query <input>.fa -db primary_chrs_GRCh38.fa -evalue 0.05 -task megablast  -outfmt  '6 qseqid
qlen sseqid slen length pident nident mismatch gaps qstart qend sstart send qseq sseq sstrand qcovs
qcovhsp qcovus evalue score'  -max_target_seqs 5 -num_threads 10 -perc_identity 90
```

We found two L1 transductions (chr21:13527247-13527278, and chrX:75628554-75628643) and one 5' SVA transduction (chr5:151705366-151705511) specific to T2T-CHM13 (tables S22-S23) while all 3' SVA transductions were present in both genome builds. However, in one case (L1Hs) we find the offspring TE in both GRCh38 and T2T-CHM13, yet the transduced sequence is missing in GRCh38. The transduced segment consists of five copies of a hexamer (TTTTTG) which was collapsed to a single copy in GRCh38.

When comparing transduction events between T2T-CHM13 and GRCh38 (table S23), we find slightly more events in T2T-CHM13 due to gap-filled regions and high confidence annotations (*11*, *12*). Interestingly, the number of 5' transduced segments mediated by SVAs exceeds the number of SVA 3' transductions, suggesting that 5' transductions by SVAs are more common in the human genome. In summary, our results indicate TE facilitated transduction is a dynamic phenomenon that has affected 0.000175% of the CHM13 genome (0.026 events per 1 Mbp). It is worth noting that our transduction annotation is likely an underestimation of the total number of events given the high stringency thresholds employed. Nevertheless, the CHM13 assembly has afforded a multi-tier analysis that can be further applied to identify bona fide transduction events in lower confidence categories (fig. S47 and table S23).

Comparison of the identified TE transductions with previous studies

To identify overlap with previously reported transductions, we extracted the sequence of transduction events reported in (*67*, *69*–*71*, *149*) and aligned them against the T2T-CHM13 (blastn -task megablast

-evalue 1e-10) followed by comparing the blast hits with the coordinates of transductions from our results. We were able only to corroborate three of the previously reported transductions with our level 3 dataset (table S24), likely due to different methodologies and filtering schemes applied therein.

Relative age of the identified transductions

To assess the relative age of each transduction, we calculated the genetic distance between all pairs of transduction progenitors and offspring using the Kimura 2-parameter (*150*) model. For this purpose, first, we used blastn with default parameters except "-task blastn" to align the sequence of each transduction against its offspring. The alignments were analyzed using the ape package (*151*) with the function "dist.dna" and the parameter "model=k80". Next, we compared these individual distances with the distribution of the genetic distances of all the retroelements of the same subfamily present in the genome (fig. S49). Consequently, we demonstrated that virtually all retroelements with transductions fall within the range of expected divergence, strongly suggesting that the transductions were caused by retrotransposition of active TEs rather than segmental duplications.

7. Repeat comparisons between CHM13v1.0 and HG002 and among non-human primate genomes

Methylation clustering

Methylation clustering was done by selecting all reads spanning a specific locus and using the mclust (v5.4.7) R package with the "VII" model to cluster methylation calls across the locus (*92*). Within mclust we specified G as being between 1 and 9 clusters. Positions with methylation calls that did not pass the threshold to be called methylated or unmethylated were assigned a value of 0.5. CpG density heatmaps were calculated by counting the total number of CpG sites per position relative to the repeat start and end and dividing by the total number of repeats in each group. Methylation single-read plots were generated in the ggplot2 R package using geom_rect() to plot individual reads with methylated CpGs as red and unmethylated CpGs as blue.

chrX liftOver analysis and repeat fasta comparison

Similar to the fasta sequence comparison of the set aside T2T-CHM13 loci, the lifted chrX to HG002 coordinates were compared. A similarity score was assigned to each repeat based on crossmatch output as a percentage of the maximum score. Sequences with a score of greater than 90% and/or shorter than 50 bp were the threshold for concordant similarity or insufficient information for comparison, respectively. All other sequences were considered as potential polymorphic loci. These remaining 778 sequences of interest were filtered for length differences between the T2T-CHM13 and HG002 chrX liftOver coordinates. Simple repeats were not considered as part of this analysis. Differences that were further analyzed were loci 20 bp or greater if the T2T-CHM13 RM annotation was *Alu*, and 50 bp or greater for all other repeat types. 64 loci remained. For these 64 loci, the fasta sequence was extracted and subjected to RepeatMasker analysis.

Copy Number Comparison across primates

Copy number comparisons across primate genomes were generated with the most recent, available primate genomes for each species: *Pan trogolodytes* (accession: GCA_002880755.3) (*84*)*, Gorilla gorilla* (accession: GCA_900006655.3)(*116*)*, Pongo abelii* (accession: GCA_002880775.3) (*84*)*, Hylobates moloch* (accession: GCA_009828535.2)*, Macaca mulatta* (accession: GCA_008058575.1)(*117*)*, Rhinopithecus roxellana* (accession: GCF_007565055.1)(*118*)*, Callithrix jacchus* (accession: GCF_009663435)(*119*)*,* and *Microcebus murinus* (accession: GCF_000165445.2) (*120*). Custom BLAST databases were generated from each genome and searched for individual instances of the corresponding repeat or composite element. Due to the varying quality and completeness of these genomes, and in order to avoid returning individual composite subunits, the search was done by requiring at least an 85% length match to the query repeat / composite monomer. Standard BLASTN parameters were used for query match divergence, except in the case of highly similar gap tandem array sequences, which were rerun with a 100% match requirement across the 85% length to assure correct counts. All results were quantified manually, and coordinates were checked within each set of results to ensure that only individual instances were counted.

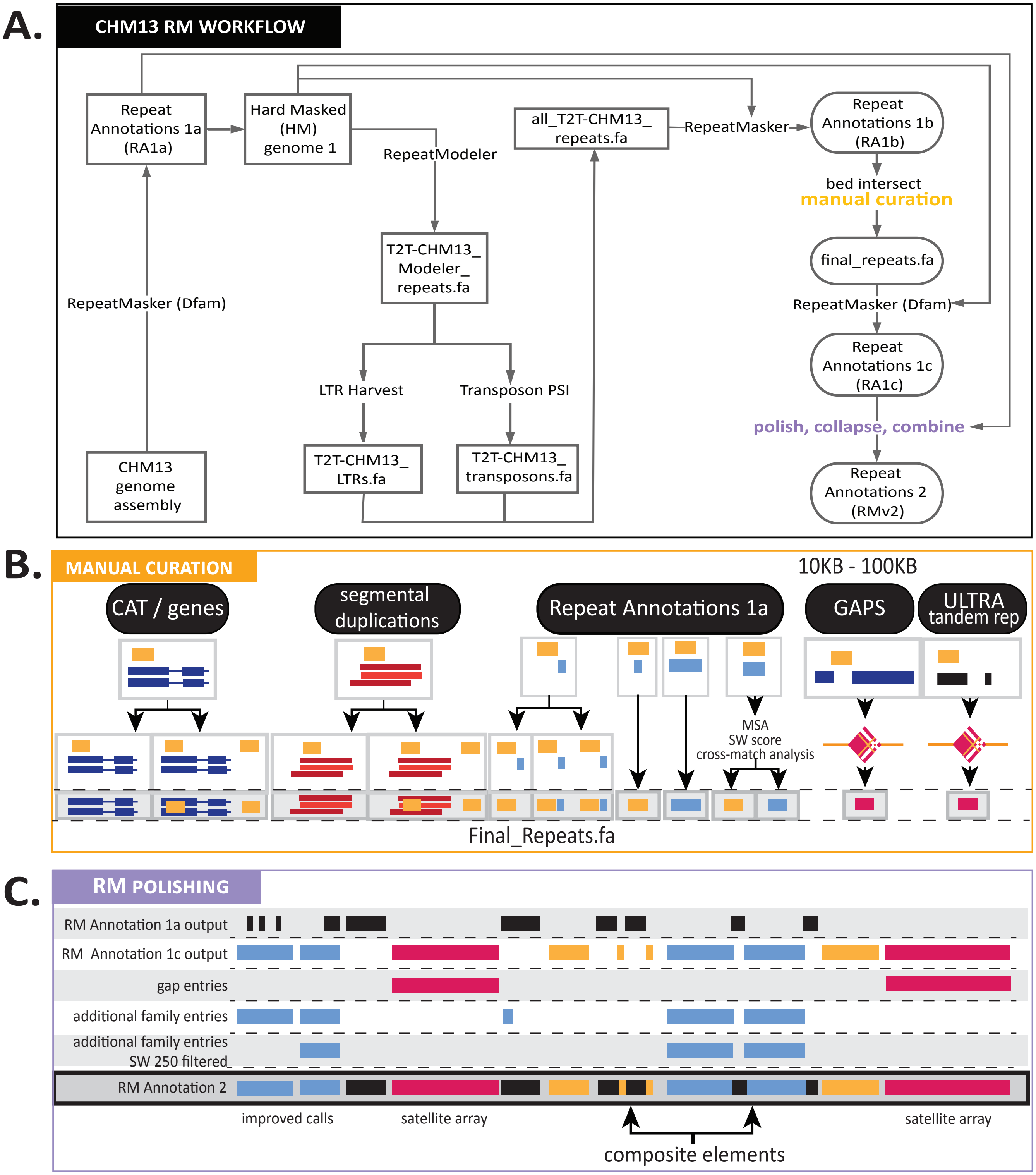
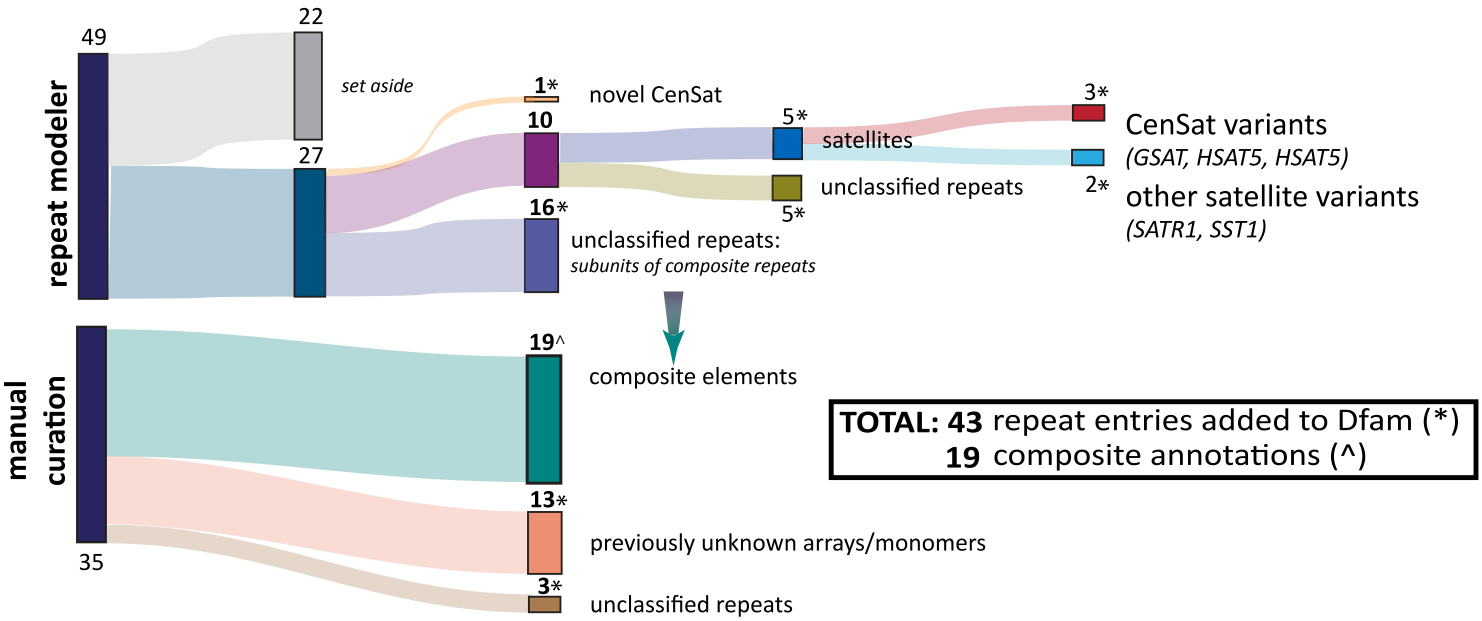


Fig. S1. A discovery workflow afforded comprehensive annotations of a complete human genome. Workflow implemented to obtain updated repeat models and the derivation of RepeatMasker Annotations 2 (RMv2), consisting of compiled and polished RM annotations submitted to Dfam **(A)** and applied to T2T-CHM13 and GRCh38 as RepeatMaskerv2 tracks. Workflow consisted of multiple iterations of RepeatMasker and RepeatModeler **(A)**. The components intersected during manual curation **(B)** include CAT/gene annotations (*11*), segmental duplications (*65*), repeats masked using Dfam (ver3.3) repeat models (*6*), tandem repeat arrays identified as gaps in annotations >10Kbp and overlap with ULTRA tandem repeat models (*9*). **(C)** Repeat model polishing was derived from a compilation of repeat masker output (previous repeat models; HM1), repeat masker 2 output (updated models; RM Annotation 2), and gap entries. Additional and previously unclassified family entries identified from RMv2 were further filtered following multiple sequence alignment (MSA) among members of the predicted category.



TOTAL: 43 repeat entries added to Dfam (*)
19 composite annotations (^)

Fig. S2. Summary of repeat annotation discovery for T2T-CHM13. Compiled annotations resulted in a final RM track for T2T-CHM13 that included the annotation of previously unknown repeats and satellite arrays outside of centromeric regions (*12*), the identification of extensions or variants to current repeat models, and the identification of composite elements consisting of multiple repeats. Plot of 49 previously unknown human repeats identified through RepeatModeler and 35 through manual curation.

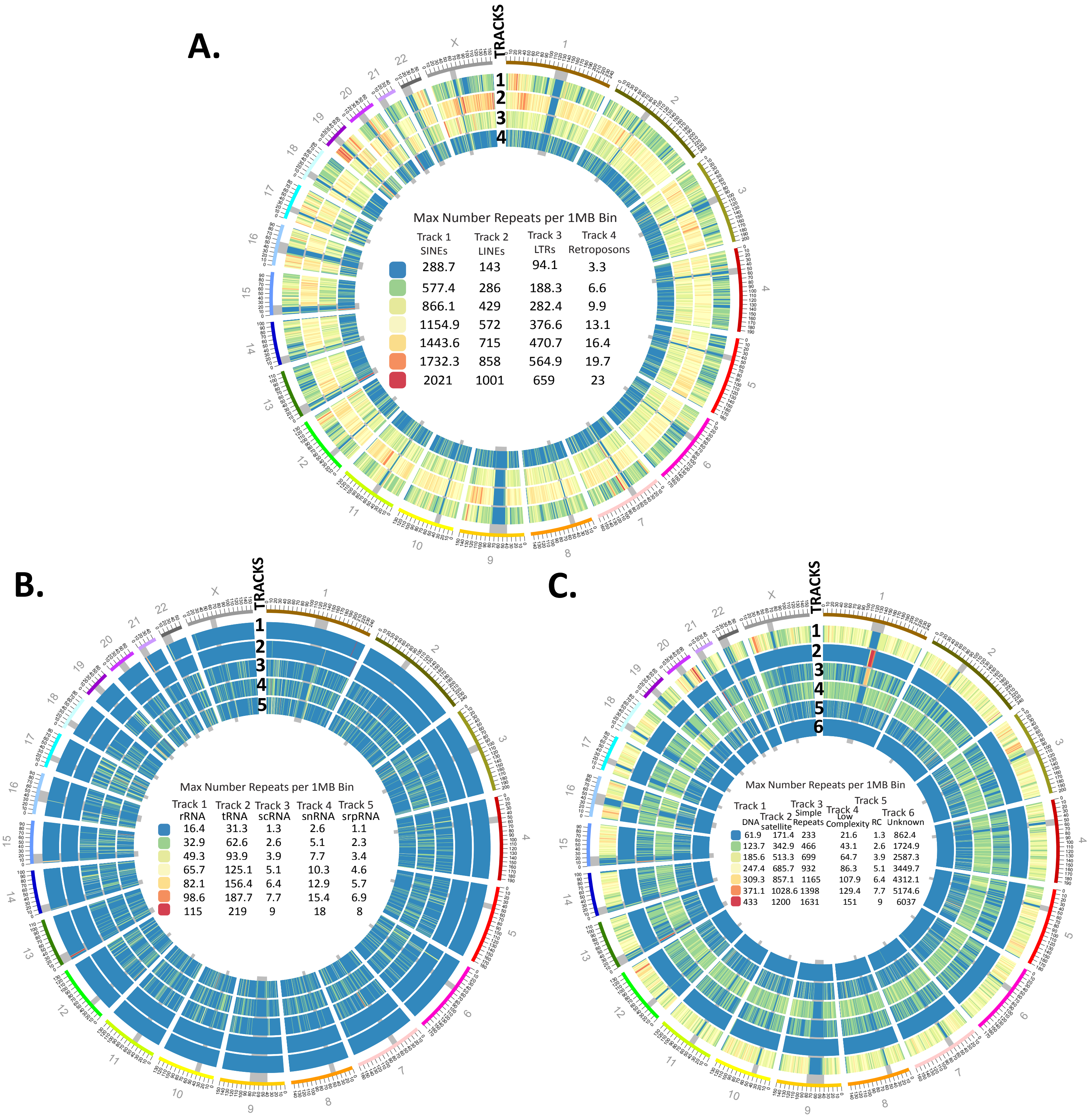


Fig. S3. Repeat density across T2T-CHM13 by family classification. Counts of all repeats identified by our repeat annotation pipeline were binned into 1Mbp windows across all chromosomes (color coded and numbered, outer ring) in CHM13v1.1 and are shown as Circos heatmaps corresponding to **(A)** retrotransposon classes, **(B)** RNAs, and **(C)** all other repeat classes. Centromere blocks (including centromere transition regions) are denoted by grey bars that span all tracks. Tracks are numbered (1, 2, …) starting from the outer ring as indicated. Each repeat class track is scaled independently with the scales located in the middle of each respective Circos.

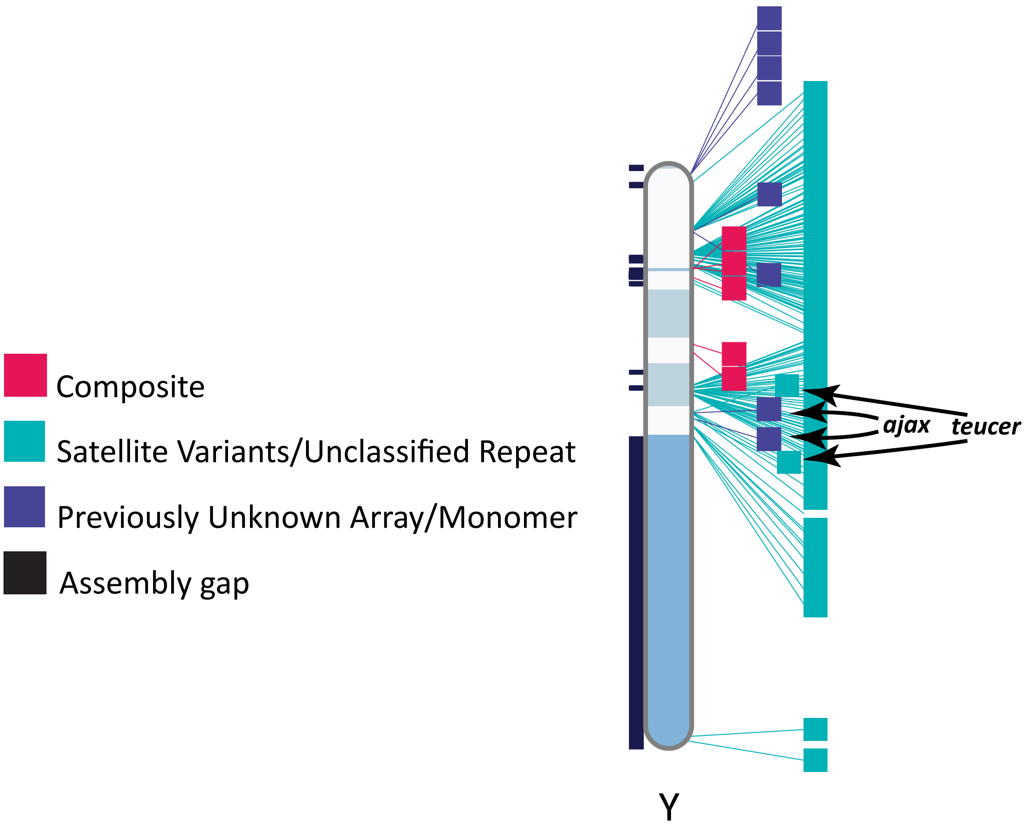

Composite
Satellite Variants/Unclassified Repeat
Previously Unknown Array/Monomer
Assembly gap
ajax
teucer
Y

Fig. S4. T2T-CHM13-based repeat annotations reveal previously unknown repeat classifications on the GRCh38 Y chromosome assembly. Ideogram of GRCh38 Chromosome Y indicating the locations of annotated composite elements (red), satellite variants and unclassified repeats (aqua), and previously unannotated/unknown arrays or monomers of sequences found within those arrays (purple). Gaps in the Chromosome Y assembly are shown in black boxes to the left of the chromosome. Notably, ajax and teucer are found together at two loci without TELO_comp as part of an inversion in the Azoospermia Factor c region of the Y chromosome, a region with recurrent de novo microdeletions linked to male infertility (*152*).

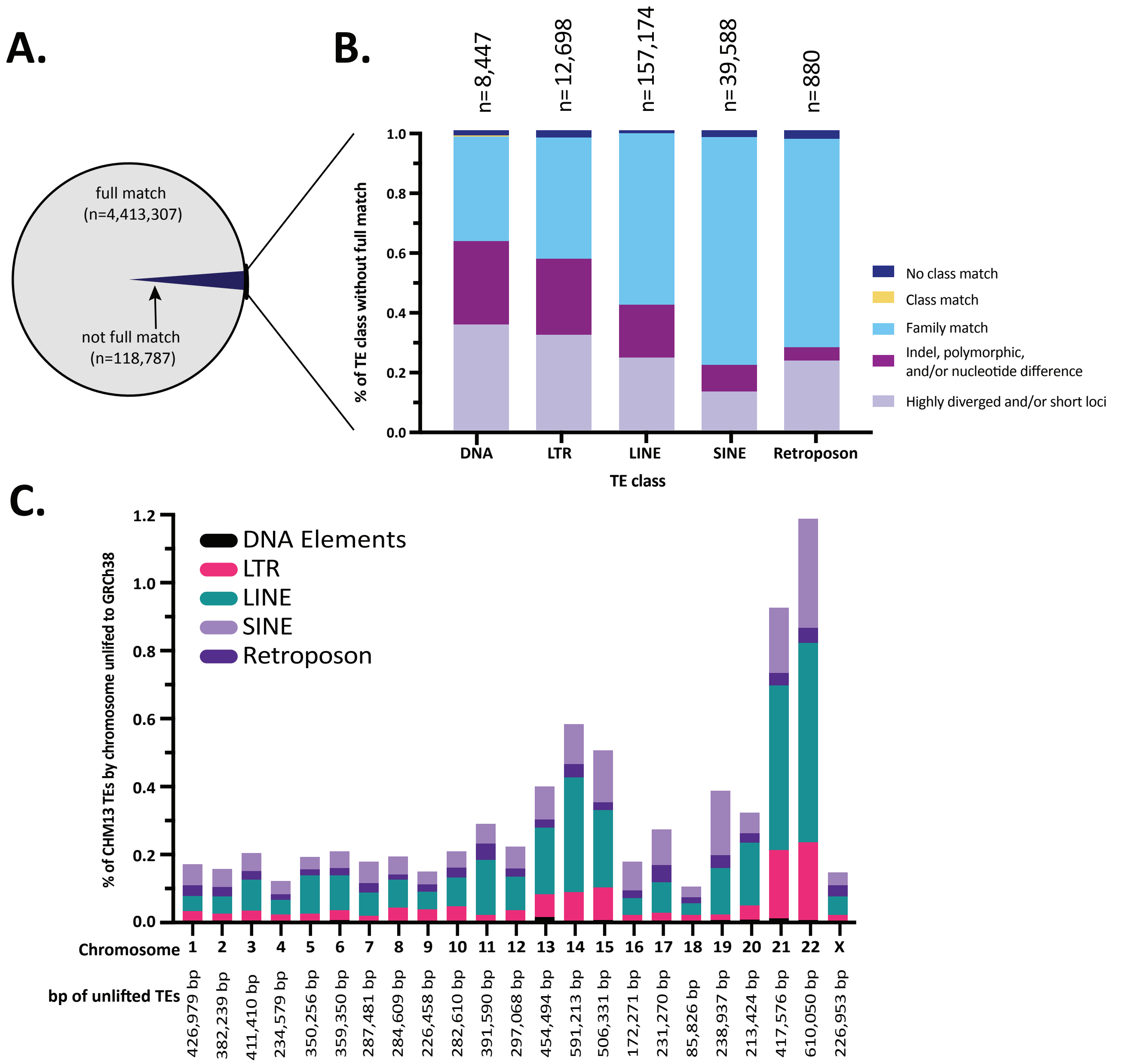


Fig. S5. Lifted TE pair annotations are discordant between T2T-CHM13 and GRCh38. (A) TEs in T2T-CHM13v1.1 with a full match with GRCh38 represent TEs with no change in annotation. The remaining 118,787 without a full match were further classified by discordance category. **(B)** "Not full match" classifications were further broken down into discordance categories as follows: 1) TEs lacking a class match (dark blue), 2) TEs with a class match but changed family (yellow), 3) TEs with a family match with a subfamily change (light blue), 4) TEs with nucleotide differences (purple) and highly diverged sequences/short loci (light purple), both of which are low confidence changes that may be the result of batch effects with RepeatMasker. **(C)** Of TE annotations in T2T-CHM13v1.1, percent unlifted to GRCh38, shown by chromosome (X axis) and repeat class normalized to bp of each chromosome in CHM13. TE family indicated by color in key inset; n= number of bp on each chromosome in CHM13v1.1 represented by unlifted TEs.

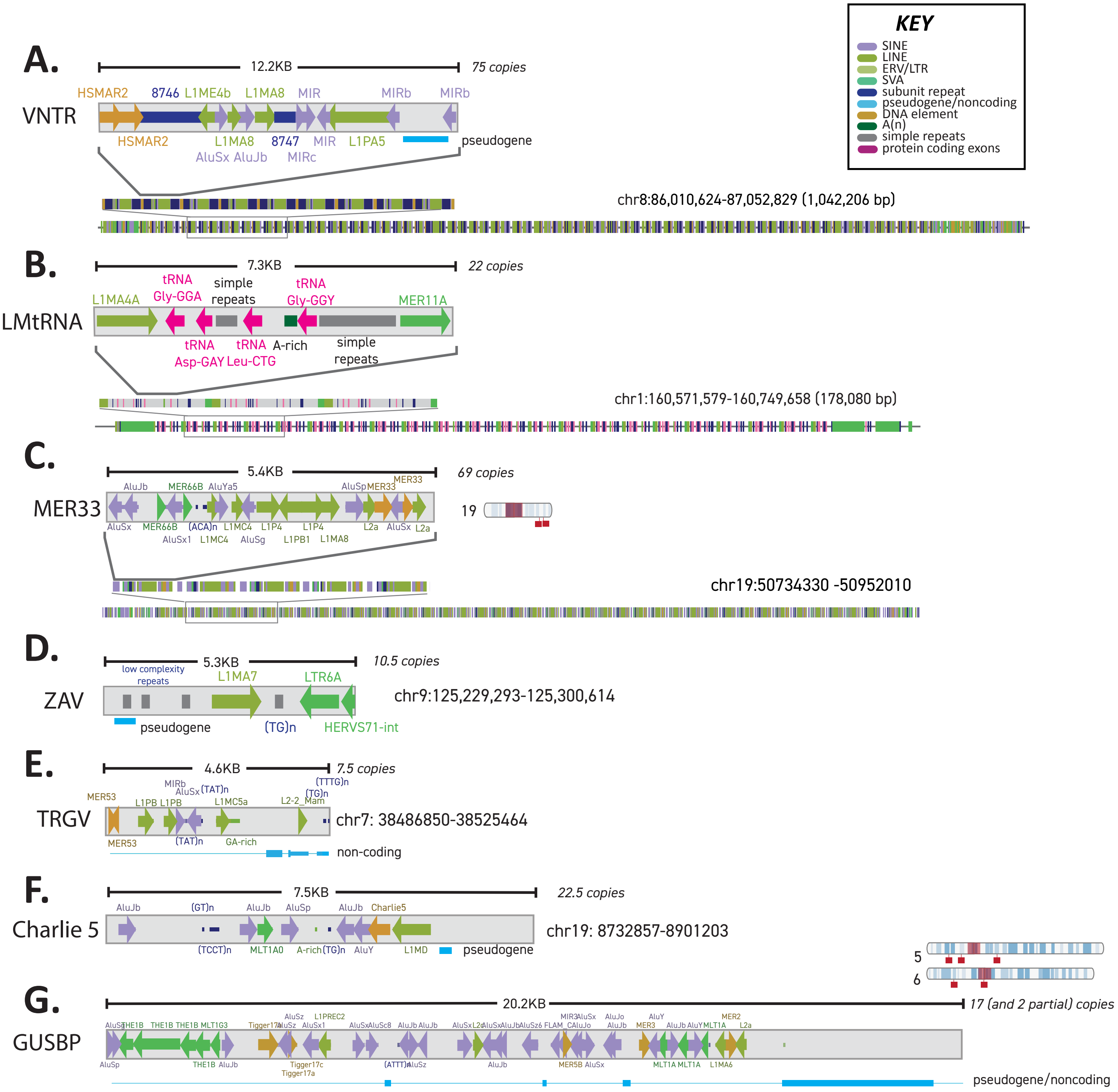


Fig. S6. Annotations for composite elements in T2T-CHM13 lacking exonic material. Repeat annotations for composite elements found without protein coding sequences: **(A)** VNTR, **(B)** LMtRNA, **(C)** MER33, **(D)** ZAV, **(E)** TRGV, **(F)** Charlie 5, **(G)** GUSP. Each subunit repeat type and orientation, including all TEs, simple and low complexity repeats, subunits and other features such as pseudogenes, color coded as per KEY (inset at top), is indicated. For each composite, the size of the composite core unit is indicated, as is the number of repeat units in the array and chromosomal location. The order of core units within the array are shown in two zoom images of RepeatMaskerv2 browser tracks for repeats found in large arrays **(A-C)**. For composites found arrayed in more than one location **(C, G)**, chromosome ideogram indicates the locations of the composite arrays in T2T-CHM13. Centromere blocks (including centromere transition regions) are indicated in orange.

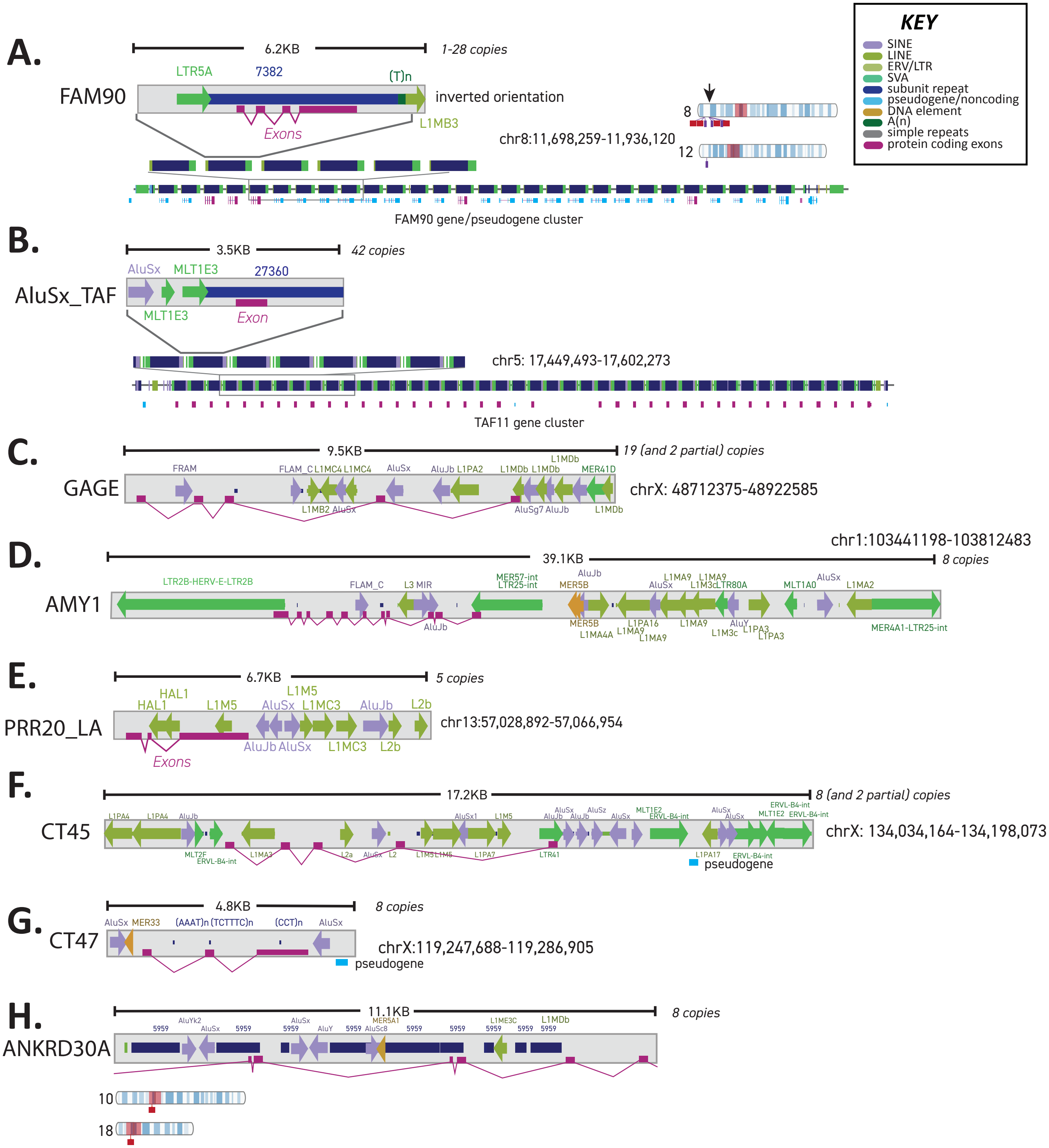


Fig. S7. Updated annotations for composite elements containing gene predictions found in T2T-CHM13. Repeat annotations for composite elements found without protein coding sequences: **(A)** FAM90, **(B)** AluSx-TAF, **(C)** GAGE, **(D)** AMY, **(E)** PRR20_LA, **(F)** CT45, **(G)** CT47, **(H)** ANKRD30A. Each subunit repeat type and orientation, including all TEs, simple and low complexity repeats, subunits and other features such as exons, color coded as per KEY (inset at top), is indicated. For each composite, the size of the composite core unit is indicated, as is the number of repeat units in the array and chromosomal location. The order of core units within the array are shown in two zoom images of RepeatMask-erv2 browser tracks for repeats found in large arrays **(A-B)**. For composites found arrayed in more than one location **(A, H)**, chromosome ideogram indicates the locations of the composite arrays in T2T-CHM13. Centromere blocks (including centromere transition regions) are indicated in orange.

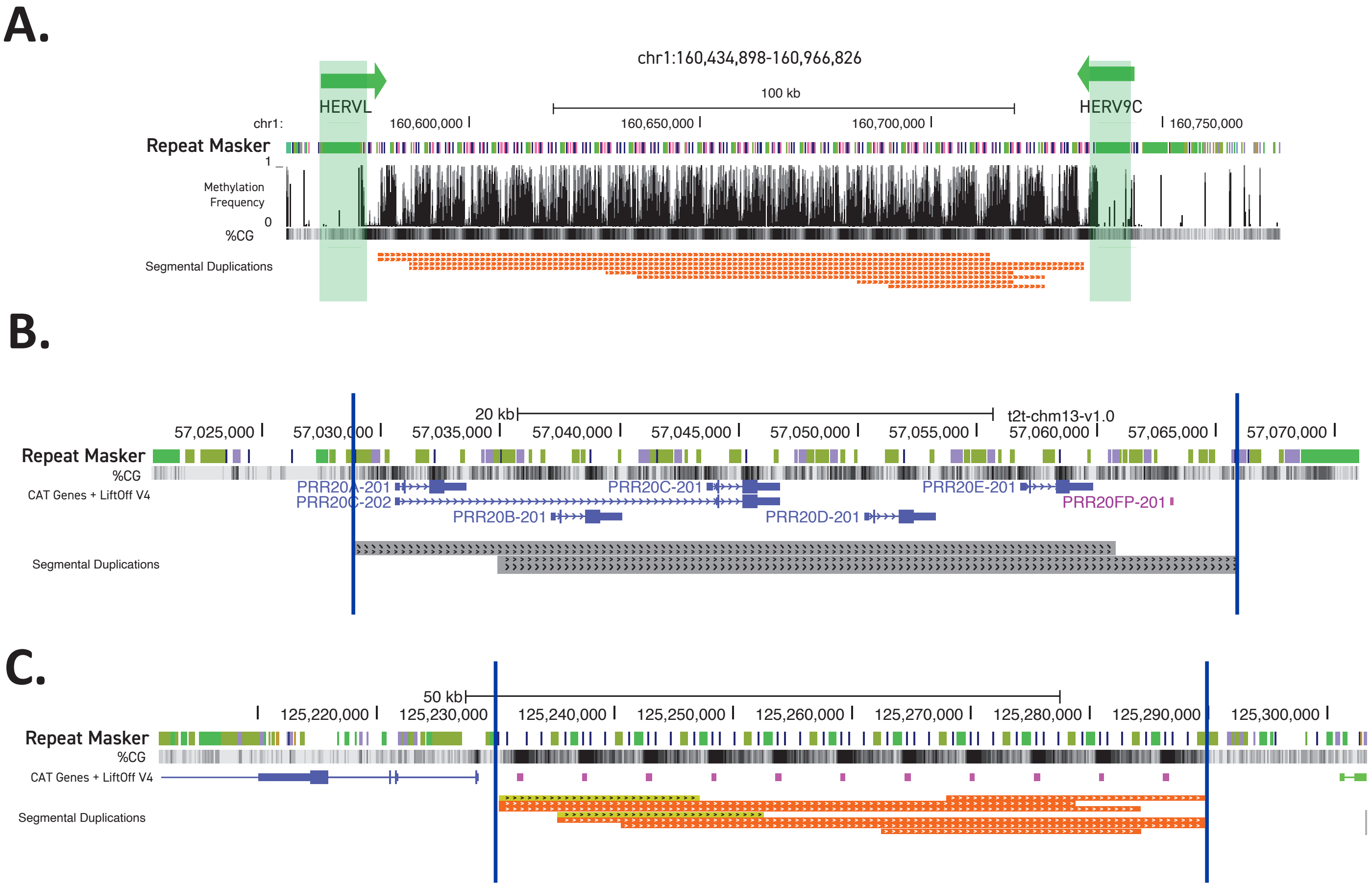


Fig. S8. Discerning composite units from segmental duplications. (A) Browser track of the LMtRNA locus showing methylation frequency, %CpG, and segmental duplication annotations. HERVs flanking the array are shown in green, with transcript unit orientation indicated with an arrow. **(B)** and **(C)** Browser tracks of PRR-LA and ZAV composite arrays showing %CpG, gene annotations, and segmental duplication annotations. Array boundaries are indicated by a vertical line.

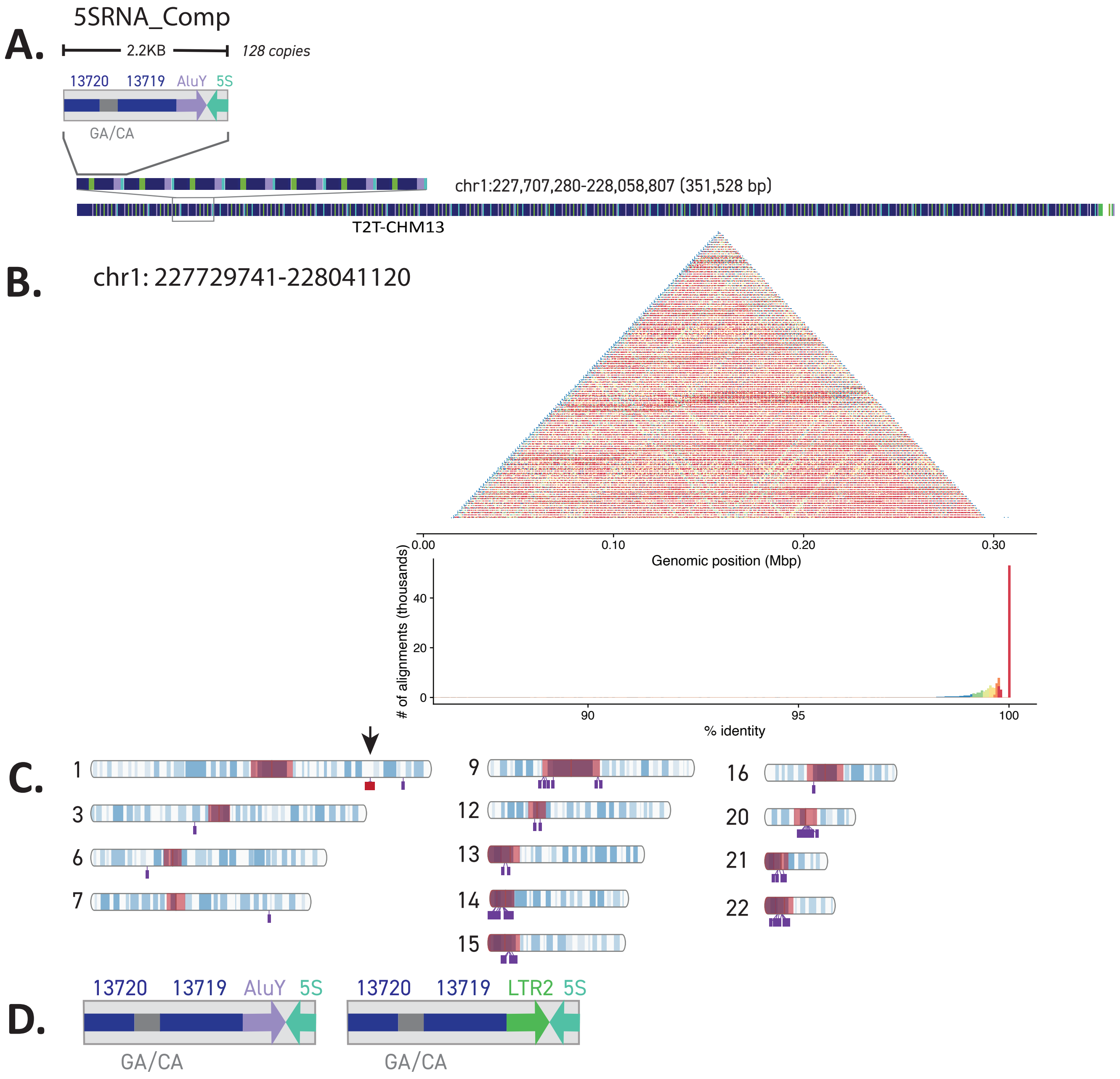


Fig. S9. A composite containing part of the 5SRNA is found at multiple loci in T2T-CHM13 and arrayed at a single locus. (A) The 5SRNA composite contains an *Alu*Y insertion, three previously known repeats (GA-rich low complexity repeat, CA simple repeat, 5S) and two other composite subunits (13719, 13720). The composite is found in an array on Chromosome 1. The order of core units within the array are shown in two zoom images of Repeat Maskerv2 browser tracks. **(B)** A self-alignment dot plot of 5SRNA composite subunits across the array. Histogram denotes the color scale and distribution of alignments for the plot showing high intra-array sequence similarity (with a peak ~100%). The array is visible in the dot plot as the brighter red triangle shape, in which connecting diagonals (the arms of the triangle) represent shared sequence identity A 5% size (bp) increase was added flanking the array, which is visible as the area with lower shared sequence identity (blue) on the left and right of the dot plot. **(C)** Ideogram of CHM13 indicates the 49 locations of the 5SRNA composites as singletons (purple) and the only array found in T2T-CHM13 (red). Arrow indicates the location of the array illustrated in **(A)**, which is also the only location containing the *Alu*Y insertion. Centromere blocks (including centromere transition regions) are indicated in orange. **(D)** The 5SRNA composite is found as two different structures: with an *Alu*Y insertion (left) in array form and with a LTR2 insertion (right) at all monomeric locations. Both contain three previously known repeats (GA-rich low complexity repeat, CA simple repeat, 5S) and two other annotated composite subunits (13719, 13720).

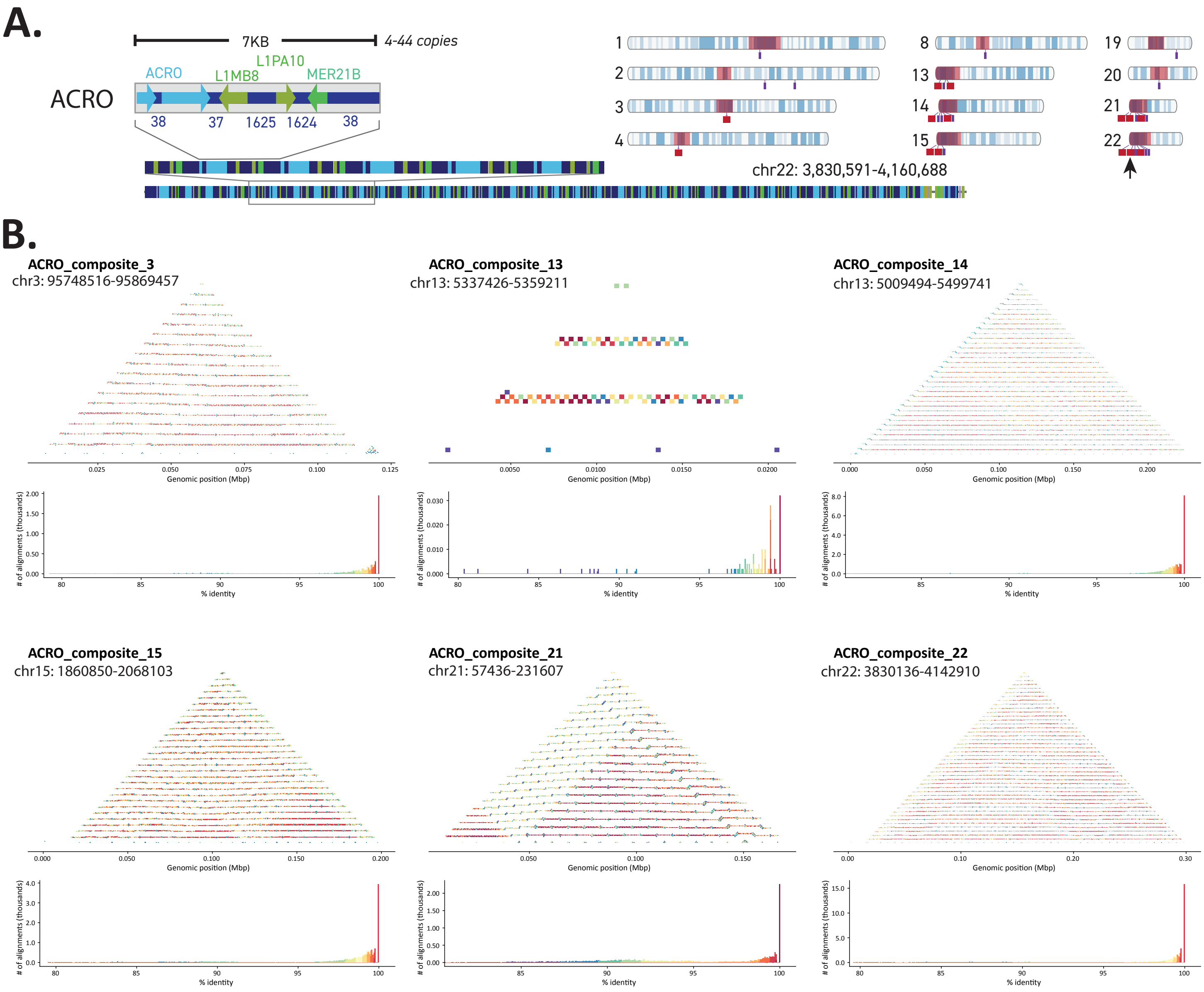


Fig. S10. ACRO_Composites are found in arrays on multiple chromosomes in CHM13. (A) Structure of the ~7Kbp ACRO composite includes four previously known repeats (ACRO, L1MB8, L1PA10, MER21B) and four composite subunits (37, 38, 1624, 1625). Ideogram of CHM13 indicates the 30 locations of the ACRO composites as both singletons (purple) and arrays (red). Centromere blocks (including centromere transition regions) are indicated in orange. Arrow indicates the location of the array illustrated on the left. **(B)** Self-alignment dot plots for the largest ACRO composite arrays on each acrocentric chromosome (chromosomes 13, 14, 15, 21, and 22) and one non-acrocentric chromosome (Chromosome 3) are shown displaying alignment similarity with histograms denoting the color scale and distribution of alignments for each independently colored plot. High sequence identity is seen within each array (>95%, excluding extra flanking region) and strong structural similarities between each array (with Chromosome 21 being a slight outlier, although the array still has the triangular structure typical of a highly repetitive sequence). A 5% size (bp) increase was added flanking each array. Note that these plots differ in structure from Fig. S9 due to the size of the arrays.

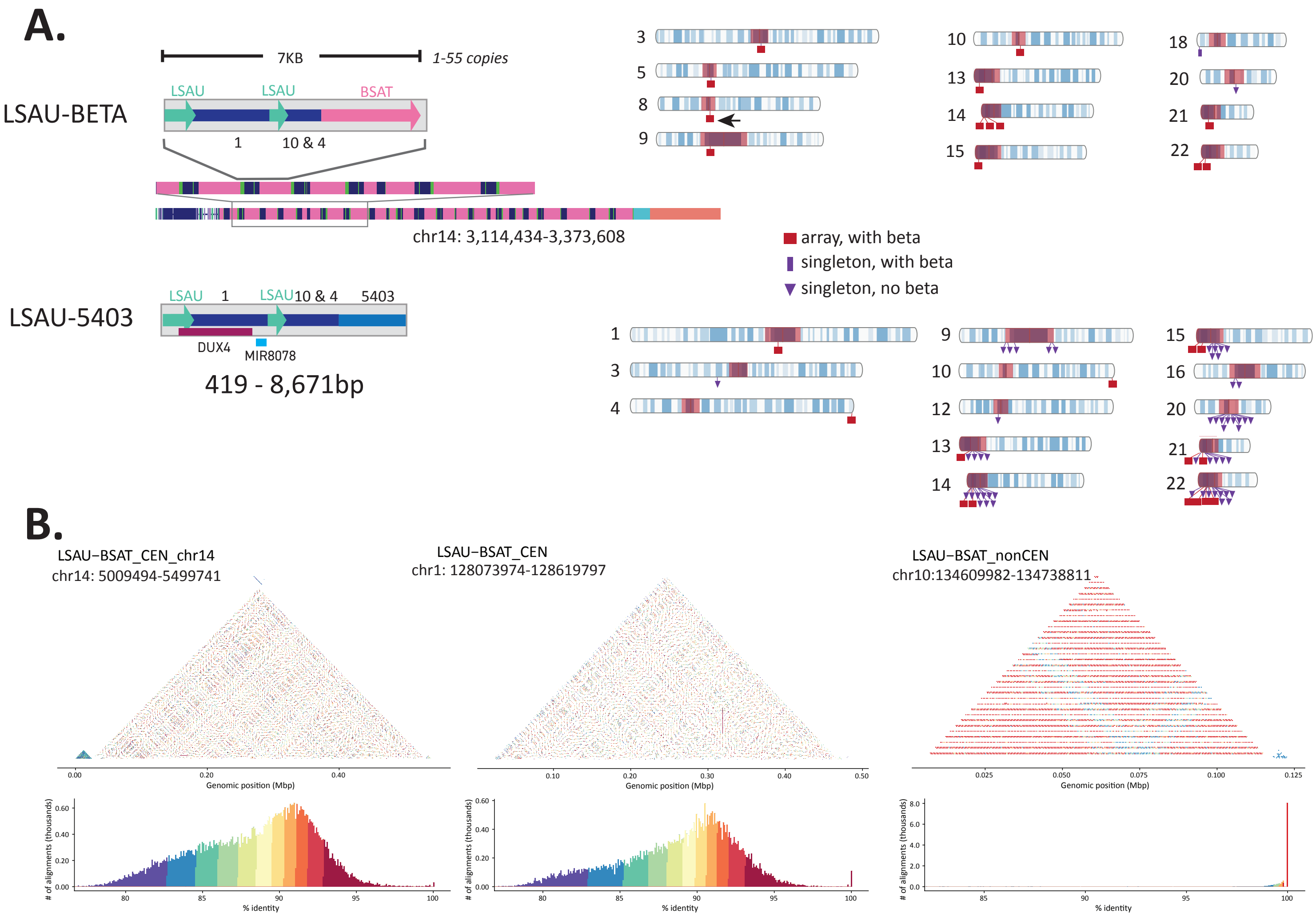


Fig. S11. LSAU-BETA_Composites are found in two forms and arrayed on multiple chromosomes in T2T-CHM13. (A) The LSAU-BETA composite is found as two different structures: without the 5403 composite subunit (Top; 15 loci) and with the repeat-5403 (Bottom; 57 loci). Both contain three other composite subunits (1, 4, 10), as well as the LSAU satellite. Ideograms of CHM13 for each of the two different structures indicate the locations of the LSAU-BETA composites as singletons (purple) and arrays (red), as well as the presence (square) or absence (triangle) of the BETA satellite. BETA is not shown in the structure containing 5403 since it is not part of the composite itself, but rather found at one edge of each of the two arrays. These two arrays (chromosomes 4 and 10) are subtelomeric and associated with the DUX4 genes (*11*). Only one locus (Chromosome 8) lacks the LSAU satellite (arrow). Centromere blocks (including centromere transition regions) are indicated in orange. Arrows indicate the location of the arrays illustrated on the left. **(B)** Self-alignment dot plots for three example arrays (two centromeric and one non-centromeric) are shown displaying alignment similarity with histograms denoting the color scale and distribution of alignments for each independently colored plot. The spectral color scheme represents a scale of 0 (purple) to 100% (red) sequence similarity. The Chromosome 1 and Chromosome 10 loci both contain the 5403 subunit, while the Chromosome 14 locus does not. The two centromeric arrays have a lower intra-array sequence similarity (with a normal distribution suggesting a lack of non-random sequence similarity) compared to the non-centromeric array (with a single peak at ~100%). This suggests a much more complex structure in centromeric arrays, regardless of the presence of the additional 5403 subunit. The complexity of this array is also apparent in the Chromosome 10 non-centromeric array, given that the array pattern is broken approximately ⅔ of the way through (blue pattern). A 5% size (bp) increase was added flanking each array and each plot is colored independently.

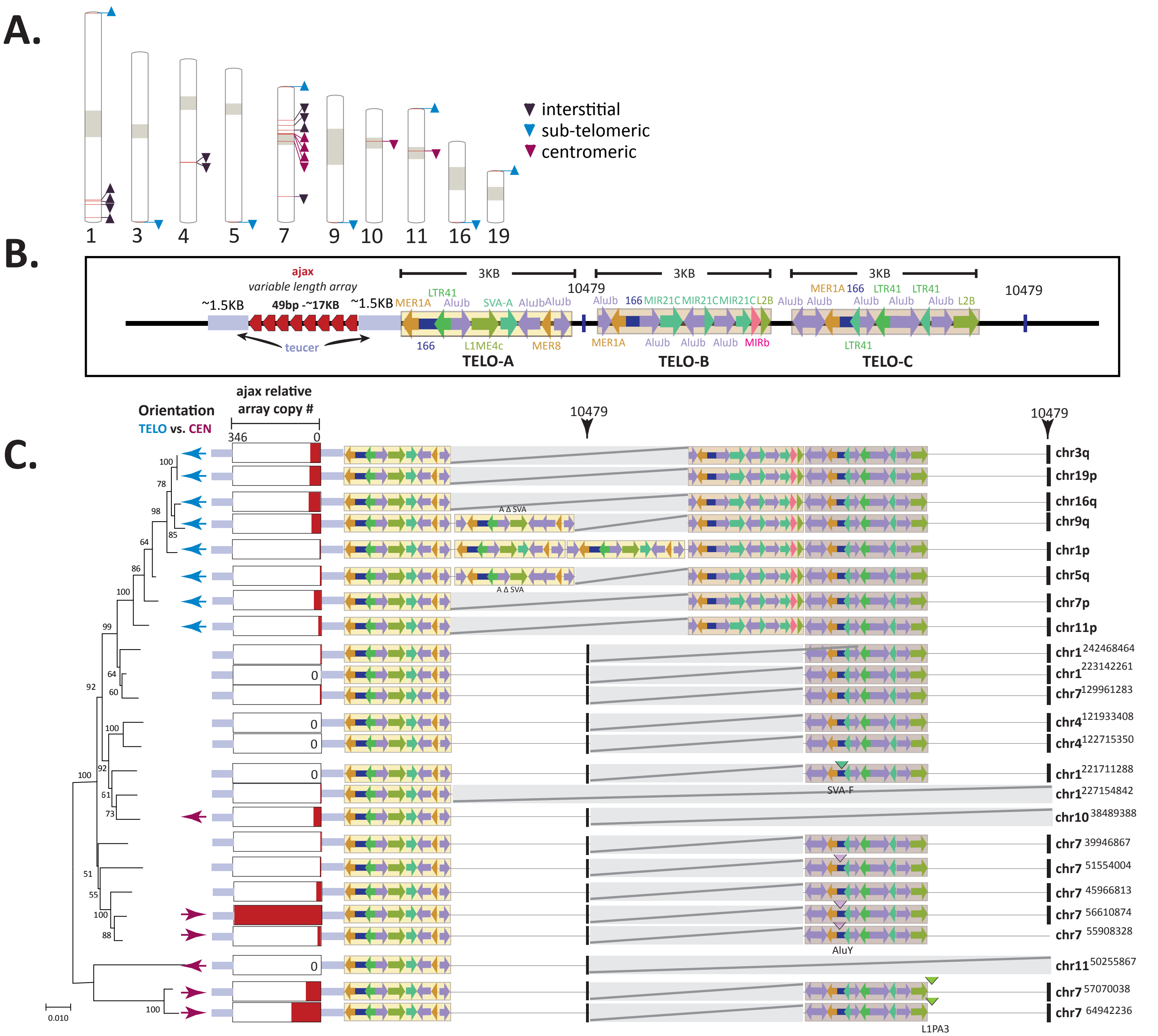


Fig. S12. TELO_comp element distribution and relationships across T2T-CHM13. (A) The location of TELO-composite elements across T2T-CHM13 is indicated by red bars on chromosomes. Tan blocks demarcate centromeres and centromere transition regions. Chromosome regions containing TELO-composites across the karyotype (21) are color coded [interstitial – purple, sub-telomeric, within 200Kbp of chromosome end– aqua, centromeric – red] with orientation indicated by arrow direction. **(B)** Each T2T-CHM13 TELO-composite element consists of a duplication of a teucer repeat (blue) separated by a variable 49bp (ajax) repeat array (red arrowheads) and three different composite subunits (TELO-A, -B, -C). Repeat and TE annotations are shown. Some copies of TELO-composite contain the repeat “10479” between the TELO-A and TELO-C subunits, and/or following the TELO-C subunit. **(C)** Schematic alignment of all complete CHM13 TELO-composite elements (location indicated to the right, centromeric and interstitial locations indicated with starting bp). Locations of “10479” (black box) repeat are arrowed (top). Relative number of each TELO subunit as pictured in **(B)** with deletions represented by grey bars. Additional TE insertions are indicated for TELO-C. Relative array size of ajax repeats shown to scale among all TELO-composite elements. Orientation of the element indicated with respect to the centromere (purple arrow) telomere (blue arrow) and interstitial (no arrow) regions. Left, RaxML phylogeny of TELO-composite elements with bootstrap values at each node and distance indicated by length of branch.

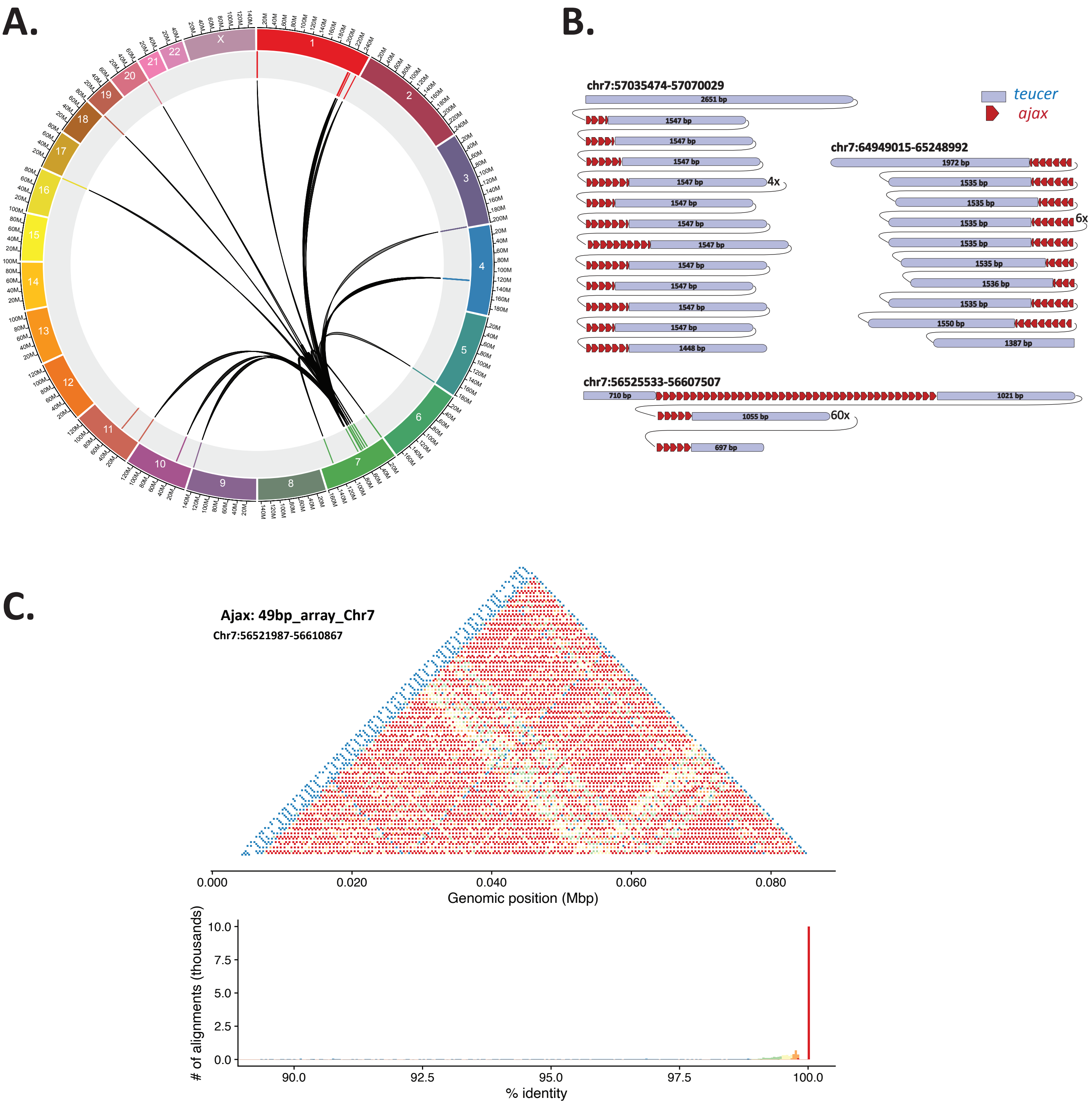


Fig. S13. Derivation of TELO_comp loci on Chromosome 7. (A) Segmental duplication synteny map connecting TELO_Comp loci found in the centromere of Chromosome 7 with other TELO_Comp loci (fig. S12, Table S9) **(B)** Organization of the Chromosome 7 locations of the ajax repeat (each red arrowhead denotes a single monomer of the repeat) and teucer (blue). "#X" indicates a subunit is arrayed at # copy number. **(C)** A self-alignment dot plot of the 49bp array (ajax) on Chromosome 7 and histogram denoting the color scale and distribution of alignments for the plot showing high intra-array sequence similarity (with a peak ~100%). The array is visible in the dot plot as the brighter red triangle shape, in which connecting diagonals (the arms of the triangle) represent shared sequence identity. A 5% size (bp) increase was added flanking the array, which is visible as the area with lower shared sequence identity (blue) on the left and right of the dot plot.

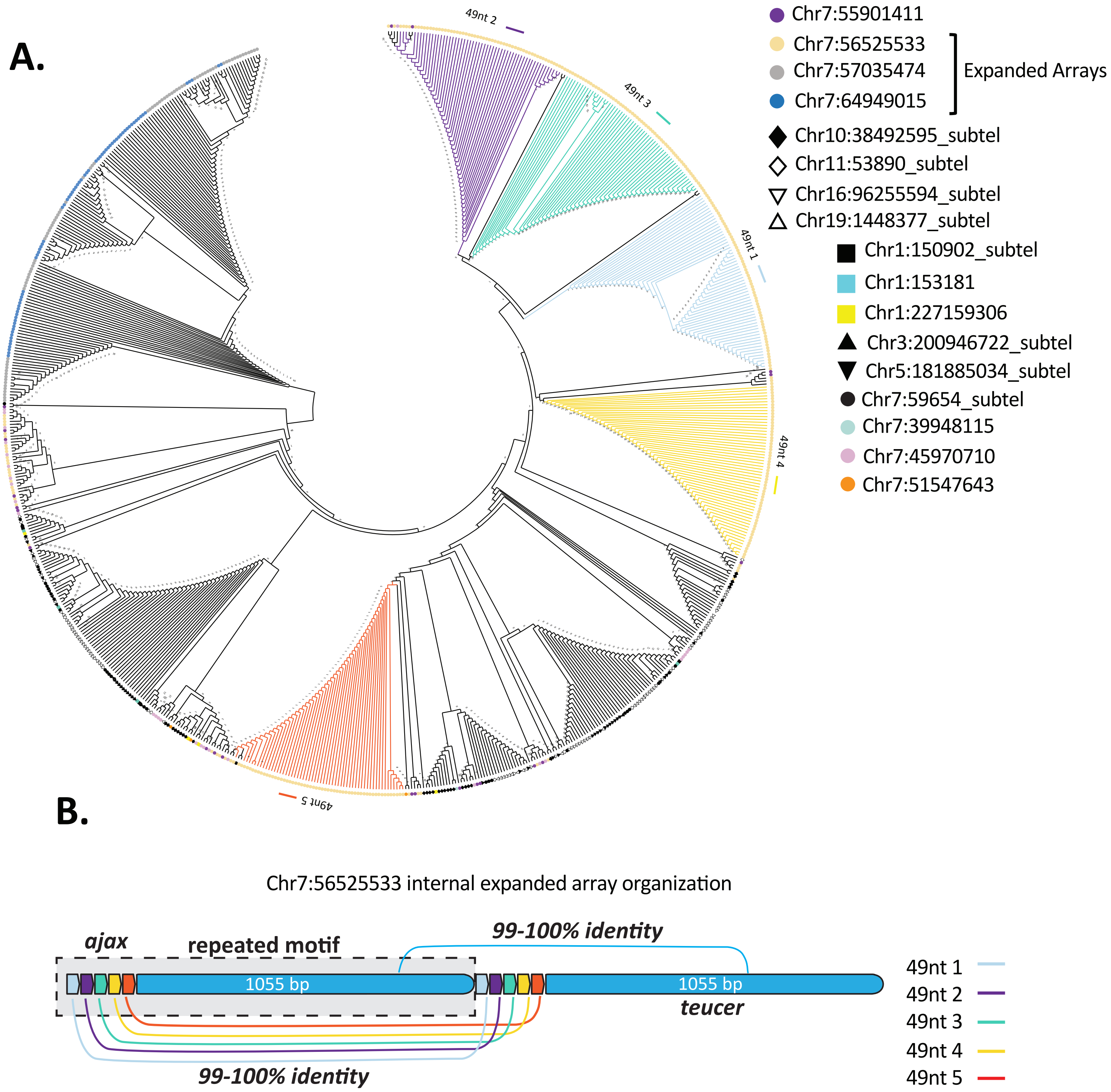


Fig. S14. Phylogenetic analyses of ajax repeats reveals different patterns of evolution influence subtelomeric vs pericentromeric arrays. (A) A Neighbor Joining unrooted phylogenetic tree using 763 full-length repeats (table S10) reveals ajax repeats in subtelomeric regions (black and open shapes) do not cluster together in array-specific or chromosome-specific subtrees. In contrast the ajax pericentromeric repeats from each of the three expanded loci tend to cluster in array-specific clades, suggesting that these arrays evolve under concerted evolution. Chr7:56525533 ajax monomers are indicated by colored branches for monomers 1 to 5 (e.g. 49nt 1, 49nt 2, ...)(see **(B)**), respectively, suggesting a higher order repeat, or super-repeat, structure across the array. **(B)** Schematic representation of Chr7:56525533 organization of ajax and the associated teucer repeat, which together comprise an element that evolves as a single unit (gre box). The 49nt monomeric ajax repeats of the Chr7:56525533 locus are indicated by colored arrowheads (key to right) for monomers 1 to 5, as per **(A)**, whereas teucer is represented in blue.

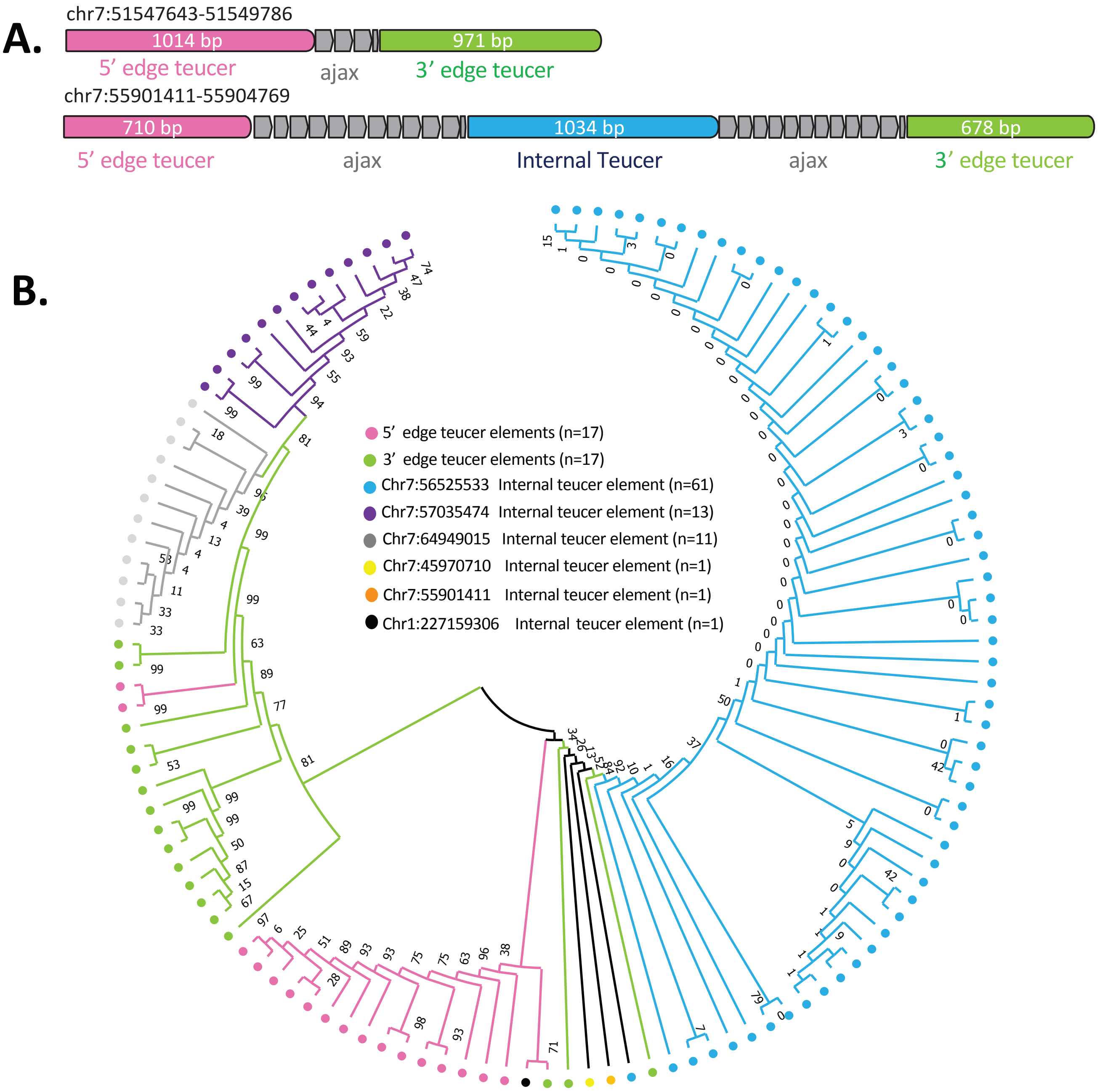

A.
chr7:51547643-51549786
1014 bp
971 bp
5’ edge teucer
ajax
3’ edge teucer
chr7:55901411-55904769
710 bp
1034 bp
678 bp
5’ edge teucer
ajax
Internal Teucer
ajax
3’ edge teucer
B.
● 5’ edge teucer elements (n=17)
● 3’ edge teucer elements (n=17)
● Chr7:56525533 Internal teucer element (n=61)
● Chr7:57035474 Internal teucer element (n=13)
● Chr7:64949015 Internal teucer element (n=11)
● Chr7:45970710 Internal teucer element (n=1)
● Chr7:55901411 Internal teucer element (n=1)
● Chr1:227159306 Internal teucer element (n=1)

Figure S15. Sequence relatedness between teucer arrays suggests location specific patterns of evolution and expansion. (A) Schematic representation of ajax+teucer composite organization indicating 5’ edge teucer elements (pink), ajax repeats (grey), internal teucer elements (blue), and 3’ edge teucer elements (green) described at two loci as indicated. Three separate consensi have been derived for the teucer element based on the internal sequence and the edge sequences. The different areas in which the element occurs appear to be under different evolutionary pressure and change at different rates. **(B)** A maximum likelihood phylogenetic tree using 122 teucer elements (table S11) reveals that teucer derived from either the 5’ edge (pink) or 3’ edge (green) form distinct clusters, suggesting independent evolution by array position. Internal teucer elements from expanded arrays show array-specific clusters that correspond with the ajax phylogenetic analyses (fig. S14). Chr7:56525533 internal teucer elements (light blue) show higher similarities with 5’ end teucer elements, suggesting the 5’ end contributed to the array expansion. The Chr7: 57035474 and Chr7: 64949015 arrays (purple and grey, respectively) cluster with 3’ end teucer elements, suggesting the 3’ end contributed to the array expansion. The relatedness of teucer elements from the three expanded arrays to the subtelomeric teucer elements suggest that independent events are responsible for the origin of Chr7:56525533 and Chr7: 57035474/Chr7: 64949015 arrays. All teucer elements were positioned in the same orientation and form the same junction with the ajax arrays.

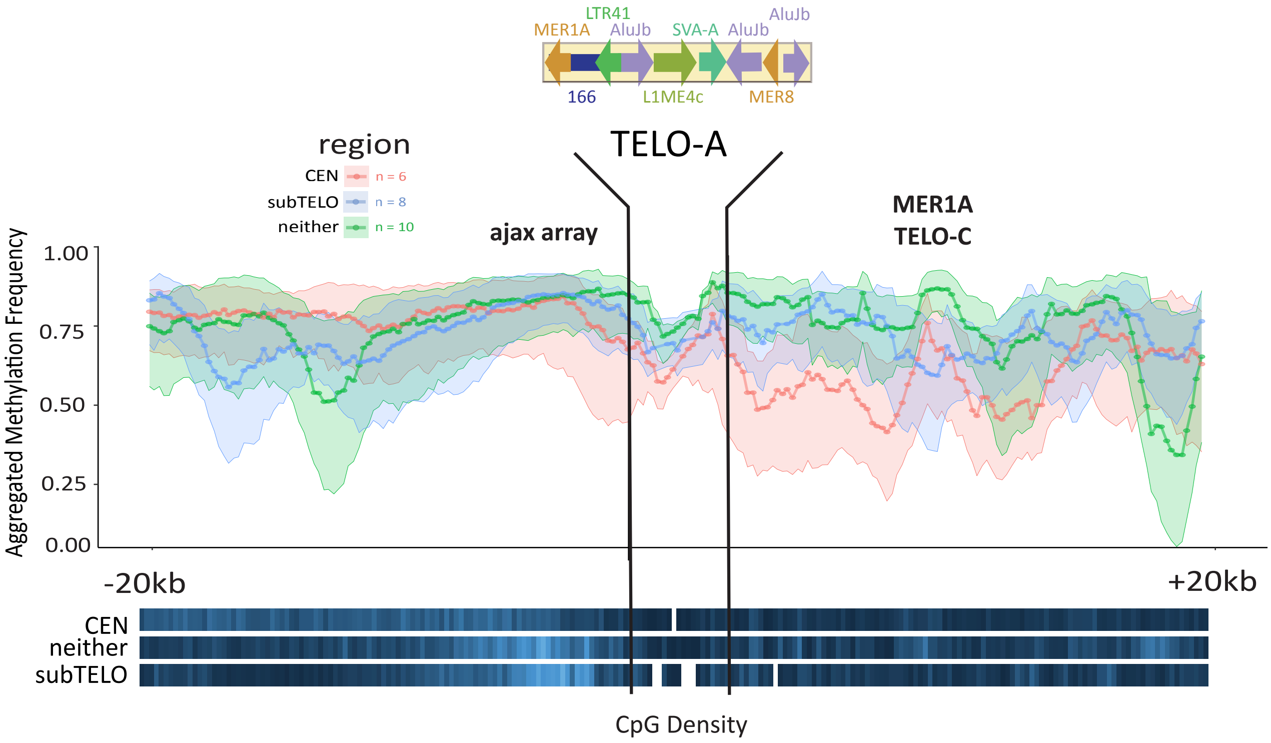


Fig. S16. Methylation metaplot for HG002 reveals a distinct epigenetic signature specific to TELO-Comp elements. Metaplot of aggregated methylation frequency (average methylation of each bin across the region, 100 bins total) centered on the TELO-A subunit (top), ±20Kbp, grouped by chromosomal location (orange – centromeric, blue – subtelomeric, green – interstitial). CpG density for each group is indicated at the bottom (white - no CpG, dark blue - low CpG, bright blue - high CpG). The location of the ajax repeat array and the MER1A element within the TELO-C subunit are indicated (top).

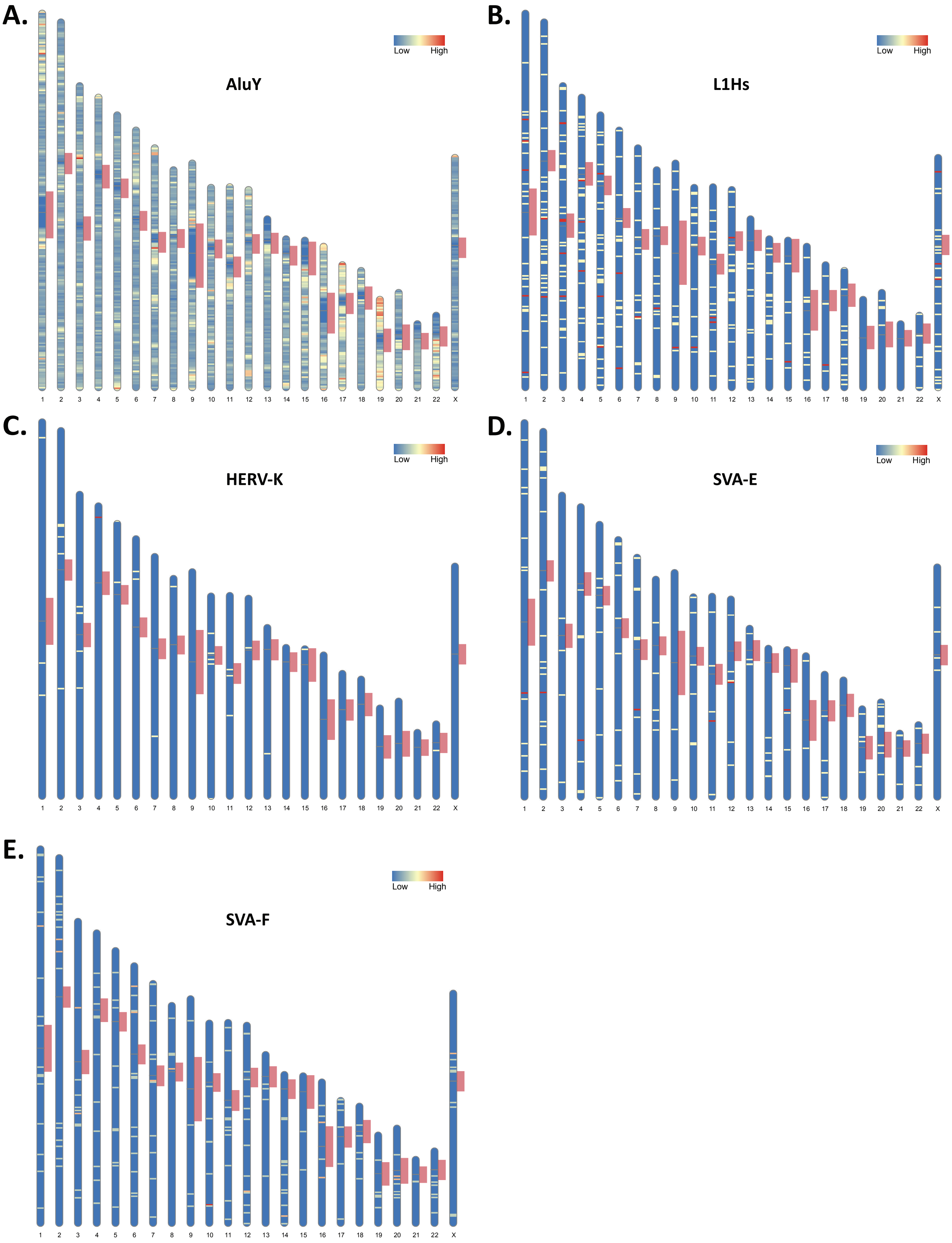
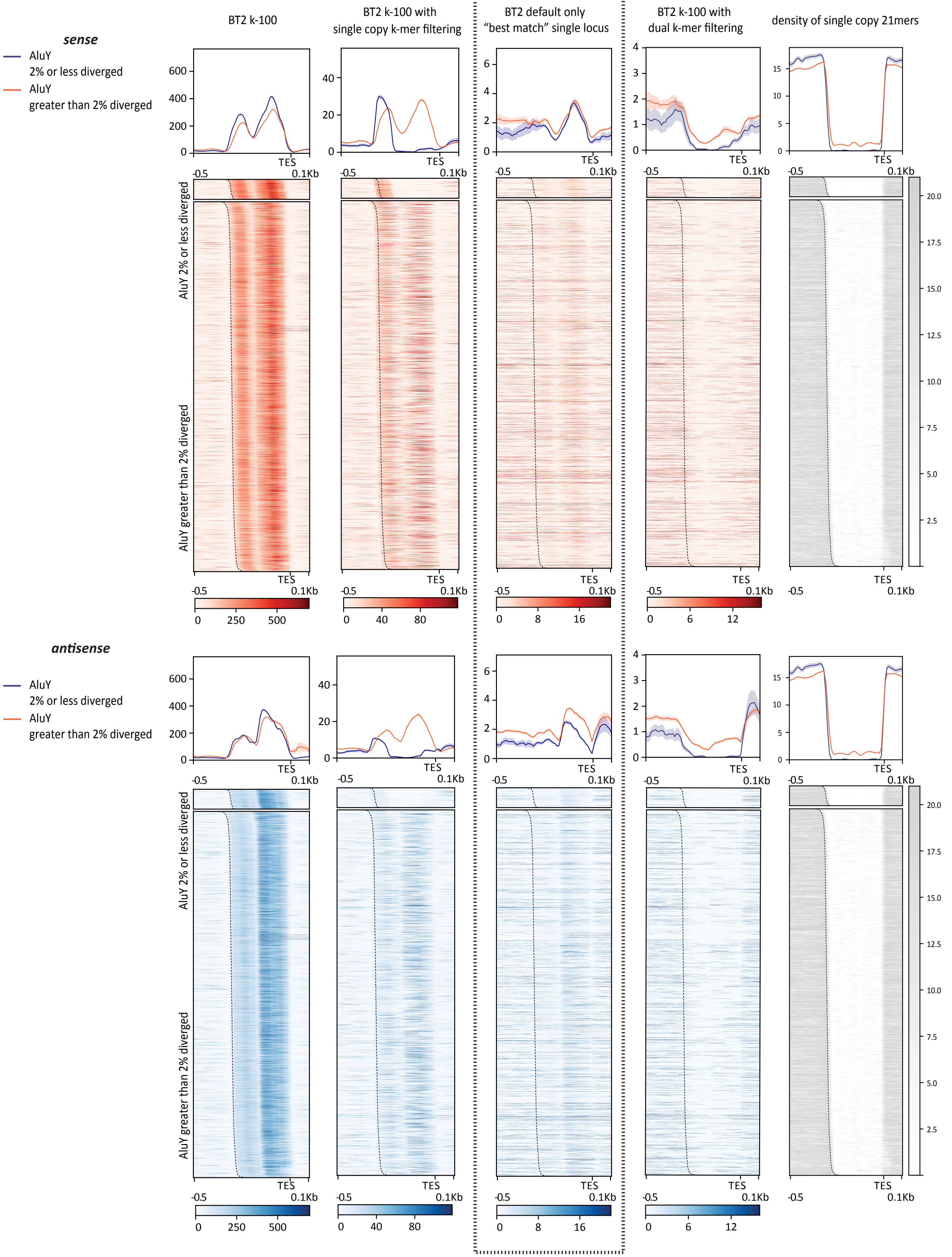
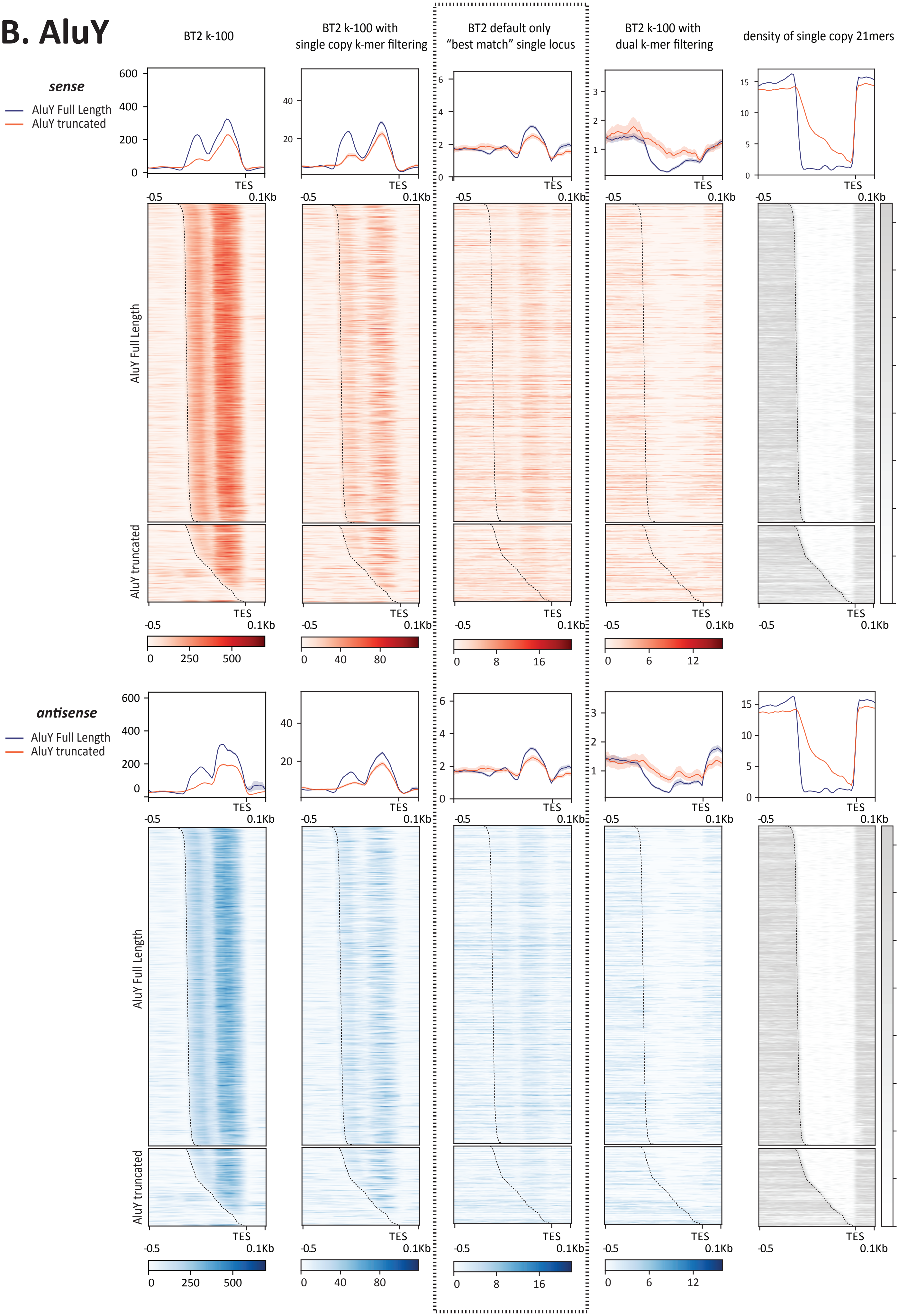


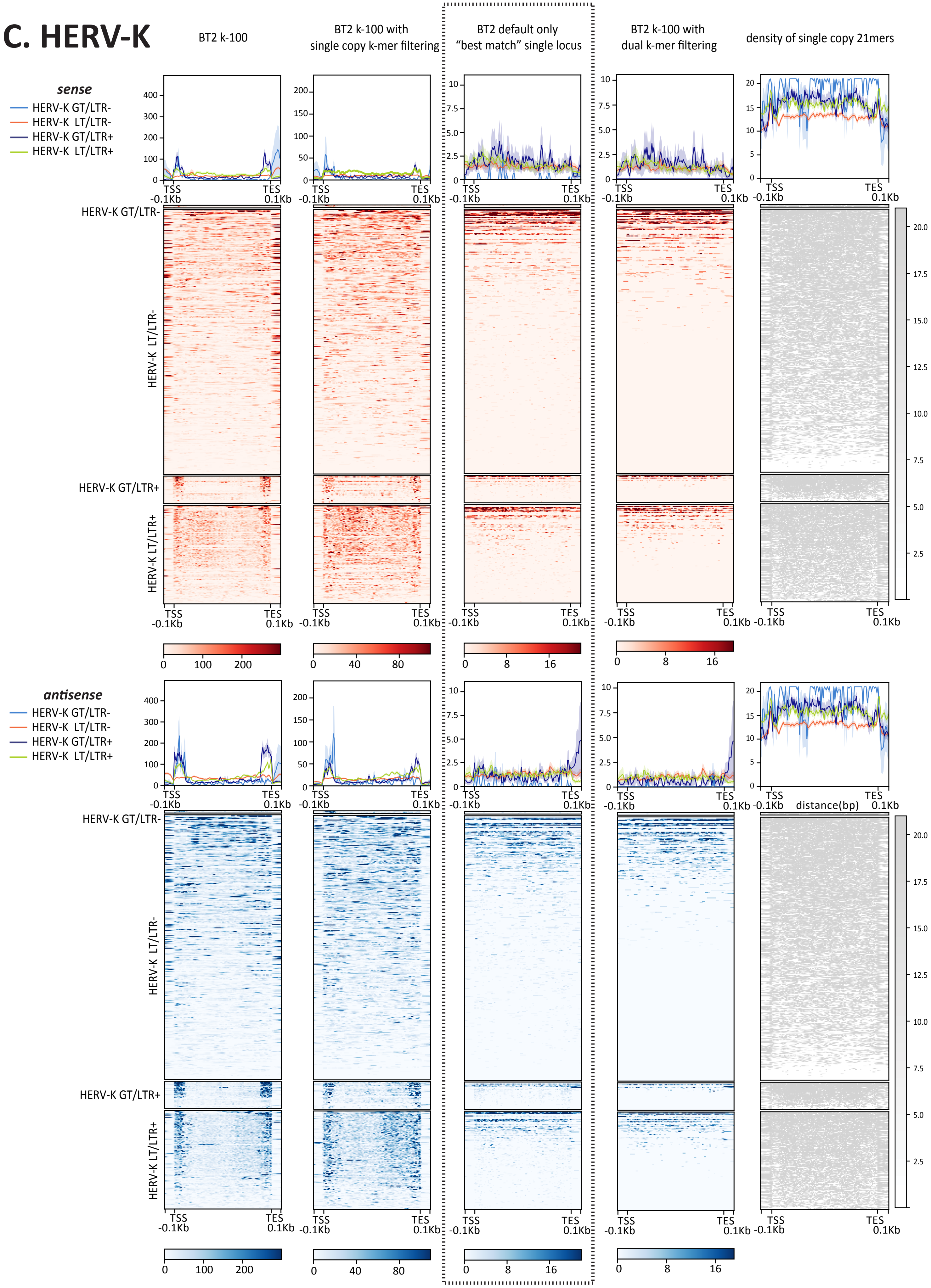
Fig. S17. Ideogram of density per 1Mbp bins of full-length retroelements in T2T-CHM13. (A) *Alu*Y, **(B)** L1Hs, **(C)** HERV-K, **(D)** SVA-E, **(E)** SVA-F. Centromere regions as per (*12*)) are shown in red. Density scale from low (blue, zero) to high (red, relative to total copy number).

A. AluY (by divergence)

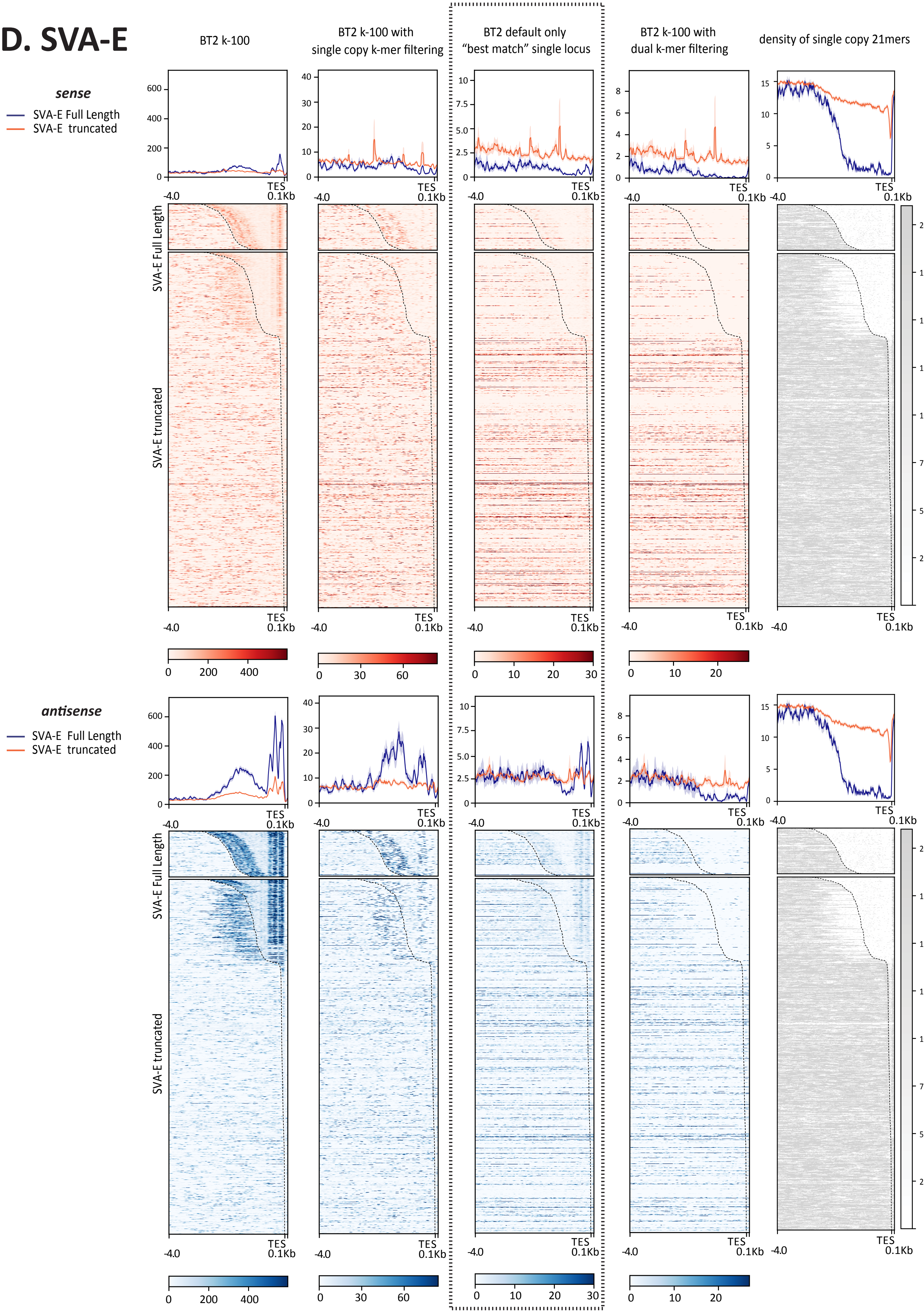




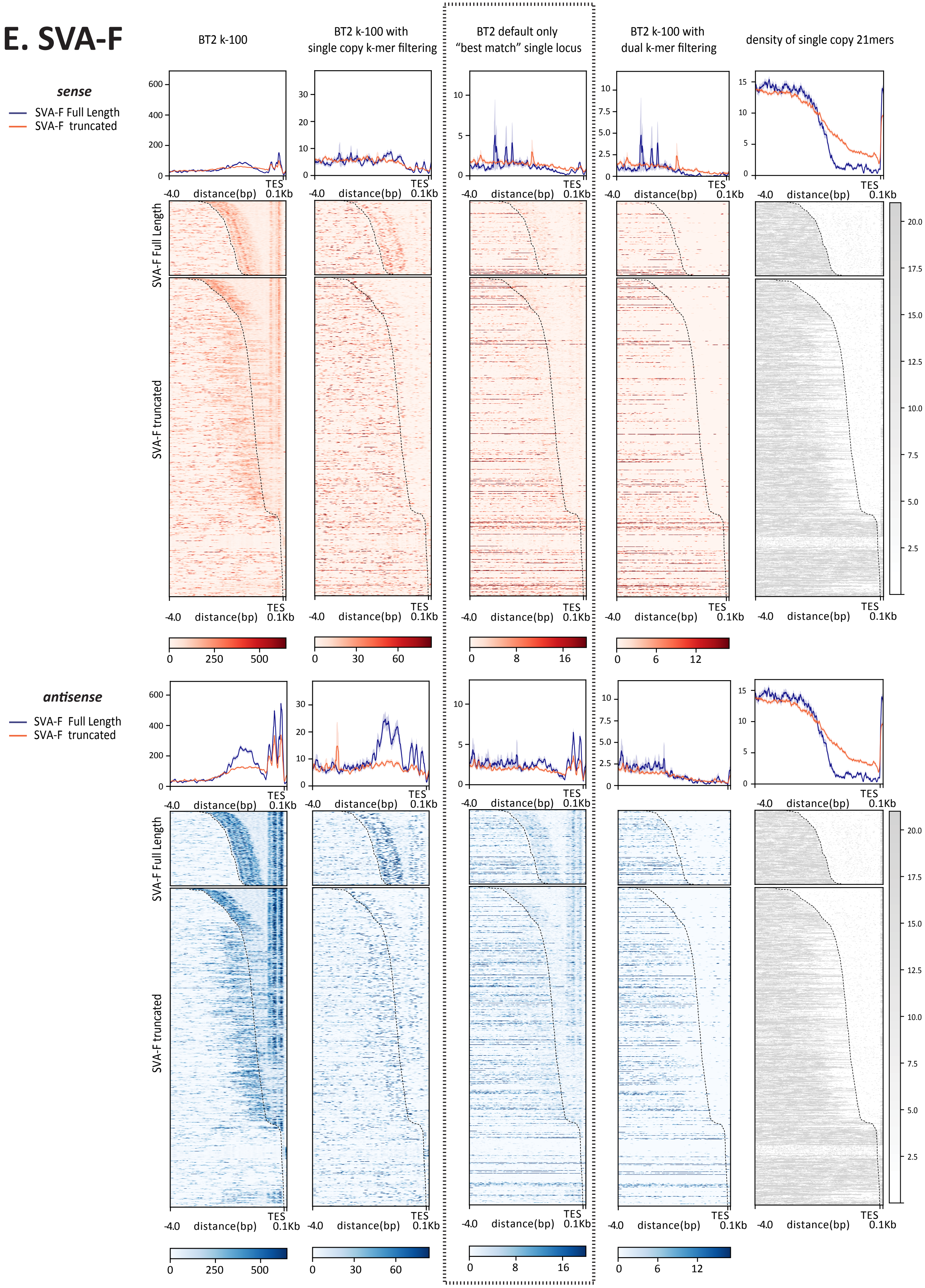
B. AluY
BT2 k-100
BT2 k-100 with single copy k-mer filtering
BT2 default only "best match" single locus
BT2 k-100 with dual k-mer filtering
density of single copy 21mers
sense
AluY Full Length
AluY truncated
600
400
200
0
40
20
6
4
2
0
3
2
1
15
10
5
0
TES
-0.5
0.1Kb
AluY Full Length
AluY truncated
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
250
500
0
40
80
0
8
16
0
6
12
antisense
AluY Full Length
AluY truncated
600
400
200
0
40
20
6
4
2
0
3
2
1
15
10
5
0
TES
-0.5
0.1Kb
AluY Full Length
AluY truncated
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
250
500
0
40
80
0
8
16
0
6
12


C. HERV-K
BT2 k-100
BT2 k-100 with single copy k-mer filtering
BT2 default only "best match" single locus
BT2 k-100 with dual k-mer filtering
density of single copy 21mers
sense
HERV-K GT/LTR-
HERV-K LT/LTR-
HERV-K GT/LTR+
HERV-K LT/LTR+
0
100
200
300
400
50
150
2.5
5.0
7.5
10
5
15
20
TSS
TES
-0.1Kb
0.1Kb
HERV-K GT/LTR-
HERV-K LT/LTR-
HERV-K GT/LTR+
HERV-K LT/LTR+
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
100
200
40
80
8
16
antisense
HERV-K GT/LTR-
HERV-K LT/LTR-
HERV-K GT/LTR+
HERV-K LT/LTR+
distance(bp)

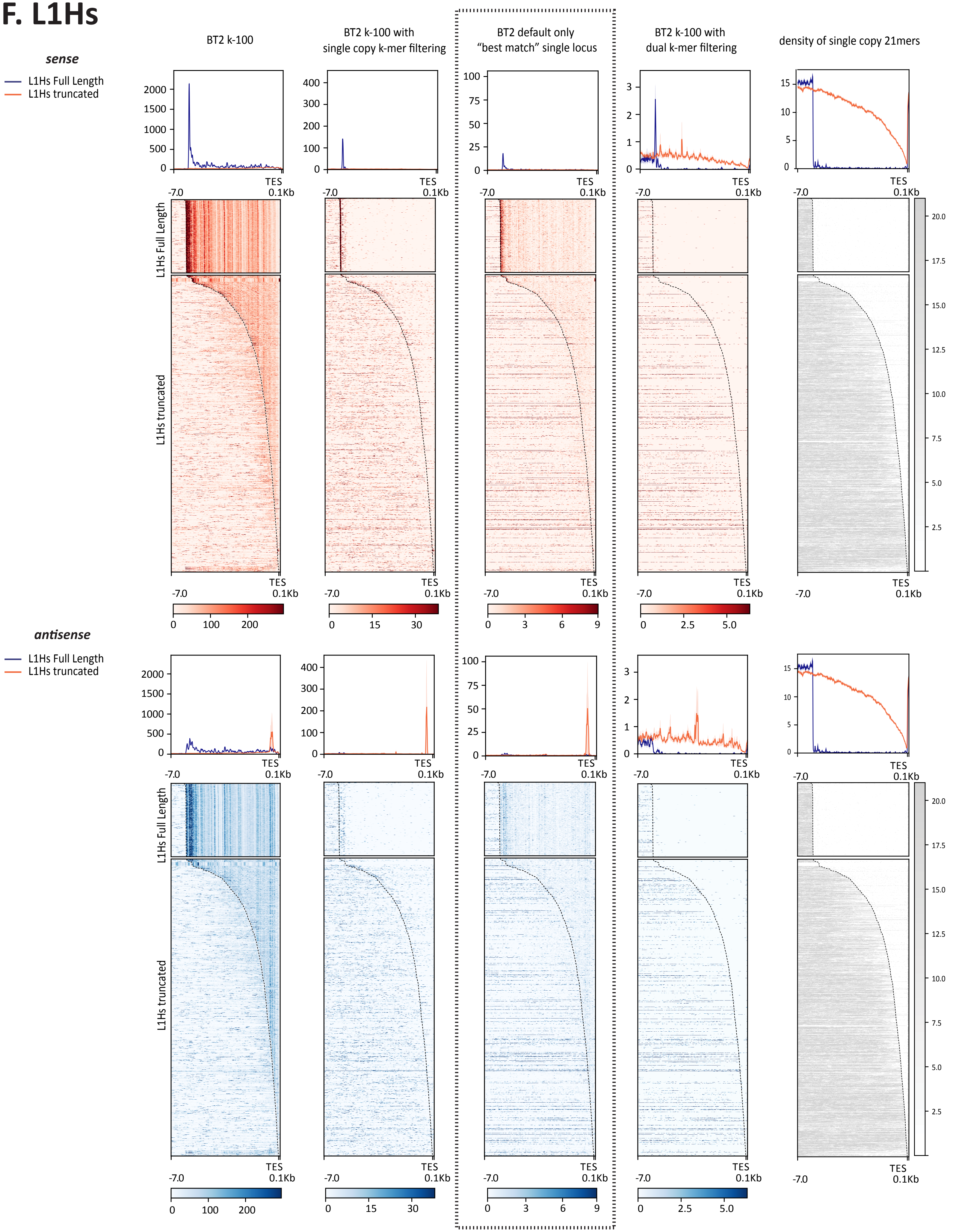
D. SVA-E



BT2 k-100
BT2 k-100 with single copy k-mer filtering
BT2 default only "best match" single locus
BT2 k-100 with dual k-mer filtering
density of single copy 21mers
sense
SVA-E Full Length
SVA-E truncated
antisense
SVA-E Full Length
SVA-E truncated
SVA-E Full Length
SVA-E truncated
TES
-4.0
0.1Kb
600
400
200
0
40
30
20
10
7.5
5.0
2.5
8
6
4
2
15
10
5
0 200 400
0 30 60
0 10 20 30
0 10 20
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5


E. SVA-F
BT2 k-100
BT2 k-100 with single copy k-mer filtering
BT2 default only "best match" single locus
BT2 k-100 with dual k-mer filtering
density of single copy 21mers
sense
SVA-F Full Length
SVA-F truncated
antisense
SVA-F Full Length
SVA-F truncated
SVA-F Full Length
SVA-F truncated
-4.0
distance(bp)
TES
0.1Kb
0
200
400
600
10
20
30
5
7.5
2.5
5.0
15
250
500
60
8
16
6
12
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5

F. L1Hs


BT2 k-100
BT2 k-100 with single copy k-mer filtering
BT2 default only "best match" single locus
BT2 k-100 with dual k-mer filtering
density of single copy 21mers
sense
L1Hs Full Length
L1Hs truncated
2000
1500
1000
500
0
400
300
200
100
0
100
75
50
25
0
3
2
1
0
15
10
5
0
TES
-7.0
0.1Kb
L1Hs Full Length
L1Hs truncated
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
100
200
0
15
30
0
3
6
9
0
2.5
5.0
antisense
L1Hs Full Length
L1Hs truncated

G. L1P

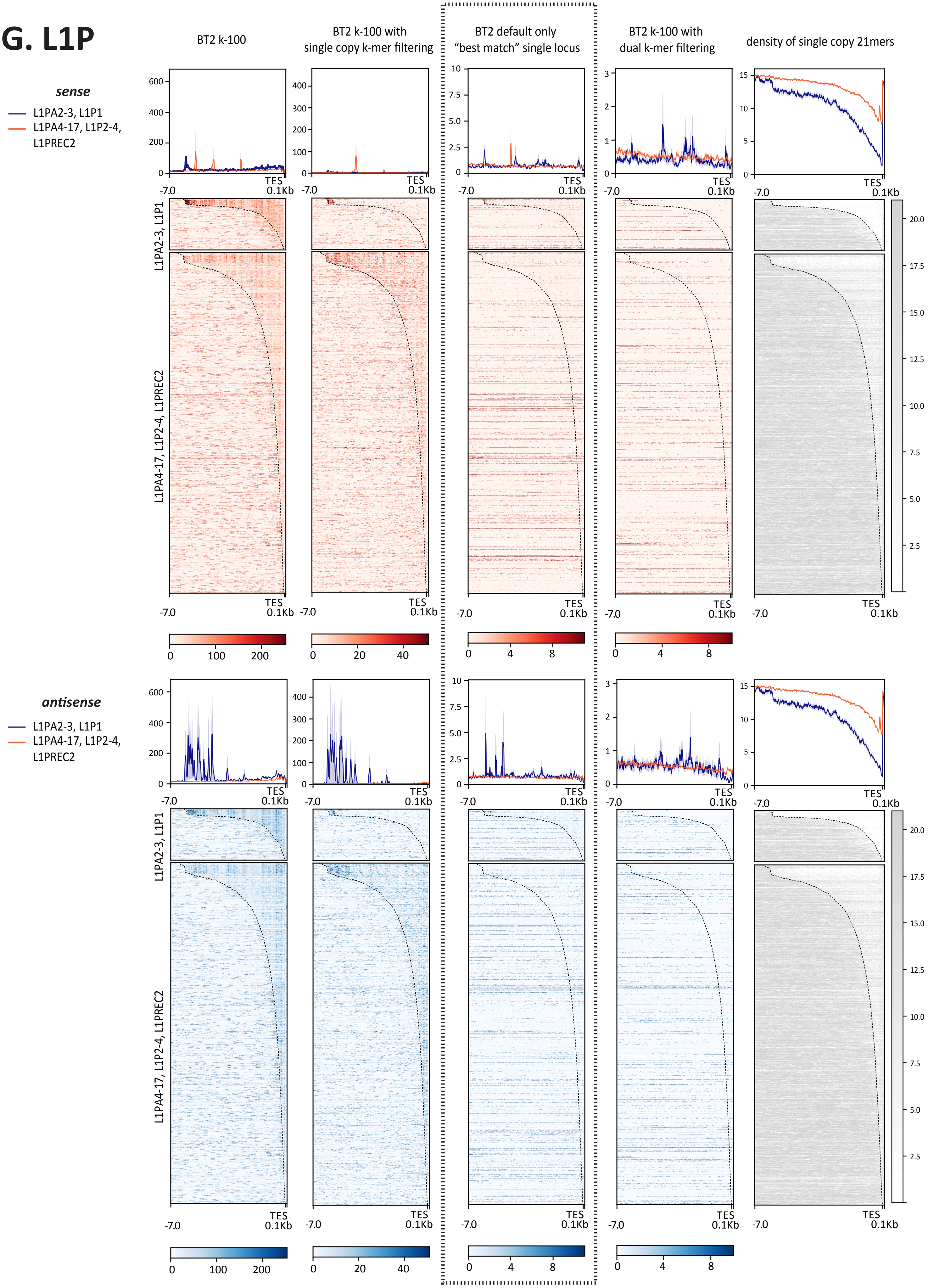

BT2 k-100
BT2 k-100 with single copy k-mer filtering
BT2 default only "best match" single locus
BT2 k-100 with dual k-mer filtering
density of single copy 21mers
sense
L1PA2-3, L1P1
L1PA4-17, L1P2-4, L1PREC2
antisense
L1PA2-3, L1P1
L1PA4-17, L1P2-4, L1PREC2
600
400
200
0
300
100
10
7.5
5.0
2.5
3
2
1
15
5
TES
-7.0
0.1Kb
L1PA2-3, L1P1
L1PA4-17, L1P2-4, L1PREC2
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
100
200
20
40
4
8

Fig. S18. Stranded PRO-seq profiles for (A-B) *Alu*Y, **(C)** HERV-K, **(D)** SVA-E, **(E)** SVA-F, **(F)** L1Hs and **(G)** L1P subfamilies. All elements are subdivided into full-length and truncated elements, with the following exceptions. **(C)** HERV-K is subdivided further into >7500 bp with both LTRs (GT/LTR+), <7500 bp with both LTRs (LT/LTR+), and elements in each category lacking one or both LTRs (LTR-). **(A-B)** *Alu*Y is subdivided based on divergence from the *Alu*Y (> or <2% diverged from RepeatMasker consensus (A) and full length vs. truncated **(B)**). CHM13 PRO-seq density for antisense (blue) and sense (red)), and average profiles (top line graphs, separated into sense and antisense read density, grey shaded portions are standard error) for TE subfamilies are shown. HERV-K elements are scaled to a fixed size, while all others are anchored to the 3' end, with a specified distance (bottom left) into the element; standard error shading (grey), TSS (transcription start site), TES (transcription end site), and ±0.1Kbp (bottom) are shown. A dotted line is included on the heatmap denoting the static -0.1Kbp from the end of the annotated element. Relative location of the VNTR in SVA elements is indicated. Mapping methods from left: Bowtie2 k-100; Bowtie2 k-100 filtered for single copy 21-mer k-mers based on locations; Bowtie2 default end-to-end, "best match"; Bowtie2 k100 dual filtered for single copy 21-mers in locations and reads. Far right, single copy 21-mer density. Boxed is the Bowtie2 "best match" panel used in Figure 2.

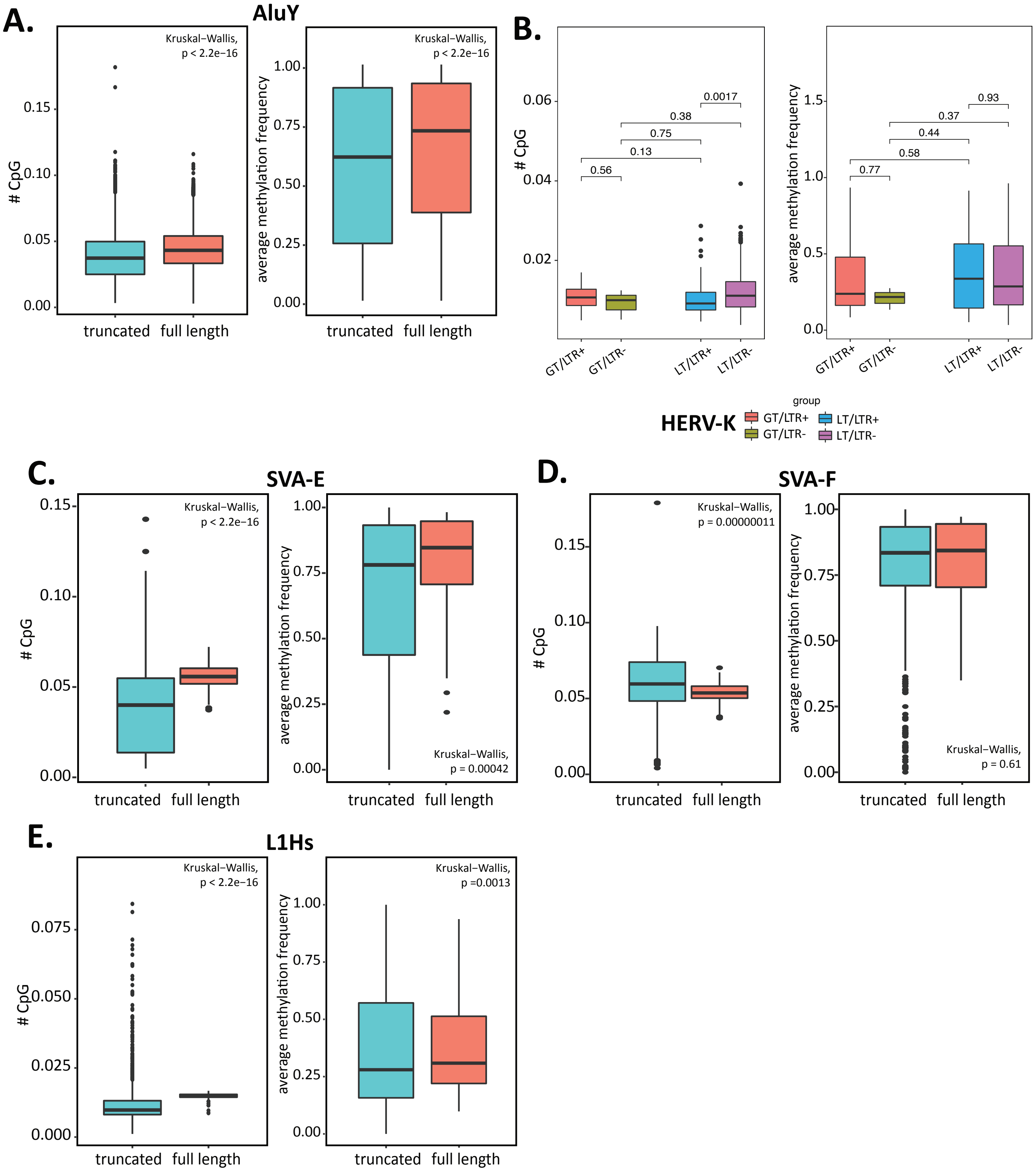
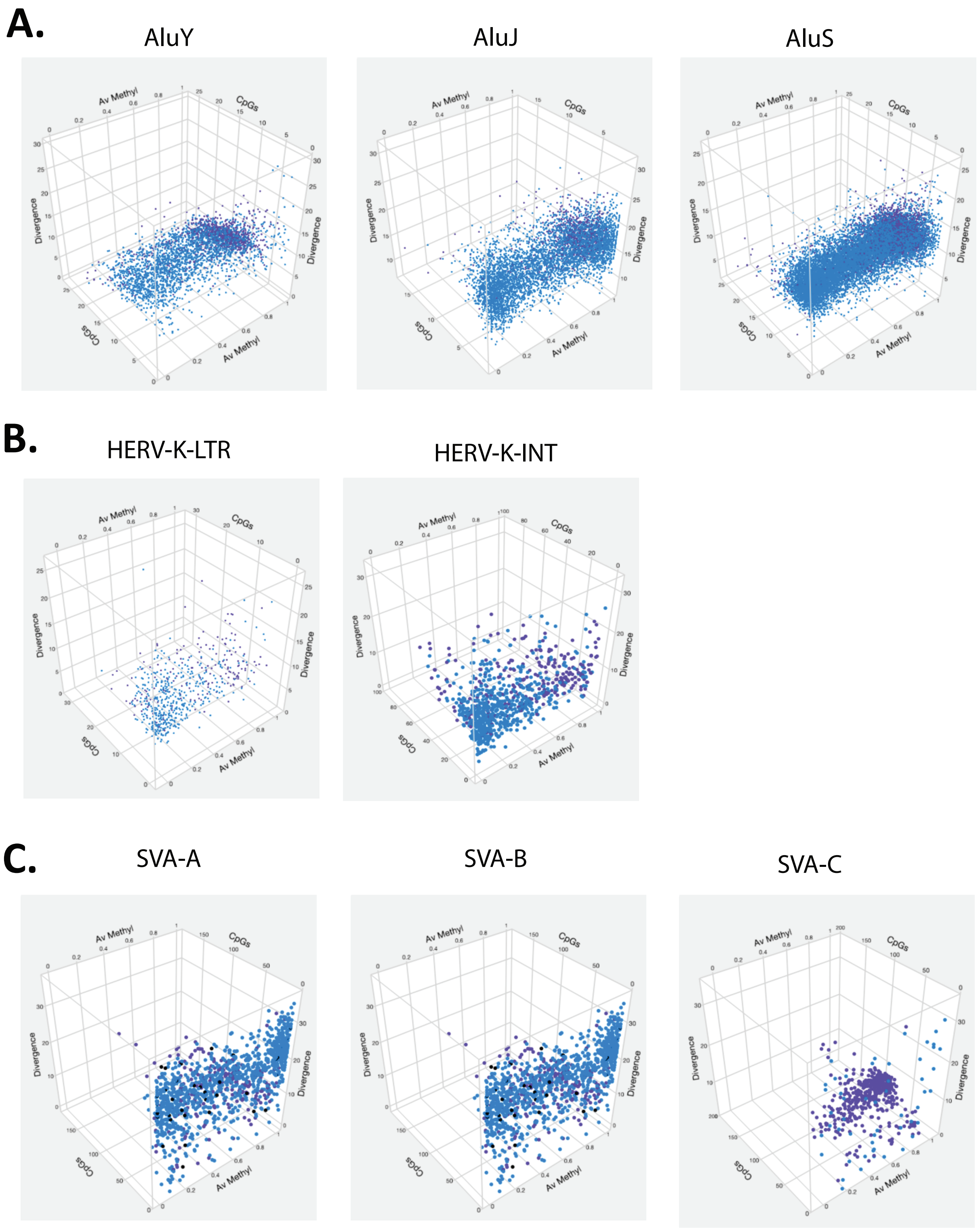


Fig. S19. Methylation and CpG density boxplot comparisons of full-length and truncated (A) *Alu*Y, **(C)** SVA-E, **(D)** SVA-F, **(E)** L1Hs elements in T2T-CHM13. **(B)** HERV-K is shown as a comparison between one of four structural groups. Methylation frequency was calculated as average methylation per repeat element and CpG density was calculated as number of CpGs normalized to total repeat length. Statistically significant differences were calculated with Kruskal-wallis one-way analysis of variance.


A.
AluY
AluJ
AluS
Av Methyl
CpGs
Divergence
B.
HERV-K-LTR
HERV-K-INT
Av Methyl
CpGs
Divergence
C.
SVA-A
SVA-B
SVA-C
Av Methyl
CpGs
Divergence

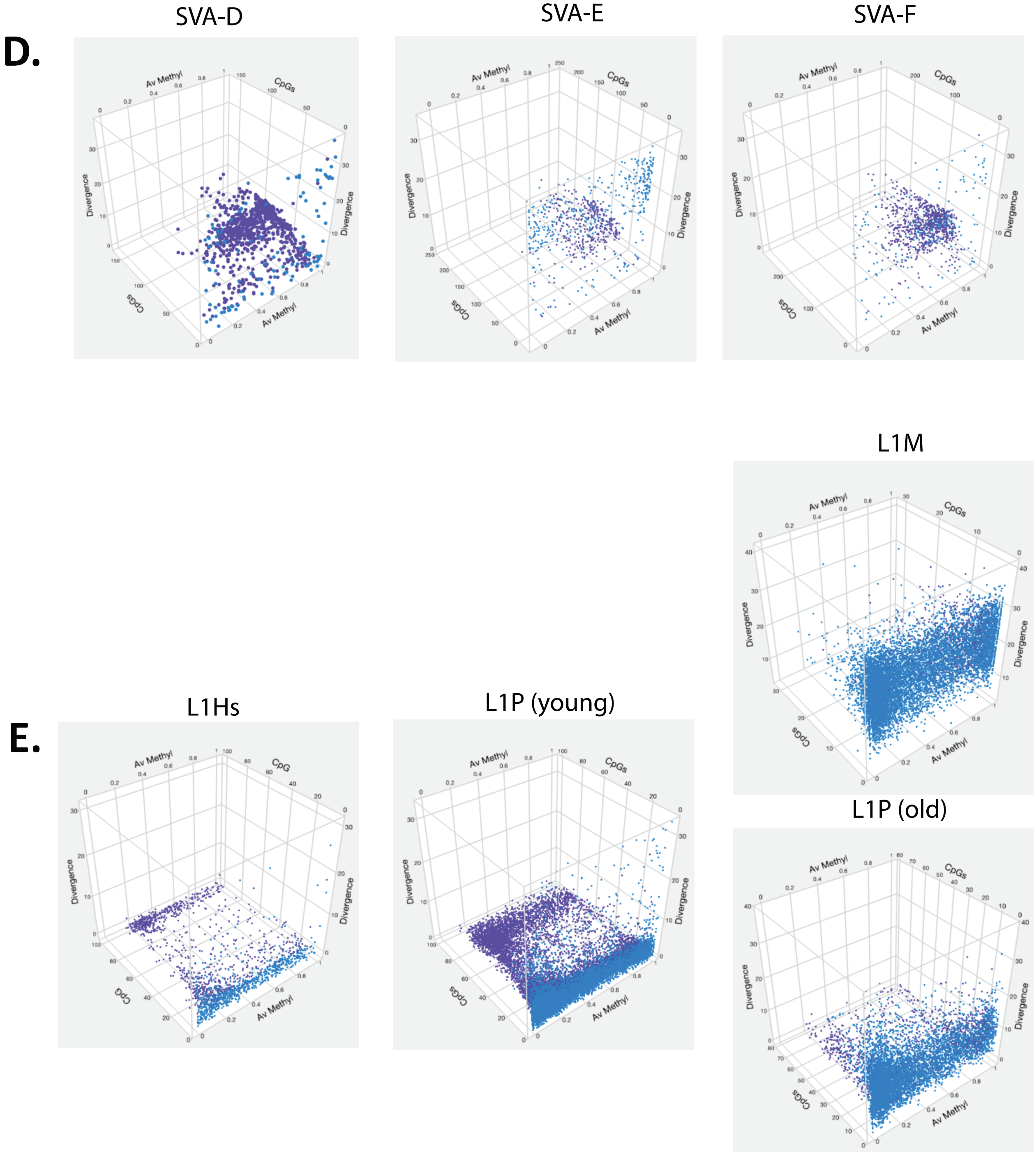
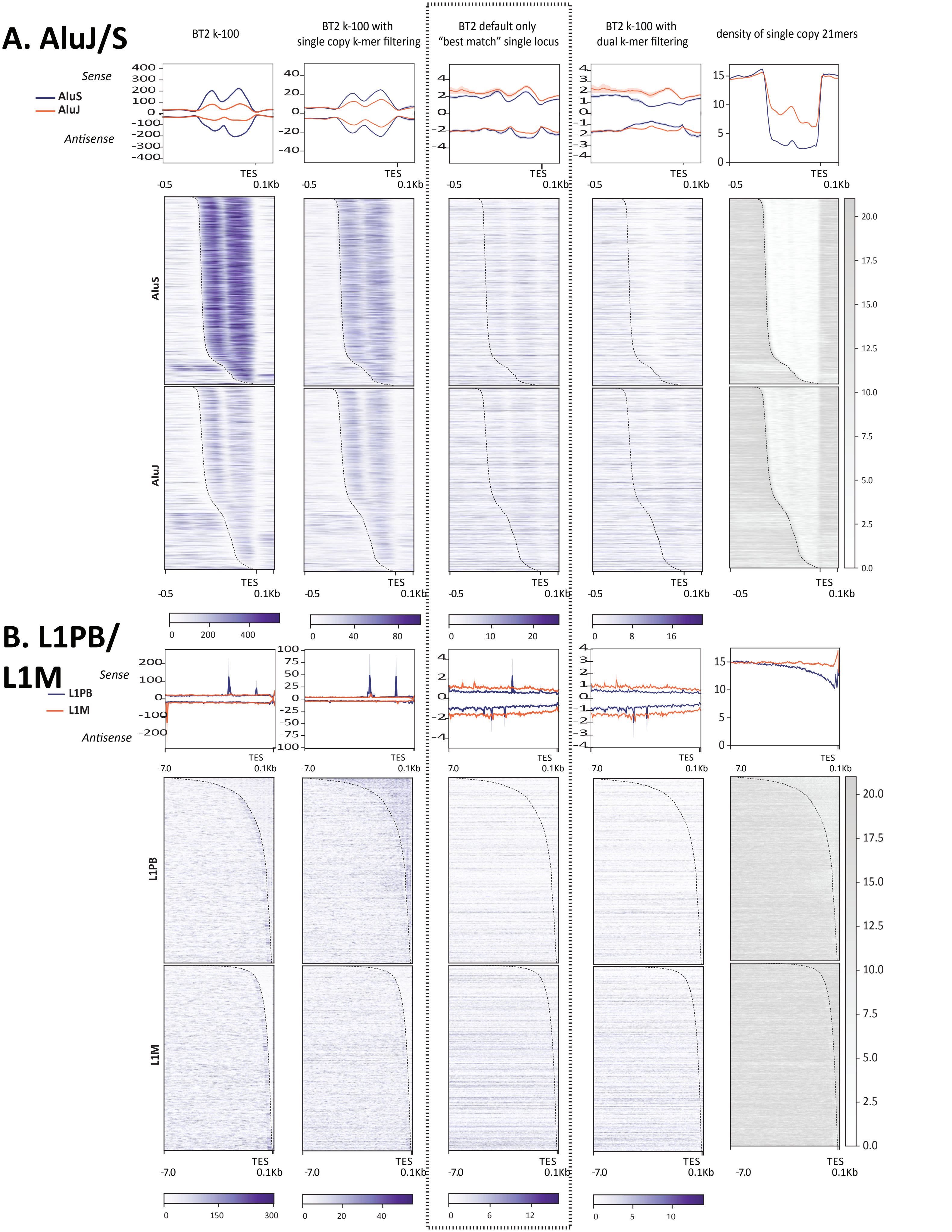


Fig. S20. 3D plots for (A) *Alu*Y, *Alu*J, *Alu*S, **(B)** HERV-K (divided into LTR and internal regions across all four structural groups), **(C)** SVA-A-C, **(D)** SVA-D-F, and **(E)** L1Hs, L1P (young and old), L1M. Axes represent scaled values for average methylation, # of CpG sites, and divergence from RepeatMasker consensus sequences for each instance of the element. Coloration by the number of overlapping PRO-seq reads (mapped with BT2 default parameters (”best match”), where purple represents the highest read overlap and blue the lowest, on the scale matching each plot. These data are the same for Figure 2A-E.


A. AluJ/S
BT2 k-100
BT2 k-100 with single copy k-mer filtering
BT2 default only "best match" single locus
BT2 k-100 with dual k-mer filtering
density of single copy 21mers
Sense
Antisense
AluS
AluJ
400
300
200
100
0
-100
-200
-300
-400
40
20
0
-20
-40
4
2
0
-2
-4
4
3
2
1
0
-1
-2
-3
-4
15
10
5
0
TES
-0.5
0.1Kb
AluS
AluJ
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
200
400
0
40
80
0
10
20
0
8
16
B. L1PB/L1M
Sense
Antisense
L1PB
L1M
200
100
0
-100
-200
100
75
50
25
0
-25
-50
-75
-100
-7.0
0.1Kb
TES
L1PB
L1M
0
150
300
0
20
40
0
6
12
0
5
10

C. SVA-A-D

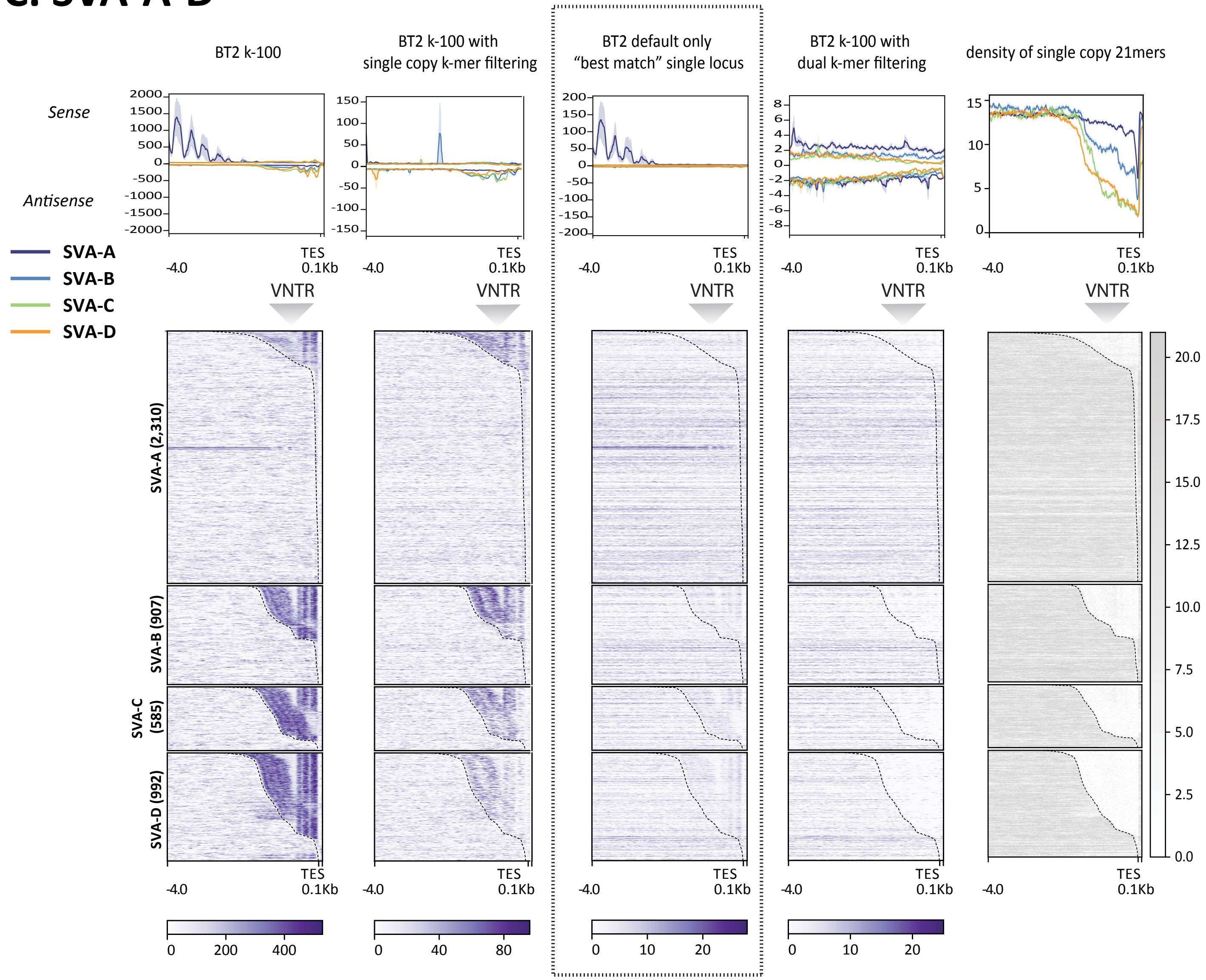


Fig. S21. PRO-seq profiles for (A) *Alu*, **(B)** L1, and **(C)** SVA subfamilies. T2T-CHM13 PRO-seq density (purple scale, reads per million both sense and antisense aggregated) and average profiles (top line graphs, separated into sense and antisense read density) for TE subfamilies. All elements are anchored to the 3' end, with a specified distance from the anchor (bottom left) into the element; standard error shading (70% opacity of respective line color), TSS (transcription start site), TES (transcription end site), and ±Kbp are shown. A dotted line is included on the heatmap denoting the starting nt of each annotated element. Relative location of the VNTR in SVA elements is indicated. Number of SVA elements in each subfamily shown to the right of each panel. Mapping methods (from left) are Bowtie2 (BT2) k-100; BT2 k-100 21nt k-mer filtered (locus level); BT2 default only ("best match" single locus); BT2 k-100 dual 21nt k-mer filtered (read and locus filtered). BT2 default only are within a dotted box. Far right is the density of single copy 21nt k-mers across each element (grey scale, number of k-mers in sense and antisense aggregated) and average profiles (top line graphs) for each TE subfamily.

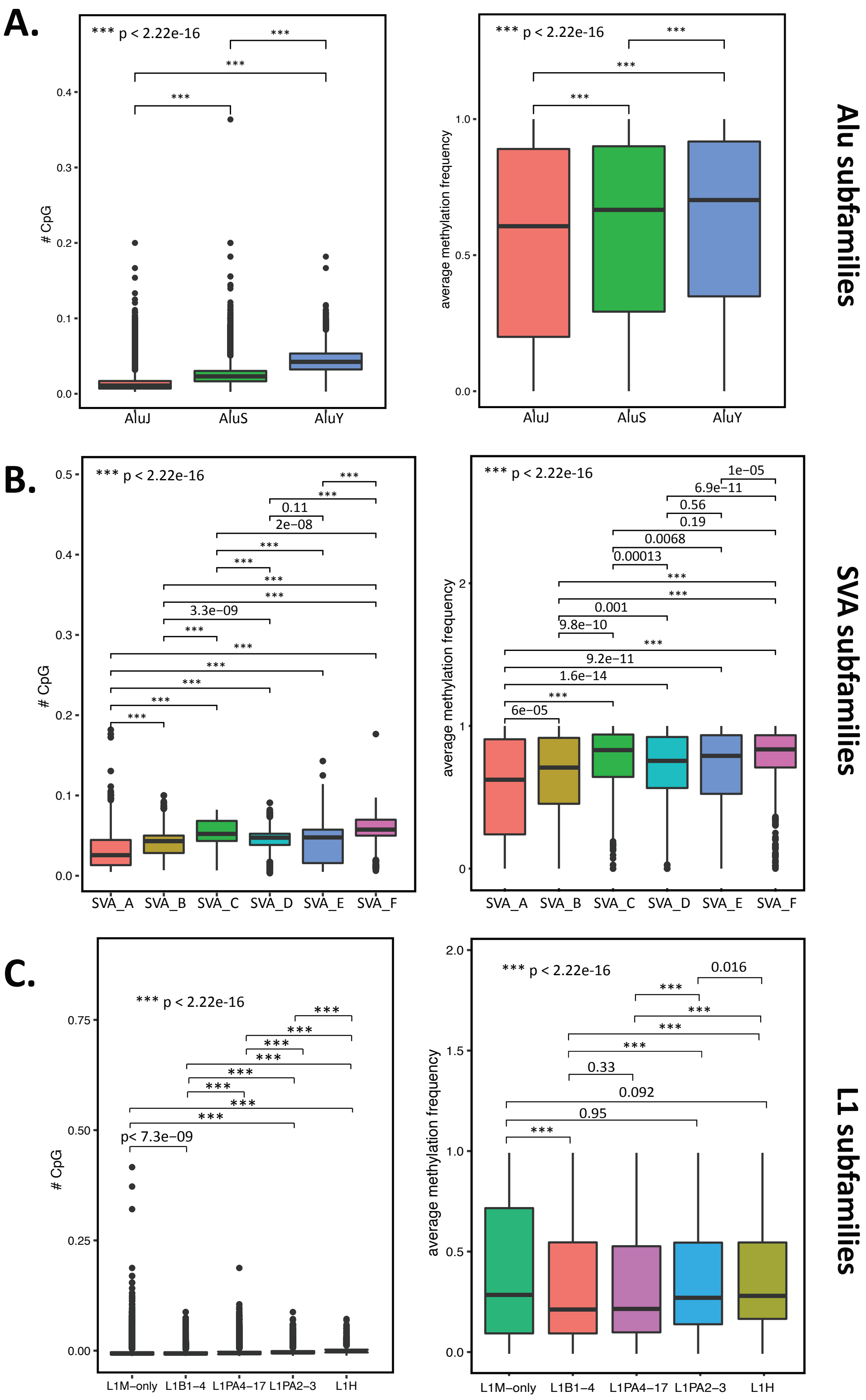


Fig. S22. Methylation and CpG density boxplots for (A) *Alu* subfamilies, **(B)** SVA subfamilies, **(C)** L1 subfamilies. Methylation frequency was calculated as average methylation per repeat element and CpG density was calculated as number of CpGs normalized to total repeat length. Statistically significant differences were calculated with Kruskal-wallis one-way analysis of variance.

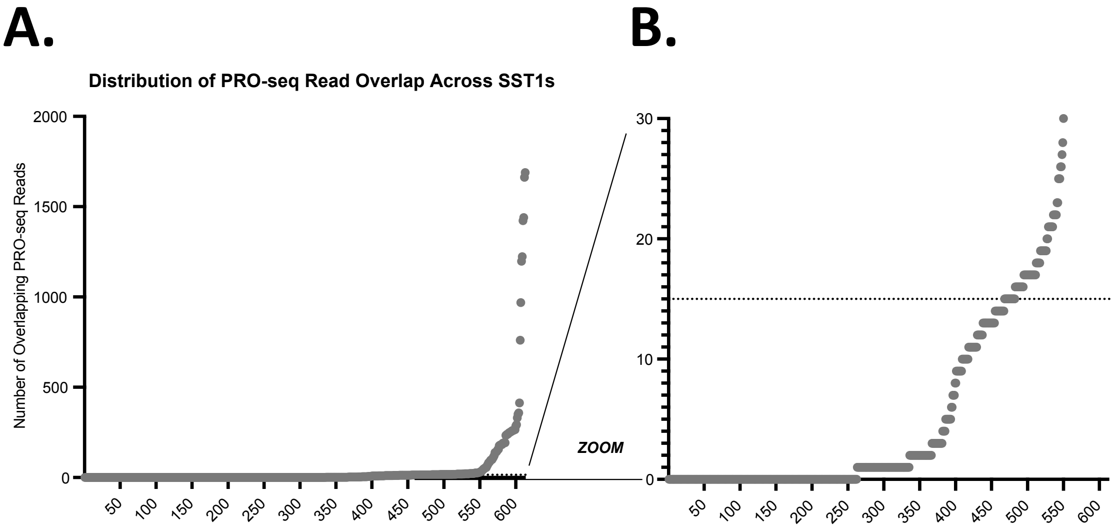
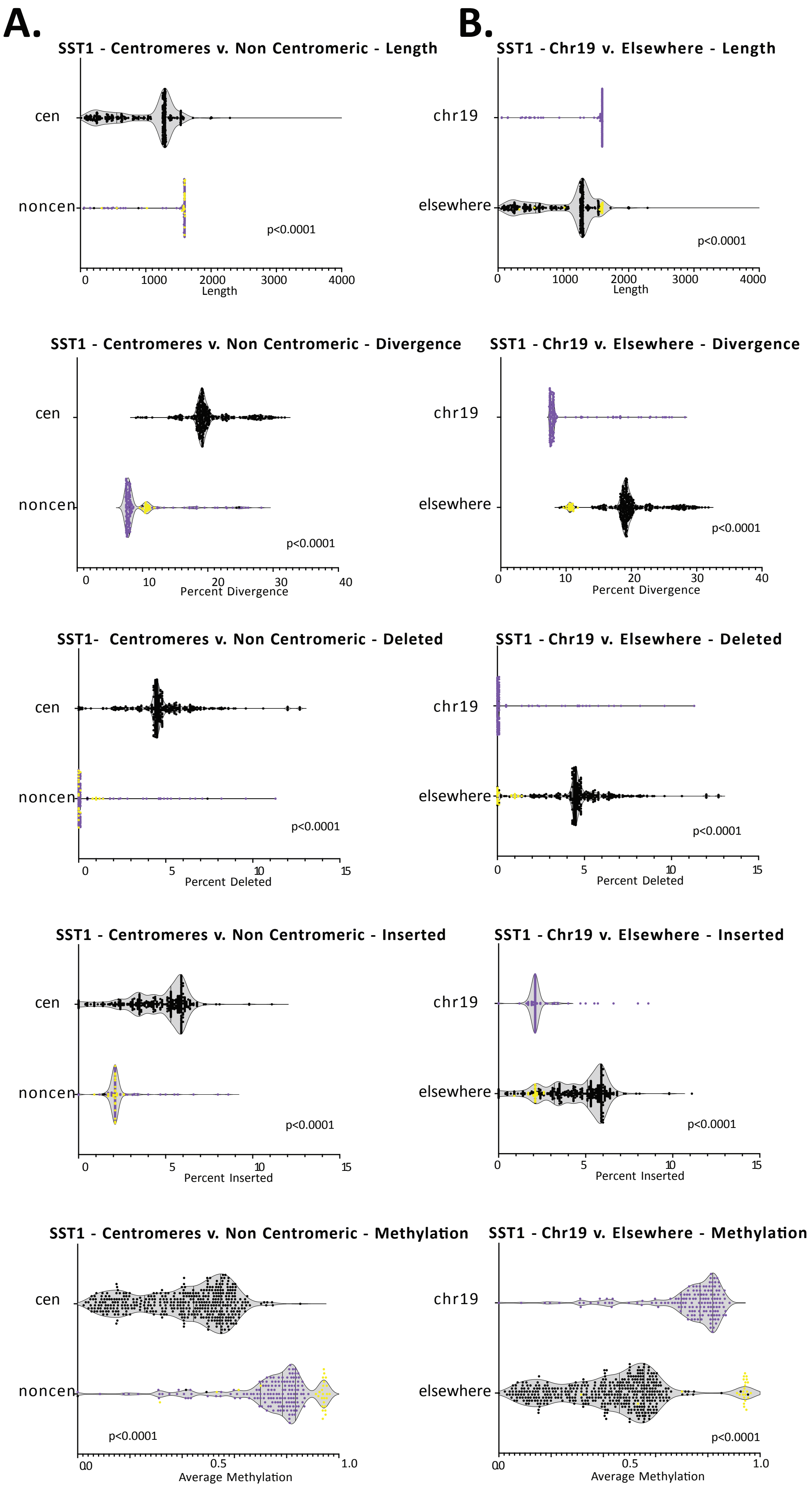


Fig S23. Distribution of PRO-seq read overlap counts over all annotated SST1 repeats. (A) SST1s with more than 15 overlapping PRO-seq reads were grouped (dotted line) and represent purple and yellow points in Fig. 3C and fig. S24. **(B)** Zoom of **(A)**; 15 overlapping PRO-seq reads (dotted line) was used as a cutoff for statistical comparisons


A.
SST1 - Centromeres v. Non Centromeric - Length
cen
noncen
p<0.0001
0
1000
2000
3000
4000
Length
B.
SST1 - Chr19 v. Elsewhere - Length
chr19
elsewhere
p<0.0001
0
1000
2000
3000
4000
Length
SST1 - Centromeres v. Non Centromeric - Divergence
cen
noncen
p<0.0001
0
10
20
30
40
Percent Divergence
SST1 - Chr19 v. Elsewhere - Divergence
chr19
elsewhere
p<0.0001
0
10
20
30
40
Percent Divergence
SST1- Centromeres v. Non Centromeric - Deleted
cen
noncen
p<0.0001
0
5
10
15
Percent Deleted
SST1 - Chr19 v. Elsewhere - Deleted
chr19
elsewhere
p<0.0001
0
5
10
15
Percent Deleted
SST1 - Centromeres v. Non Centromeric - Inserted
cen
noncen
p<0.0001
0
5
10
15
Percent Inserted
SST1 - Chr19 v. Elsewhere - Inserted
chr19
elsewhere
p<0.0001
0
5
10
15
Percent Inserted
SST1 - Centromeres v. Non Centromeric - Methylation
cen
noncen
p<0.0001
0.0
0.5
1.0
Average Methylation
SST1 - Chr19 v. Elsewhere - Methylation
chr19
elsewhere
p<0.0001
0.0
0.5
1.0
Average Methylation

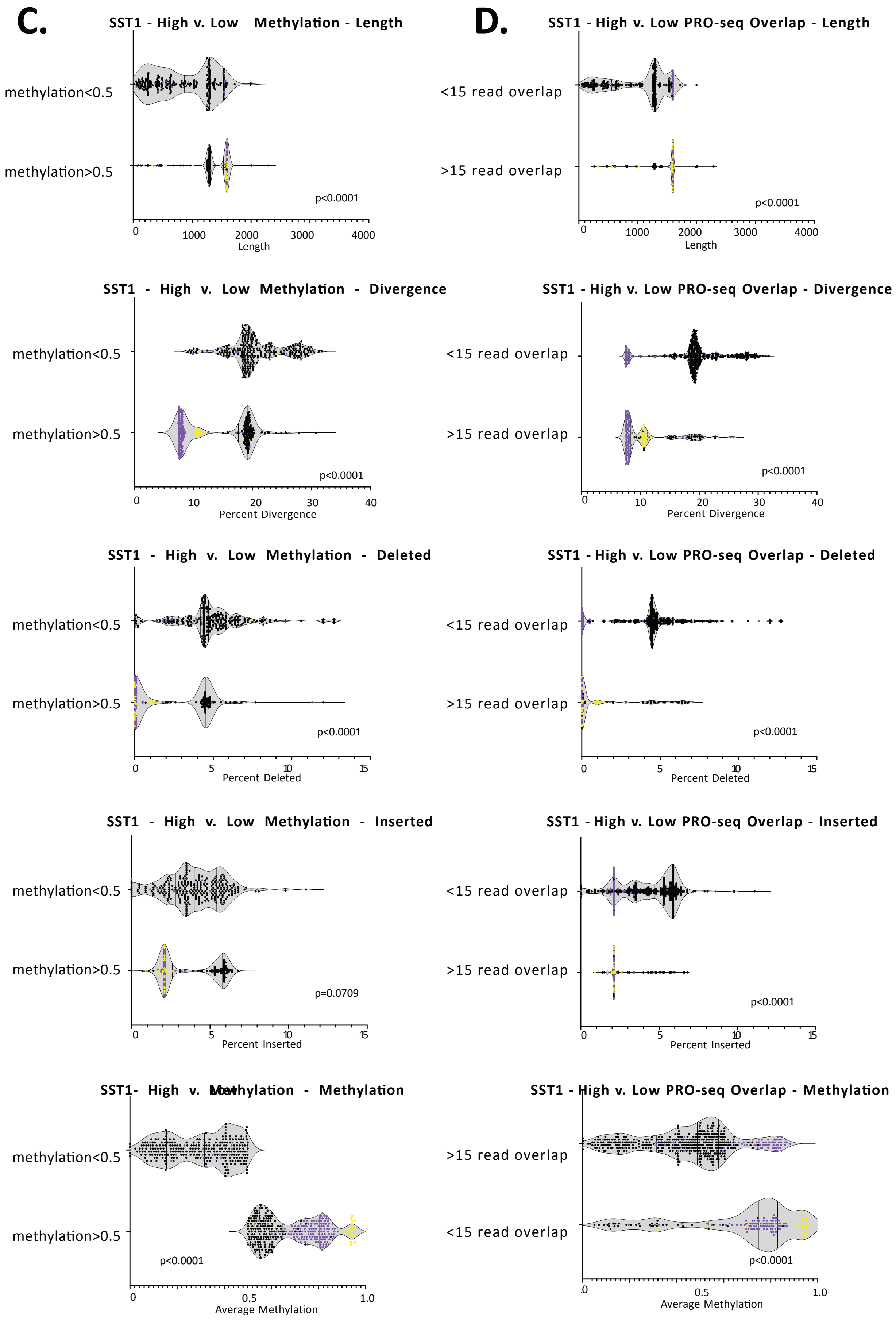

C.
SST1 - High v. Low Methylation - Length
methylation<0.5
methylation>0.5
p<0.0001
0
1000
2000
3000
4000
Length
D.
SST1 - High v. Low PRO-seq Overlap - Length
<15 read overlap
>15 read overlap
p<0.0001
0
1000
2000
3000
4000
Length
SST1 - High v. Low Methylation - Divergence
methylation<0.5
methylation>0.5
p<0.0001
0
10
20
30
40
Percent Divergence
SST1 - High v. Low PRO-seq Overlap - Divergence
<15 read overlap
>15 read overlap
p<0.0001
0
10
20
30
40
Percent Divergence
SST1 - High v. Low Methylation - Deleted
methylation<0.5
methylation>0.5
p<0.0001
0
5
10
15
Percent Deleted
SST1 - High v. Low PRO-seq Overlap - Deleted
<15 read overlap
>15 read overlap
p<0.0001
0
5
10
15
Percent Deleted
SST1 - High v. Low Methylation - Inserted
methylation<0.5
methylation>0.5
p=0.0709
0
5
10
15
Percent Inserted
SST1 - High v. Low PRO-seq Overlap - Inserted
<15 read overlap
>15 read overlap
p<0.0001
0
5
10
15
Percent Inserted
SST1- High v. Low Methylation - Methylation
methylation<0.5
methylation>0.5
p<0.0001
0
0.5
1.0
Average Methylation
SST1 - High v. Low PRO-seq Overlap - Methylation
>15 read overlap
<15 read overlap
p<0.0001
.0
0.5
1.0
Average Methylation

Fig. S24. Statistical tests for SST1. Violin plots of SST1 elements show differences in length, divergence, deletions, insertions, and methylation of repeats found in the centromere **(A)**, on Chromosome 19 **(B)**, with varying PRO-seq expression levels **(C)** and with varying methylation patterns **(D)**. Dot colors represent interstitial arrays on Chromosome 19 (purple), and Chromosome 4 (yellow); all other SST1 repeats are colored black. Non-centromeric SST1s, particularly those on Chromosome 19, are longer, less diverged, and possess higher average methylation than those situated in the centromeres. Similarly, SST1s with > 0.5 methylation and > 15 PRO-seq reads are longer and less diverged than those with lower transcription and lower methylation. All differences are statistically significant (p<0.0001).

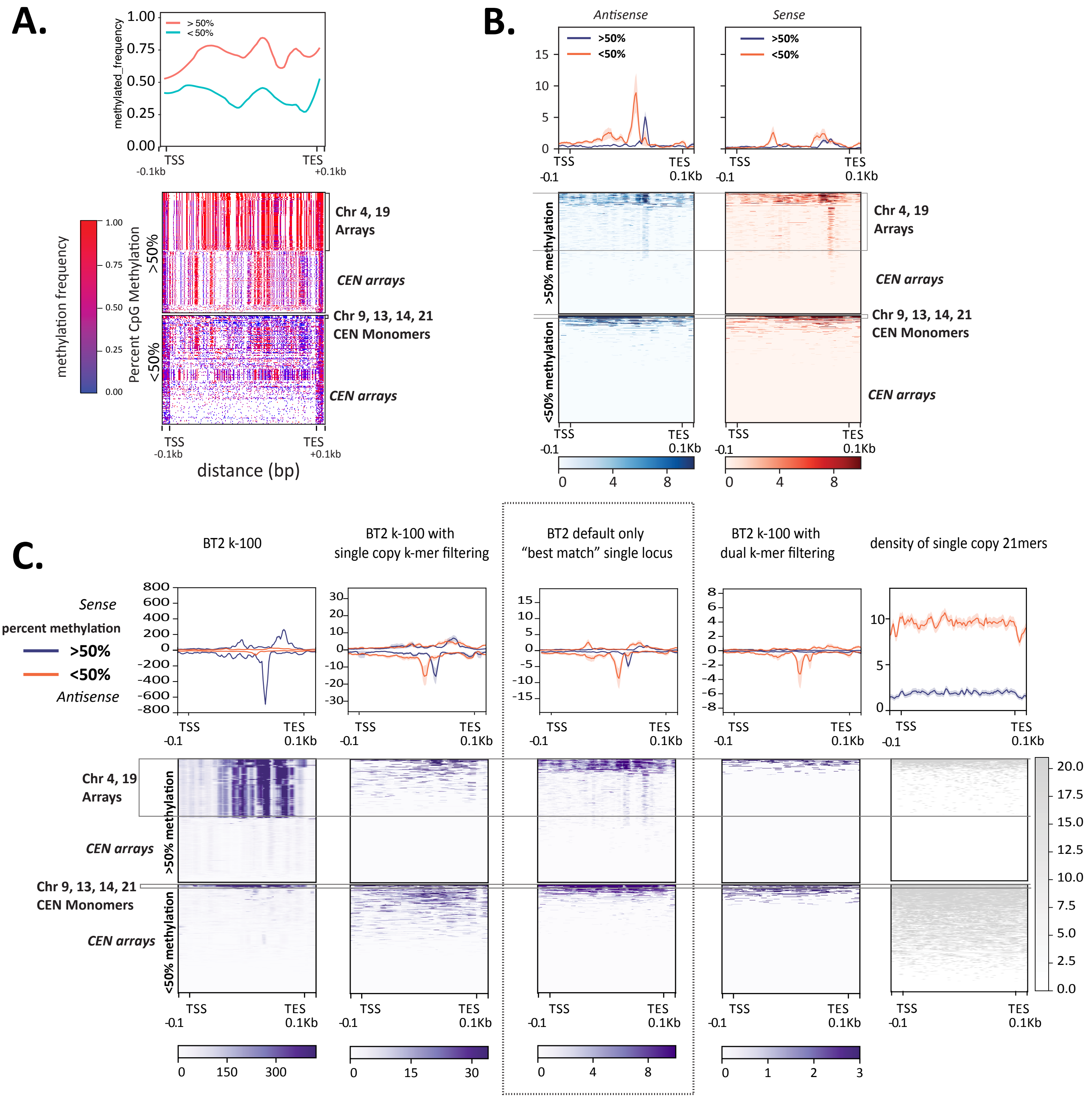


Fig. S25. T2T-CHM13 methylation and stranded PROseq profiles for SST1. (A) Methylation profiles for SST1 grouped by average methylation levels (>50% top, <50% bottom). Each element is scaled to a fixed size; TSS (transcription start site), TES (transcript end site), and ±0.1Kbp are shown. Clusters of specific SST1 loci are indicated to the right. Methylation frequency scale is on the left. **(B,C)** Heatmaps of PRO-seq density (heatmaps, reads per million) and average profile read density (top line graphs) grouped by average methylation levels (< and > 50%). Each element is scaled to a fixed size; standard error shading (70% opacity of respective line color), TSS (transcription start site), TES (transcription end site), and ±0.1Kbp are shown. **(B)** Heatmaps of PRO-seq density for BT2 default only (purple scale, reads per million both sense and antisense aggregated). **(C)** Heatmaps of PRO-seq density for all mapping methods, from left: Bowtie2 (BT2) k-100; BT2 k-100 21nt k-mer filtered (locus level); BT2 default only ("best match" single locus); BT2 k-100 dual 21nt k-mer filtered (read and locus filtered). BT2 default only corresponding to Fig. 3 are within a dotted box. Far right is the density of single copy 21nt k-mers across each element (grey scale, number of k-mers in sense and antisense aggregated) and average profiles (top line graphs).

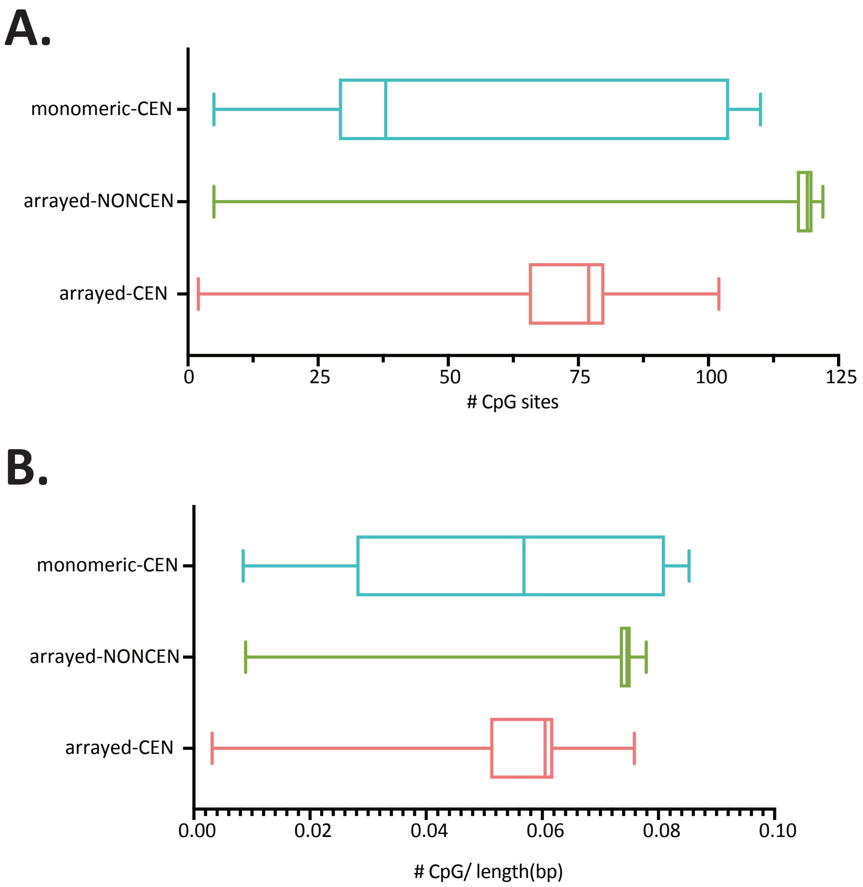


Fig. S26. SST1 CpG density. Box plots of SST1 repeats showing CG density distribution in **(A)** CpG density for SST1 repeats 500bp-2Kbp in length delineated by location and repeat density (centromeric (CEN) vs non-centromeric (NONCEN), monomeric vs arrayed). **(B)** CpG density for SST1 repeats 500bp-2Kbp in length, normalized by element length, delineated by location and density (monomeric vs arrayed, centromeric (CEN) vs non-centromeric (NONCEN).

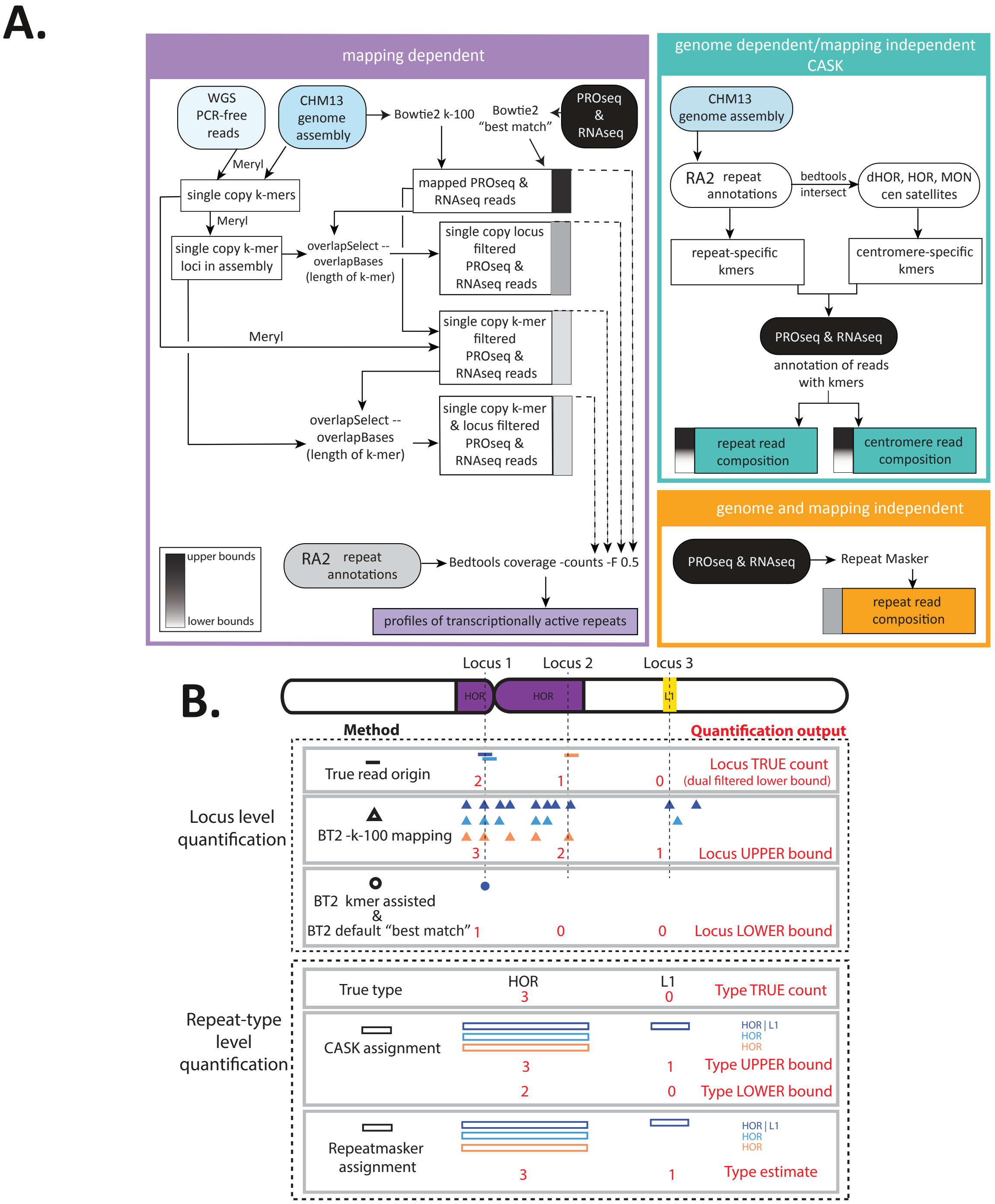
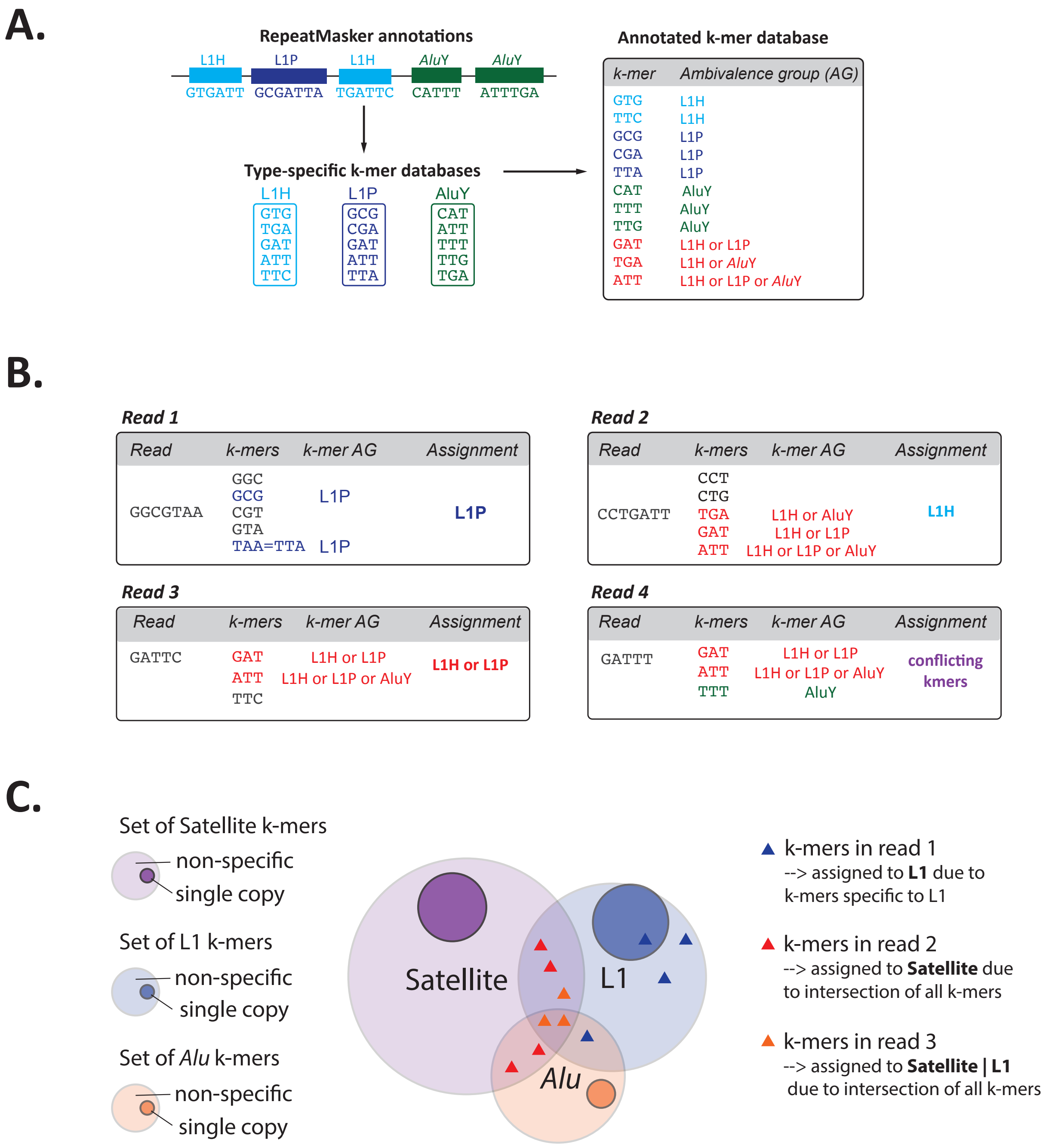


Fig. S27 (A). Overview of three-pronged repeat transcription pipeline. Two pipelines were developed to assess repeat transcription levels across the T2T-CHM13 genome, including 1) a mapping-dependent method relying on Bowtie2 and single copy k-mers (*11*) (purple) and 2) a mapping independent method (teal, CASK; see below for details and fig. S28). Both methods are reliant upon having a genome assembly. Repeatmasker (orange) was simultaneously used to determine repeat content in the reads as a genome and mapping-independent method as per (*43*). **(B)** Overview of methods used in this study to quantify expression of specific repeat elements. In this representation, 2 reads originating from the same locus L1 within a HOR region, and 1 read from another locus L2 within a HOR region are considered. Locus-level methods quantify expression across the genome, with all non-zero loci shown (L1, L2, L3). Bowtie2 k100 output can be used to obtain an upper bound ("over fit") expression estimate at each locus. Bowtie2 k-mer-assisted or Bowtie2 default "best match" provides a lower bound expression estimate at each locus, while read- and locus- filtered provide a strict dual filtered lower bounds. CASK and RM are repeat type-level quantification which provide lower and upper bound of expression (CASK), and best-estimate expression (RM) for all repeat types.



k-mer	*Ambivalence group (AG)*
GTG	L1H
TTC	L1H
GCG	L1P
CGA	L1P
TTA	L1P
CAT	AluY
TTT	AluY
TTG	AluY
GAT	L1H or L1P
TGA	L1H or *AluY*
ATT	L1H or L1P or *AluY*

Read 1

Read	*k-mers*	*k-mer AG*	*Assignment*
GGCGTAA	GGC		**L1P**
	GCG	L1P	
	CGT		
	GTA		
	TAA=TTA	L1P	

Read 2

Read	*k-mers*	*k-mer AG*	*Assignment*
CCTGATT	CCT		**L1H**
	CTG		
	TGA	L1H or AluY	
	GAT	L1H or L1P	
	ATT	L1H or L1P or AluY	

Read 3

Read	*k-mers*	*k-mer AG*	*Assignment*
GATTC	GAT	L1H or L1P	**L1H or L1P**
	ATT	L1H or L1P or AluY	
	TTC		

Read 4

Read	*k-mers*	*k-mer AG*	*Assignment*
GATTT	GAT	L1H or L1P	**conflicting kmers**
	ATT	L1H or L1P or AluY	
	TTT	AluY	

Fig. S28. The CASK (Classification of Ambivalent Sequences using k-mers) algorithm. (A) Main steps of the algorithm and **(B)** examples showing the repeat-type assignment for four reads with different k-mers composition. **(C)** Simplified schematic describing the CASK algorithm. The sets of k-mers found within each repeat type are shown as circles and contain both single copy k-mers (dark color) and non-specific k-mers (light color). Only sets for satellite (purple), L1 (blue) and *Alu* (orange) are shown for simplicity. These sets overlap as some k-mers are found across different repeat types, generating a partition of the k-mer space. Triangles indicate k-mers present in three example reads. For each read, repeat-type of origin is determined by taking the intersection of the subsets in the partition to which these k-mers belong. For example, for Read2, 2 k-mers belong to the subset {Satellite OR L1} and 2 k-mers to the subset {Satellite OR *Alu*}. The intersection of these subsets is {Satellite}.

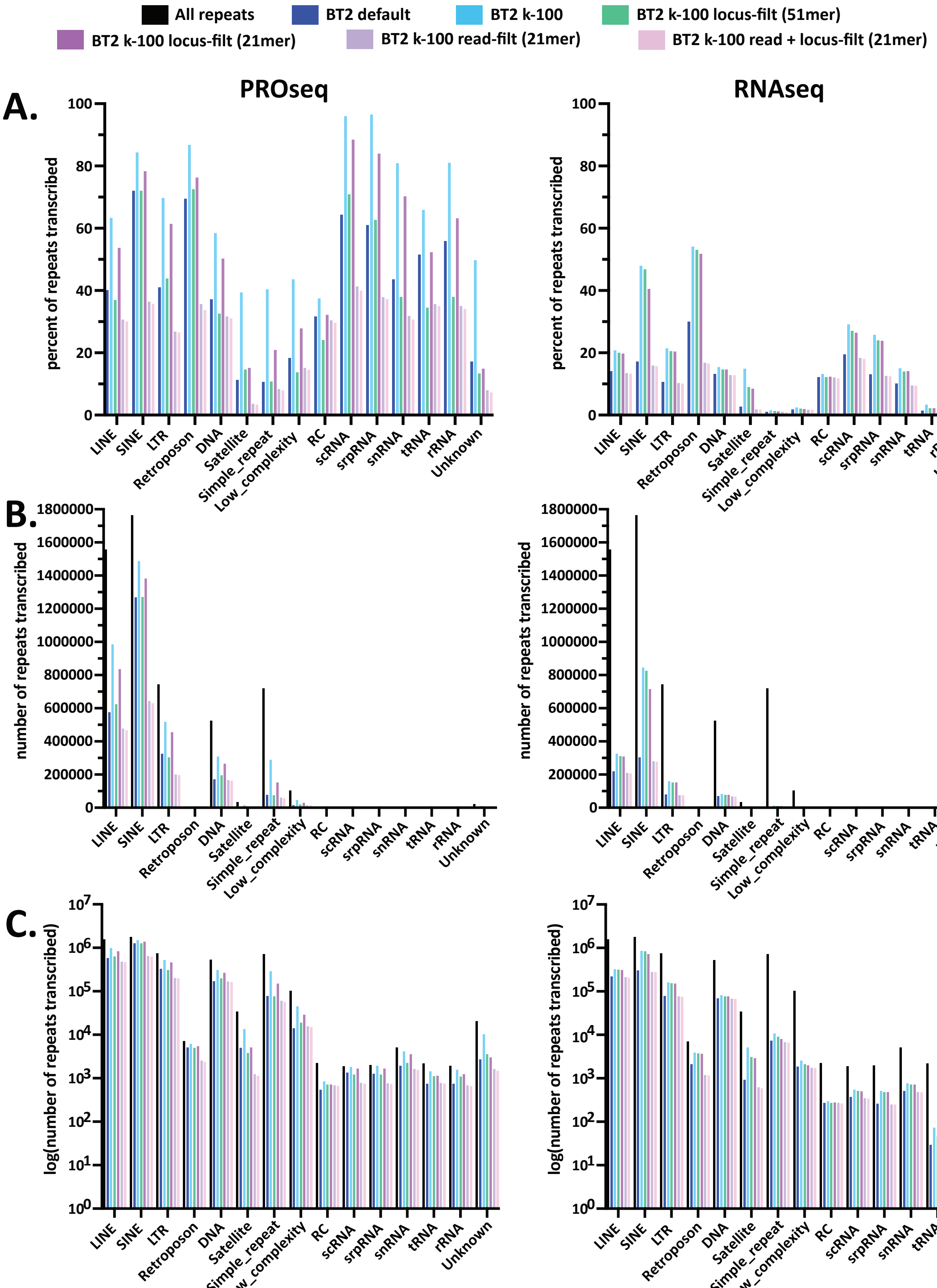


Fig. S29. Mapping of repeats assayed by PRO-seq and RNA-seq. Profile of transcriptionally active repeats across Bowtie2 mapping methods for PRO-seq (Left) and RNA-seq (right) data shown as **(A)** percent of repeats transcribed, **(B)** number of repeats transcribed, and **(C)** Log transformed number of repeats transcribed.

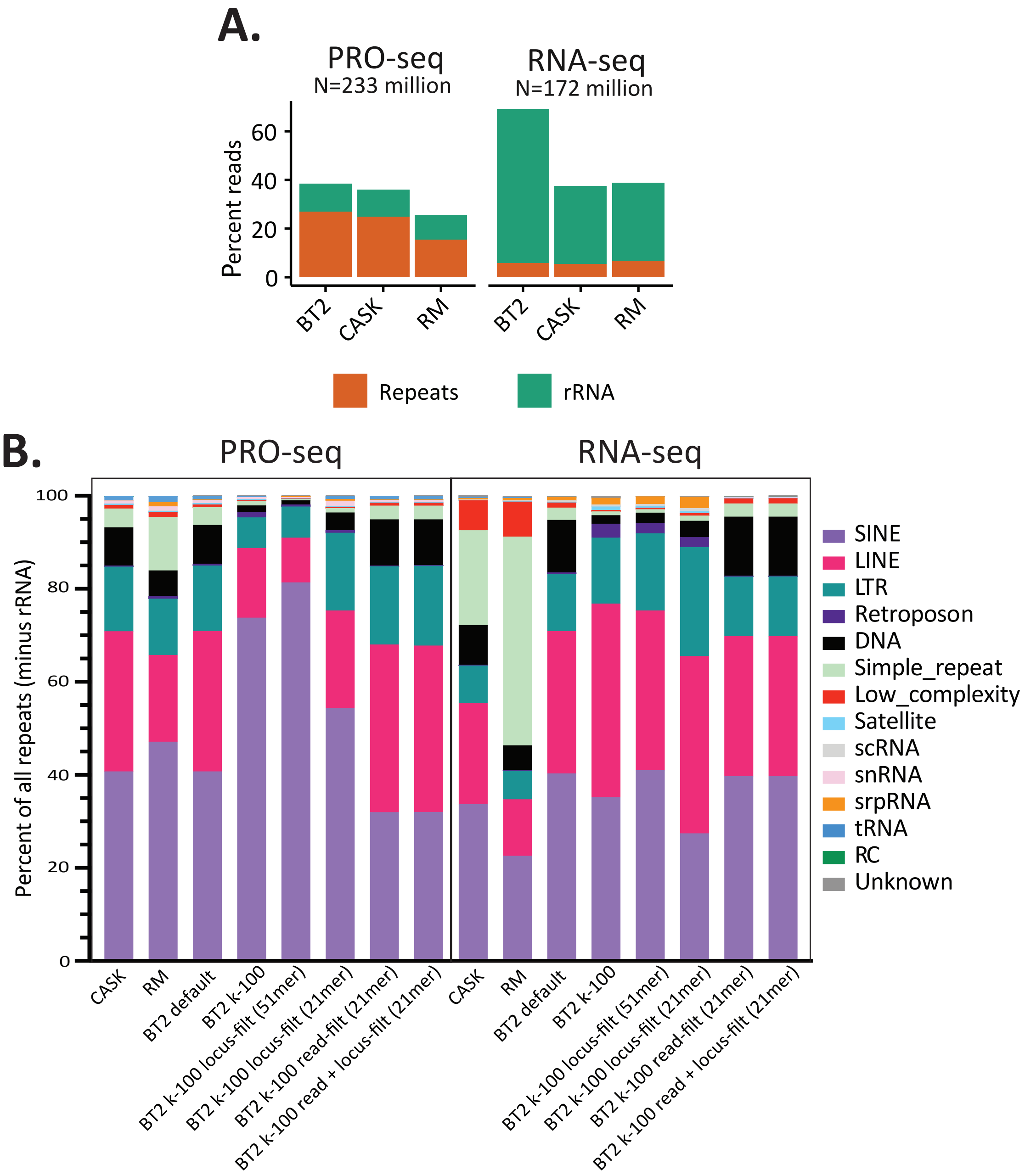


Fig. S30. Transcription of repeats assayed by PRO-seq and RNA-seq. (A) Percentage of reads assigned to repeat elements or rRNA by Bowtie2, CASK, and RepeatMasker (RM) in theT2T-CHM13 PRO-seq and RNA-seq datasets (replicates combined). **(B)** Relative abundance of the 14 repeat classes defined in RMv2 (excluding rRNA), as quantified by Bowtie2, CASK, and RM. Six settings were used for Bowtie2 differing in the handling of multi-mappers: default (default Bowtie2 options, resulting in single alignment randomly chosen for multi-mappers); k-100 (-k-100 option, up to 100 alignments chosen for multi-mappers); k-100 51mer (-k 100 option, but reads were post-filtered to overlap a genome-wide single copy 51-mer); k-100 locus-filt (-k-100 option, but reads were post-filtered to overlap a genome-wide single copy 21-mer); k-100 read-filt (-k-100 option, but reads were post-filtered to contain a genome-wide single copy 21-mer); k-100 read + locus-filt (-k-100 option, but reads were post-filtered to contain and overlap a genome-wide single copy 21-mer).

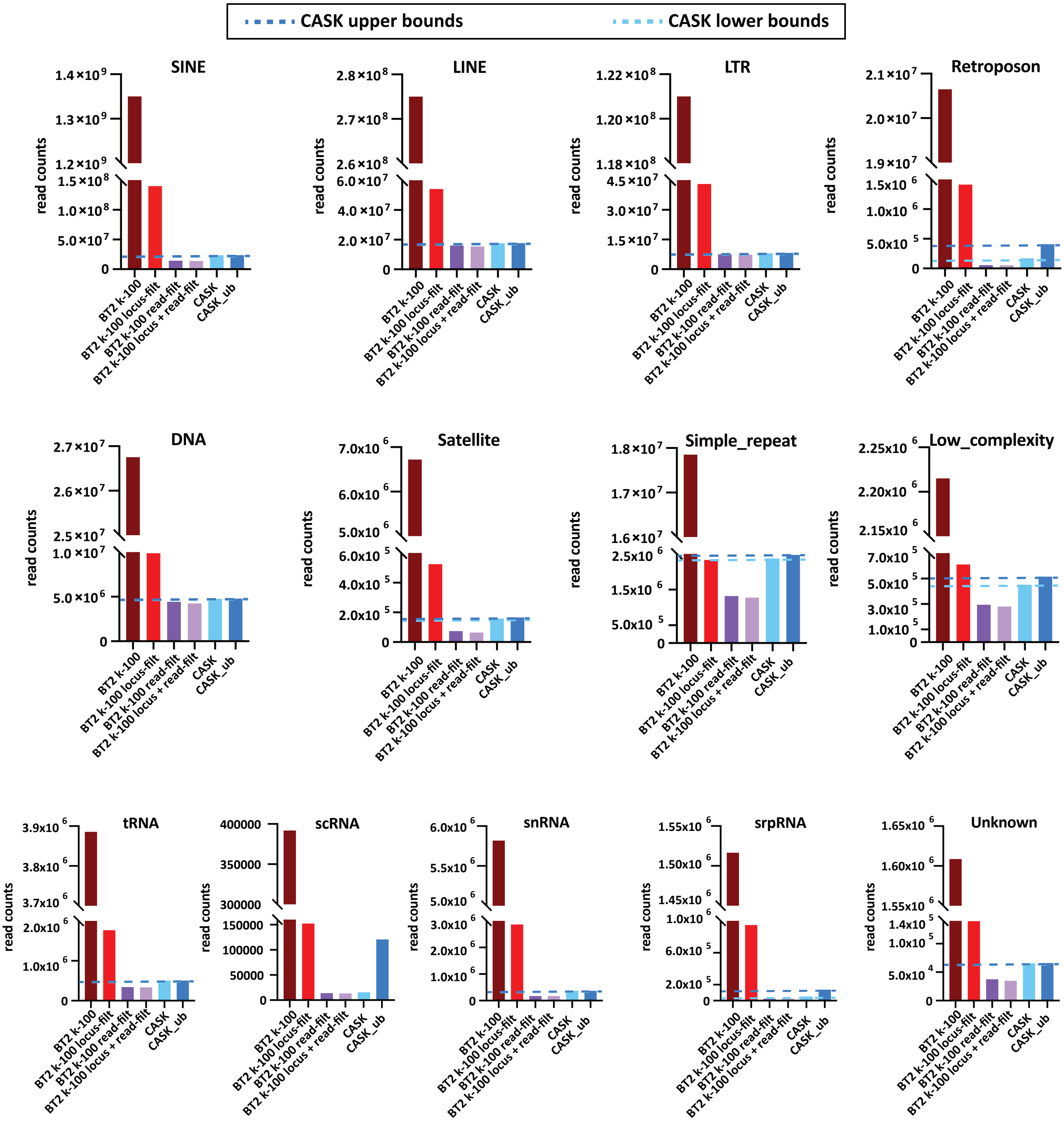


Fig. S31. Repeat transcription pipeline captures a range of upper and lower bounds. Raw read counts per repeat class (excluding rRNA and Rolling Circle (RC)) as quantified by Bowtie2 and CASK in the T2T-CHM13 PRO-seq datasets (replicates combined). Four settings were used for Bowtie2 differing in the handling of multi-mappers and are shown in decreasing stringency from left to right: k-100, k-100 locus-filt (21-mer), k-100 read-filt (21-mer), k-100 read + locus-filt (21-mer). CASK is shown as CASK (lower bounds) and CASK_ub (upper bounds) with dotted lines included to highlight these boundaries across all methods. Since k-100 blows out the axes making it challenging to visualize the other bars, a broken y-axis was used.

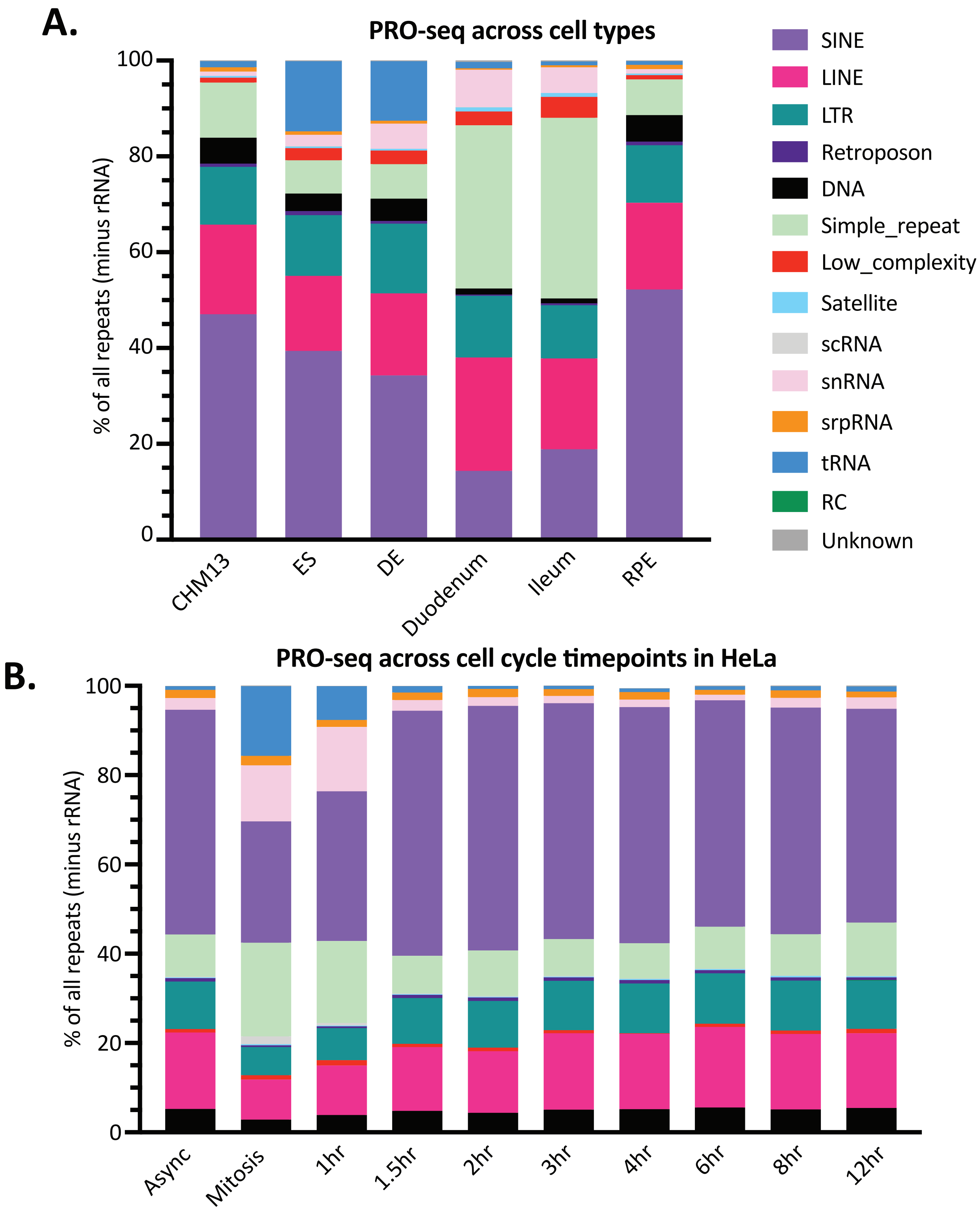


Fig. S32. RepeatMasked PRO-seq reads across (A) cell types and a **(B)** HeLa time-course. Relative abundance of the 14 repeat classes defined in RMv2 (excluding rRNA) as quantified by RepeatMasker. While **(A)** reveals changes in relative repeat abundance between cell types, **(B)** reveals that within an individual cell line (HeLa) relative repeat abundance is highly consistent across time points with slight differences during Mitosis and 1hr post-Mitosis.

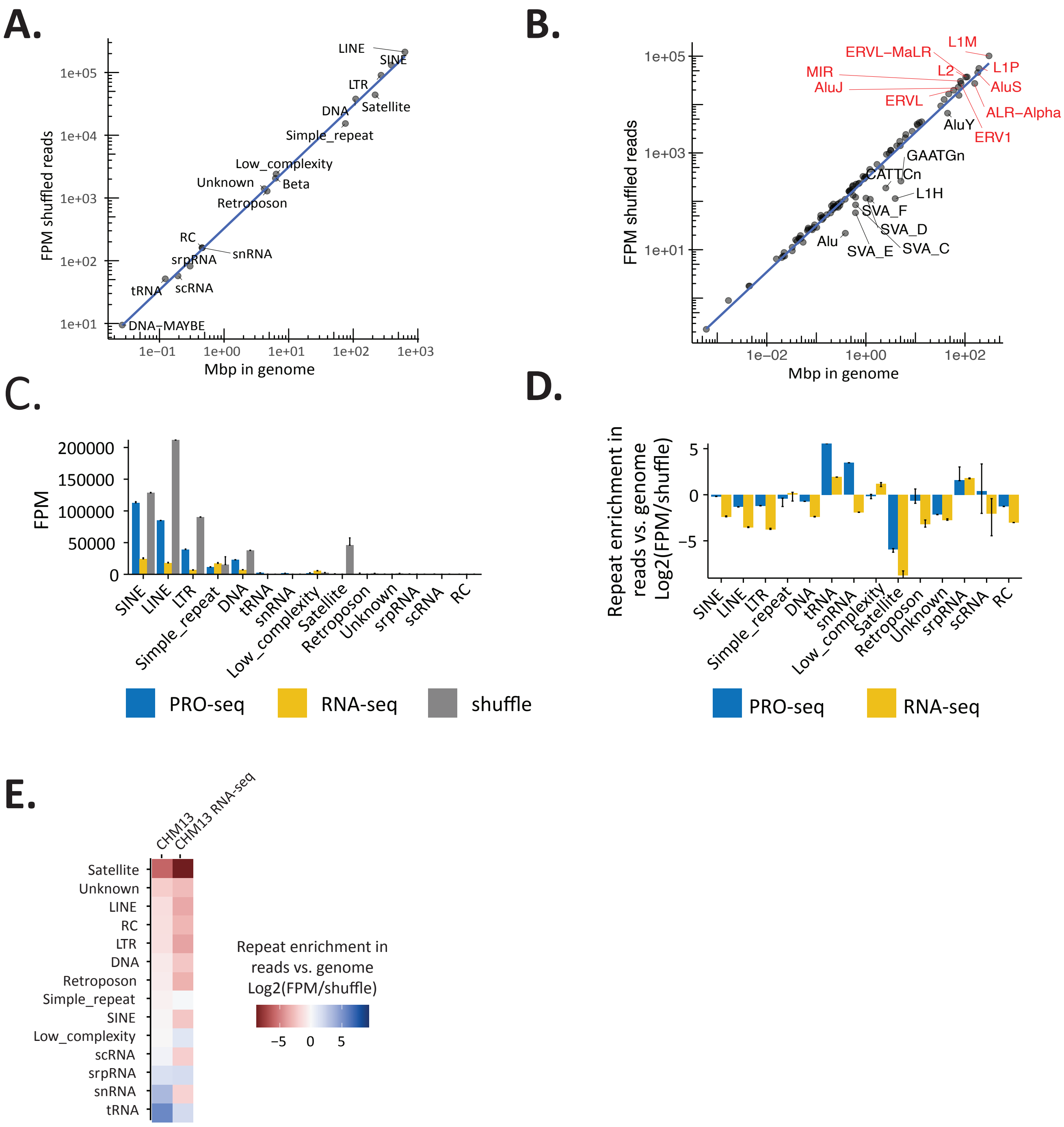


Fig. S33. Repeat expression normalized to genome content. (A) Linear scaling between the genomic abundance of each repeat family **(A)** and class **(B)** (Mbp), and the FPM values obtained by CASK for 100 million 65bp reads sampled from random positions in the genome (shuffled reads). Top 10 classes by Mbp in T2T-CHM13 are indicated in red in **(B)**. **(C)** Comparison between the observed repeat expression in CHM13 and the expression that would be expected from reads originating from random loci (shuffle), both quantified by CASK. Error bars represent upper bounds on the repeat expression, obtained by including reads with ambivalent CASK repeat assignment in the tally. **(D)** Repeat enrichment in the CHM13 transcriptome vs. T2T-CHM13 genome defined as the log2 ratio of observed expression over shuffle. Error bars represent lower and upper bounds on the ratio, obtained by including shuffled or true reads with ambivalent CASK repeat assignment, respectively, in the tally. **(E)** Repeat enrichment across PRO-seq and RNA-seq data ranked from least (red) to most enriched (blue) in CHM13.

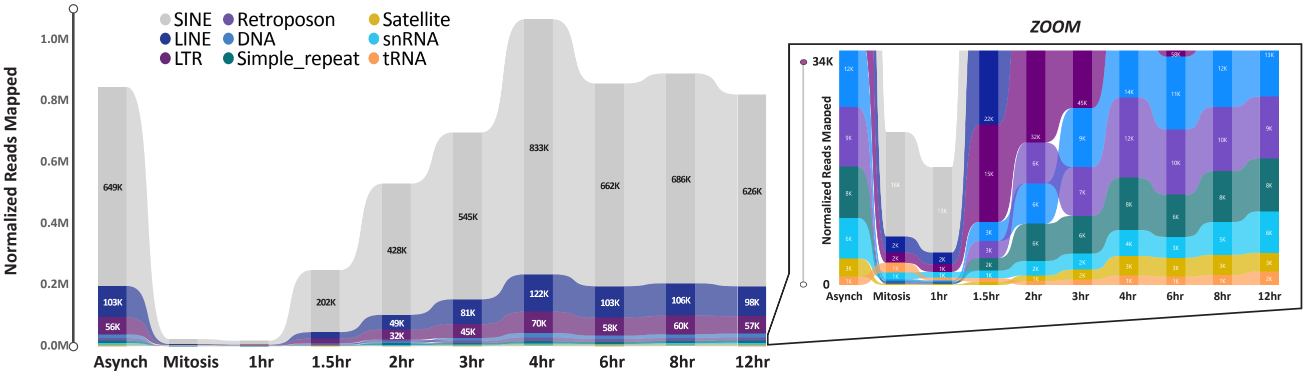


Fig. S34. Repeat abundance in normalized PRO-seq data for HeLa time course. Ribbon plots of repeat abundance in normalized PRO-seq data (shown as Reads per Million RPM) assessed by Bowtie2 (-k-100) in asynchronous and synchronized HeLa cells collected at time points across the cell cycle (key in inset). Zoom shows the reads for the lower range of expressed repeats, including all satellites classified in T2T-CHM13 (tan).

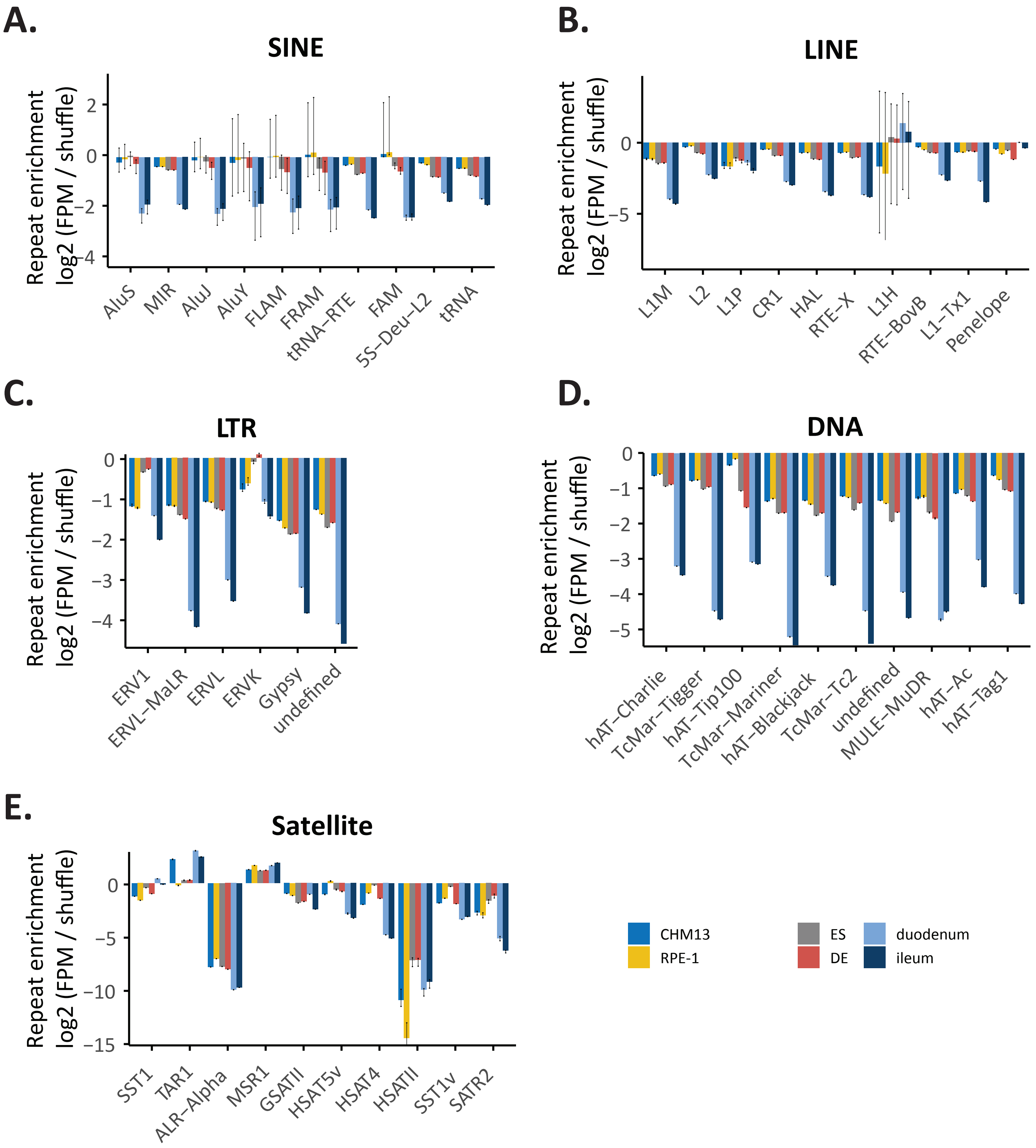


Fig. S35. Repeat enrichment by repeat family in cell specific transcriptomes (as per key, lower right) vs. T2T-CHM13 genome for **(A)** SINE, **(B)** LINE, **(C)** LTR, **(D)** DNA elements, and **(E)** Satellite repeat classes. Within each class, repeats are displayed by decreasing abundance (average expression across cell lines) from left to right, with the top 10 repeat families in each category shown. The repeat enrichment is defined as in fig. S33D.

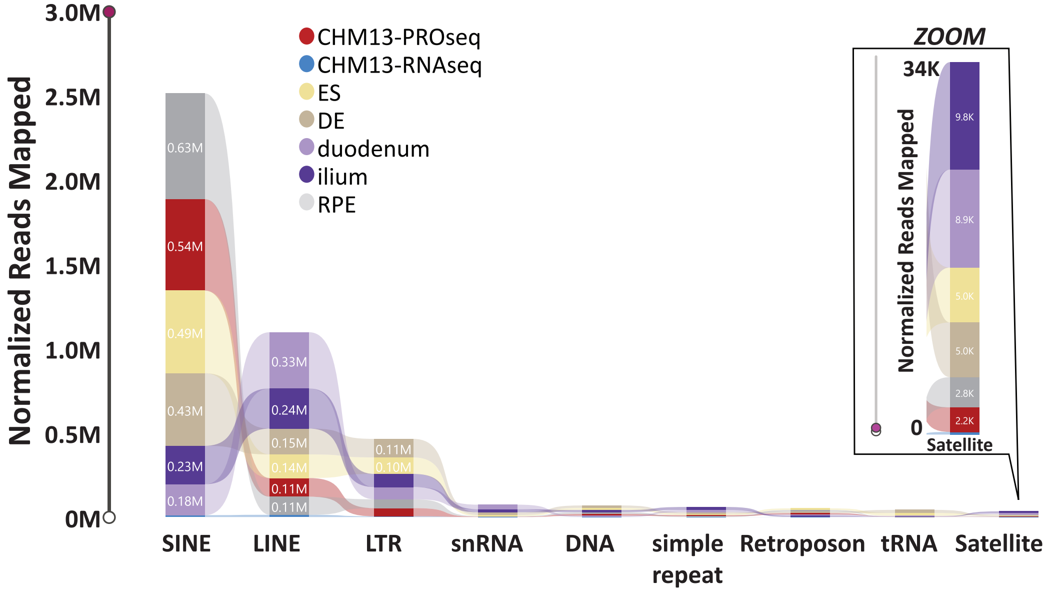


Fig. S36. Repeat abundance in normalized PRO-seq data for H9 developmental time course. Ribbon plots of repeat abundance in normalized PRO-seq data (shown as Reads per Million RPM) assessed by Bowtie2 (-k-100) across cell types (key in inset). Zoom shows the reads for the lower range of expressed repeats, including all satellites classified in T2T-CHM13.

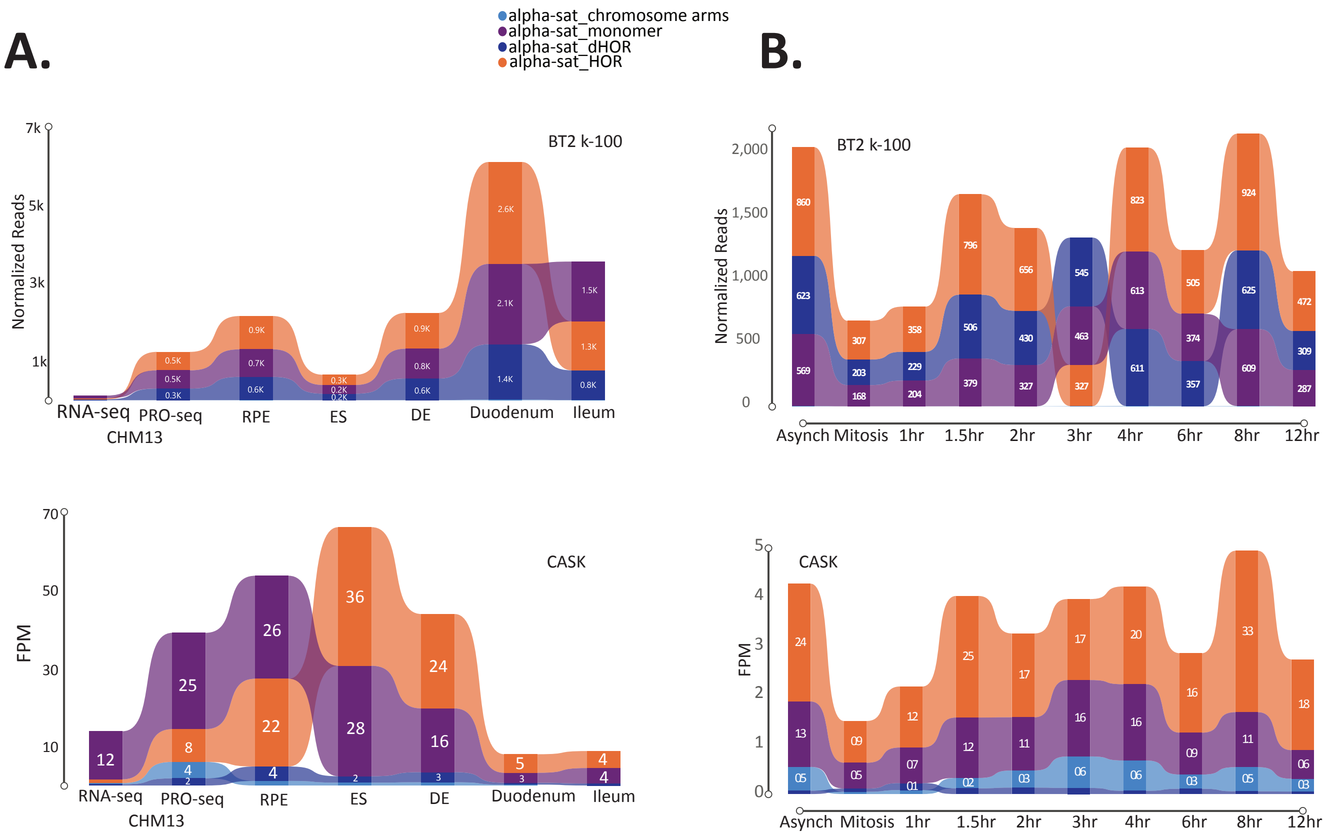
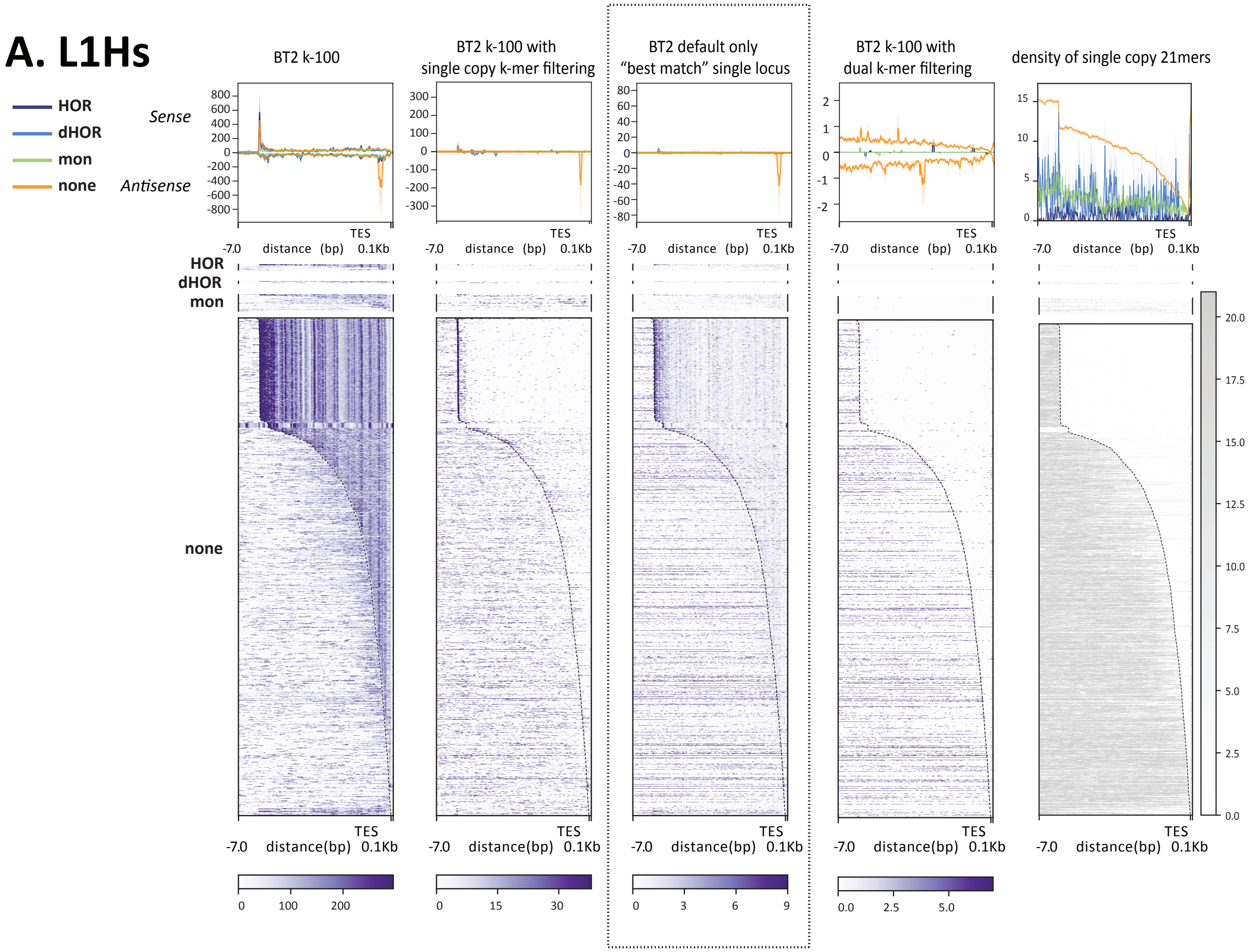
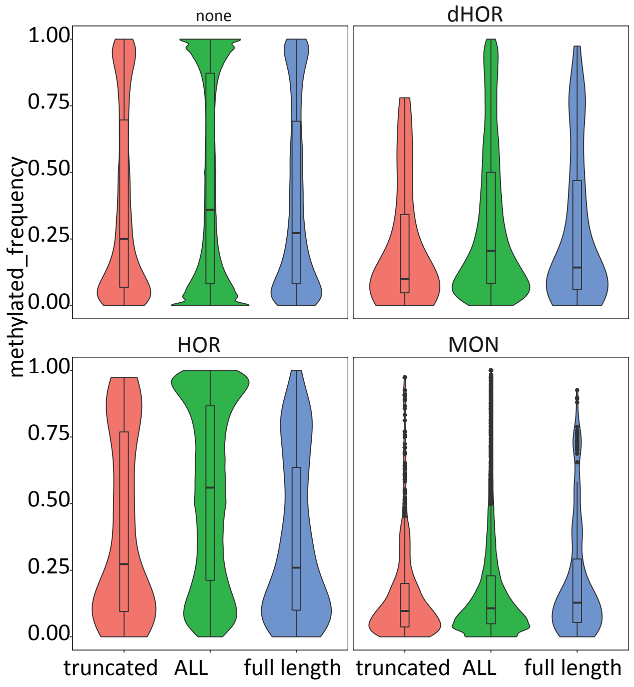
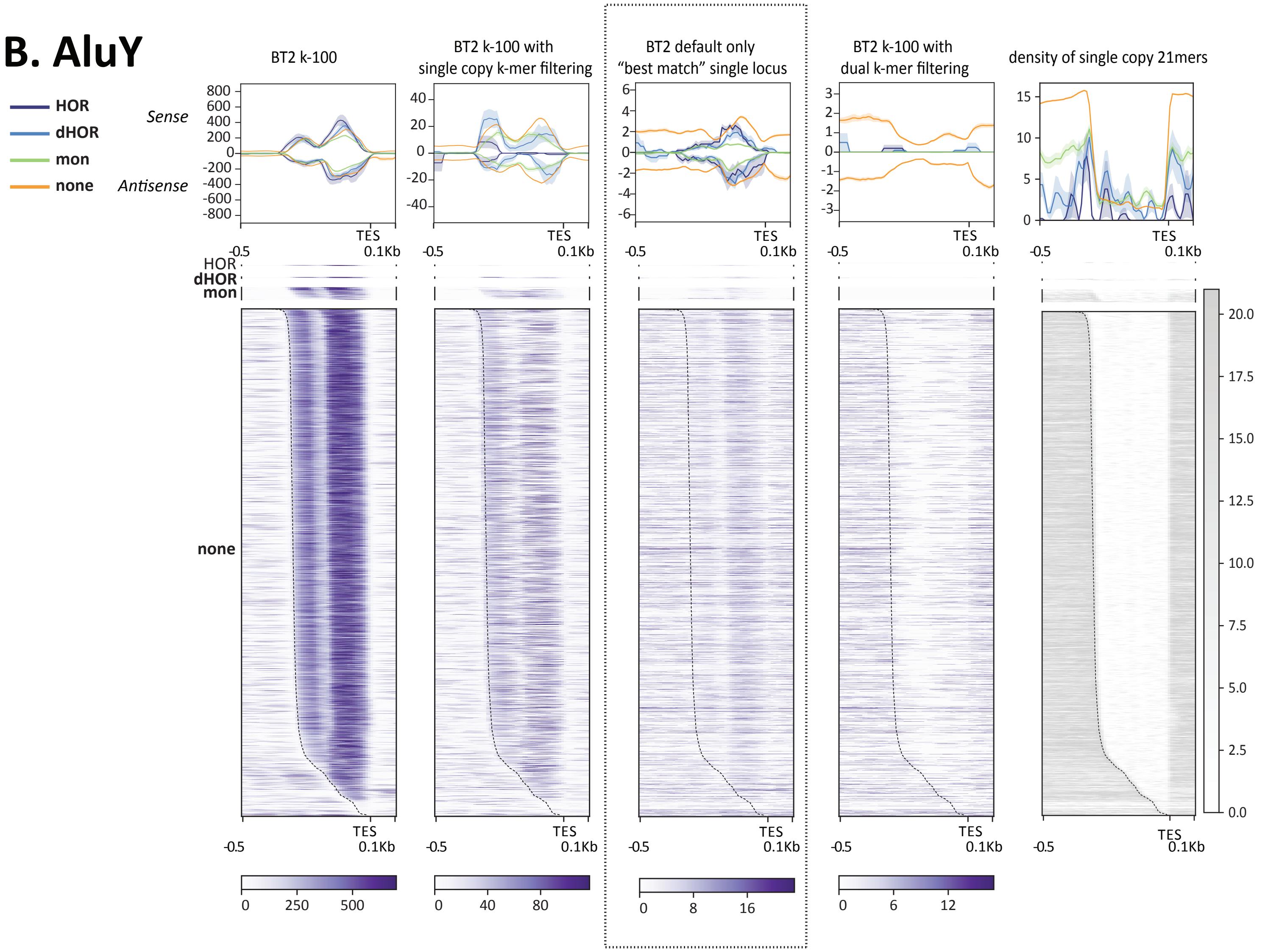


Fig. S37. Alpha satellite abundance in normalized PRO-seq data. Ribbon plots of alpha satellite repeat abundance in PRO-seq data (shown as Reads per Million RPM) assessed by Bowtie2 (-k-100) (top) and CASK (bottom) across cell types and developmental stages **(A)** and time points **(B)** (key at top).


A. L1Hs
HOR
dHOR
mon
none
Sense
Antisense
BT2 k-100
BT2 k-100 with single copy k-mer filtering
BT2 default only "best match" single locus
BT2 k-100 with dual k-mer filtering
density of single copy 21mers
800
600
400
200
0
-200
-400
-600
-800
300
200
100
0
-100
-200
-300
80
60
40
20
0
-20
-40
-60
-80
2
1
0
-1
-2
15
10
5
0
TES
-7.0
distance (bp)
0.1Kb
HOR
dHOR
mon
none
TES
-7.0
distance(bp)
0.1Kb
0
100
200
0
15
30
0
3
6
9
0.0
2.5
5.0
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0


none
dHOR
HOR
MON
methylated_frequency
1.00
0.75
0.50
0.25
0.00
truncated
ALL
full length
truncated
ALL
full length


B. AluY
HOR
dHOR
mon
none
Sense
Antisense
BT2 k-100
BT2 k-100 with single copy k-mer filtering
BT2 default only "best match" single locus
BT2 k-100 with dual k-mer filtering
density of single copy 21mers
800
600
400
200
0
-200
-400
-600
-800
40
20
0
-20
-40
6
4
2
0
-2
-4
-6
3
2
1
0
-1
-2
-3
15
10
5
0
TES
-0.5
0.1Kb
HOR
dHOR
mon
none
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
250
500
0
40
80
0
8
16
0
6
12

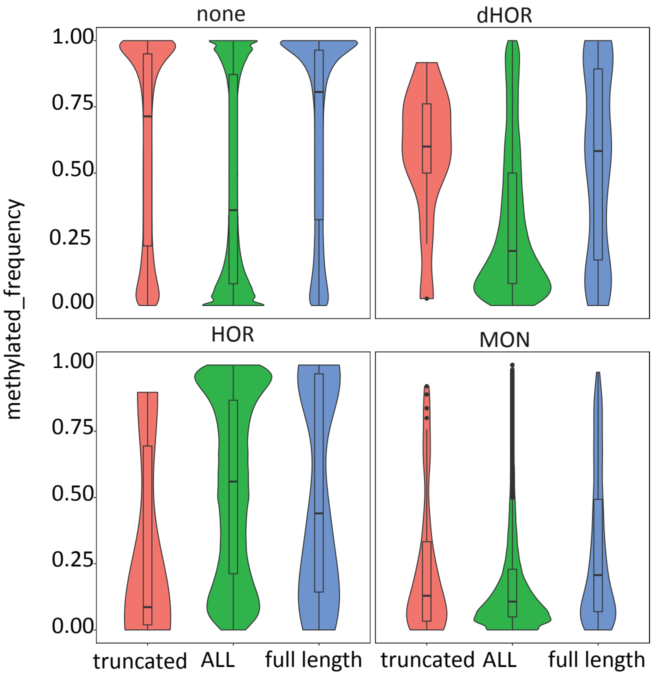

none
dHOR
HOR
MON
methylated_frequency
1.00
0.75
0.50
0.25
0.00
truncated
ALL
full length
truncated
ALL
full length

Fig S38. Transcription and methylation profiles of embedded elements. (A) L1Hs and **(B)** *Alu*Y. Top Panel: PRO-seq density (purple scale, reads per million both sense and antisense aggregated) and average profiles (top line graphs, separated into sense and antisense read density) for TE subfamilies. All elements are anchored to the 3' end, with a specified distance from the anchor (bottom left) into the element; standard error shading (70% opacity of respective line color), TSS (transcription start site), TES (transcription end site), and ±Kbp are shown. A dotted line is included on the heatmap denoting the starting nt of each annotated element. HOR, dHOR and monomeric embeds are separated from all non-embedded elements (subsampled for *Alu*Y). Mapping methods (from left) are Bowtie2 (BT2) k-100; BT2 k-100 21nt k-mer filtered (locus level); BT2 default only ("best match" single locus); BT2 k-100 dual 21nt k-mer filtered (read and locus filtered). BT2 default only are within a dotted box. Far right is the density of single copy 21nt k-mers across each element (grey scale, number of single copy k-mers in sense and antisense aggregated) and average profiles (top line graphs) for each TE subfamily.

Bottom panel: Violin plots of methylated CpGs for TEs grouped by their potential for mobility and compared across embedded categories (dHOR, HOR, monomeric) vs not embedded.

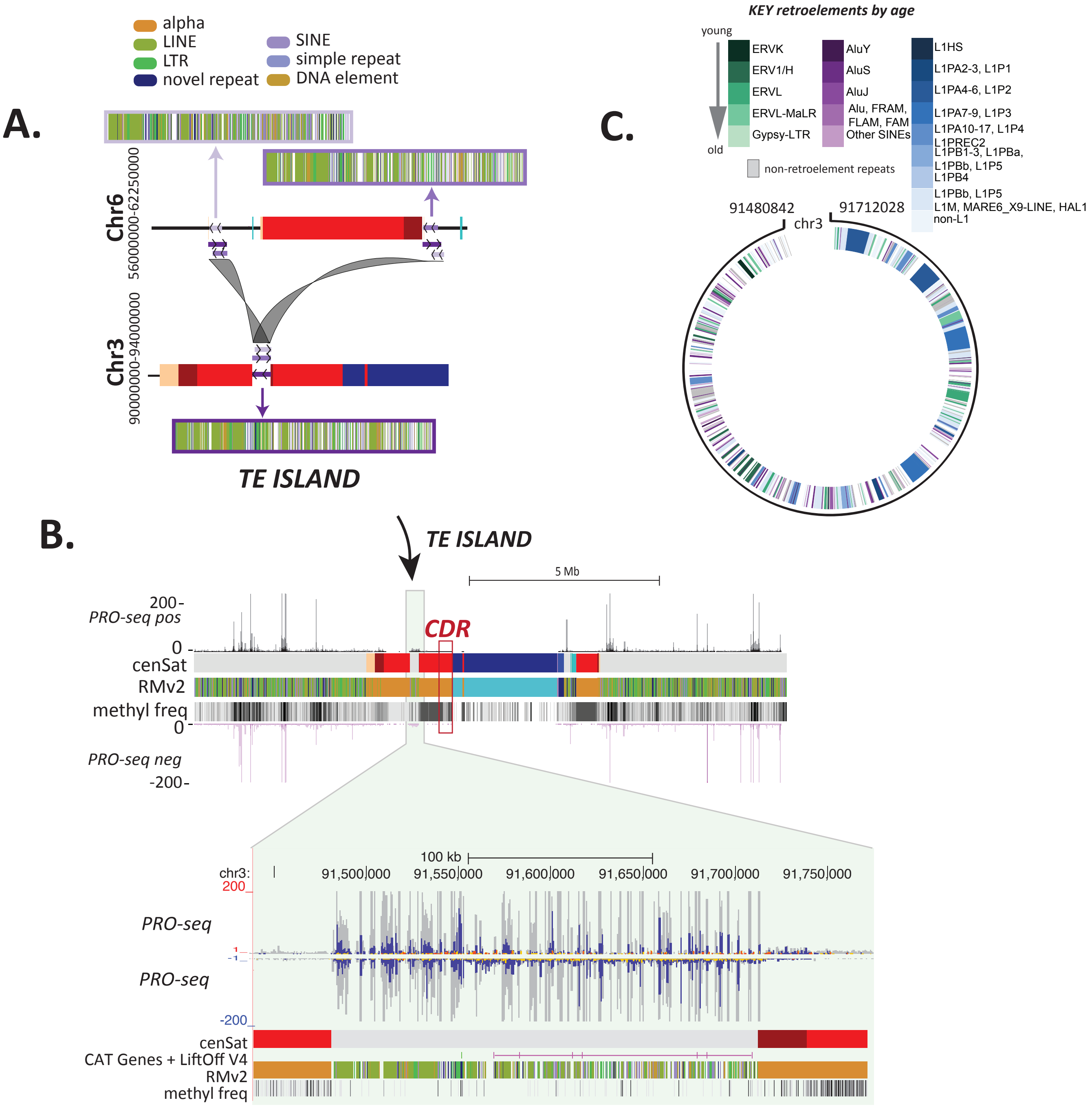


Fig. S39. TE island found in HOR of Chromosome 3. (A) Island of TE elements found within the HOR of T2T-CHM13 Chromosome 3 has undergone segmental duplication to pericentomeres of Chromosome 6. Colored blocks are RM2 tracks for the duplications as indicated by purple (light- Chromosome 6, dark – Chromosome 3). **(B)** Browser track of T2T-CHM13 Chromosome 3 centromere including cen transition (gray) and satellite arrays (colored as per key to left). The TE island (black box) resides within the alpha satellite HOR yet does not overlap with the CDR (red box) (21). Top: censat, RM2, methylation frequency and PRO-seq (Bowtie2 default, "best match"), are shown. Zoom inset (right) shows PRO-seq signal across TEs (denoted as per RM2 track) associated with low methylation levels. PRO-seq tracks show Bowtie2 "best match" mapping (yellow), k-100 over-fit mapping (grey), and single (blue) and dual filtered (red) k-100 as per Fig. 4D-E. **(C)** The relative age of retroelements (left) shows the island contains no elements with recent activity, but rather has elements that were active during the divergence of the hominoid lineages.

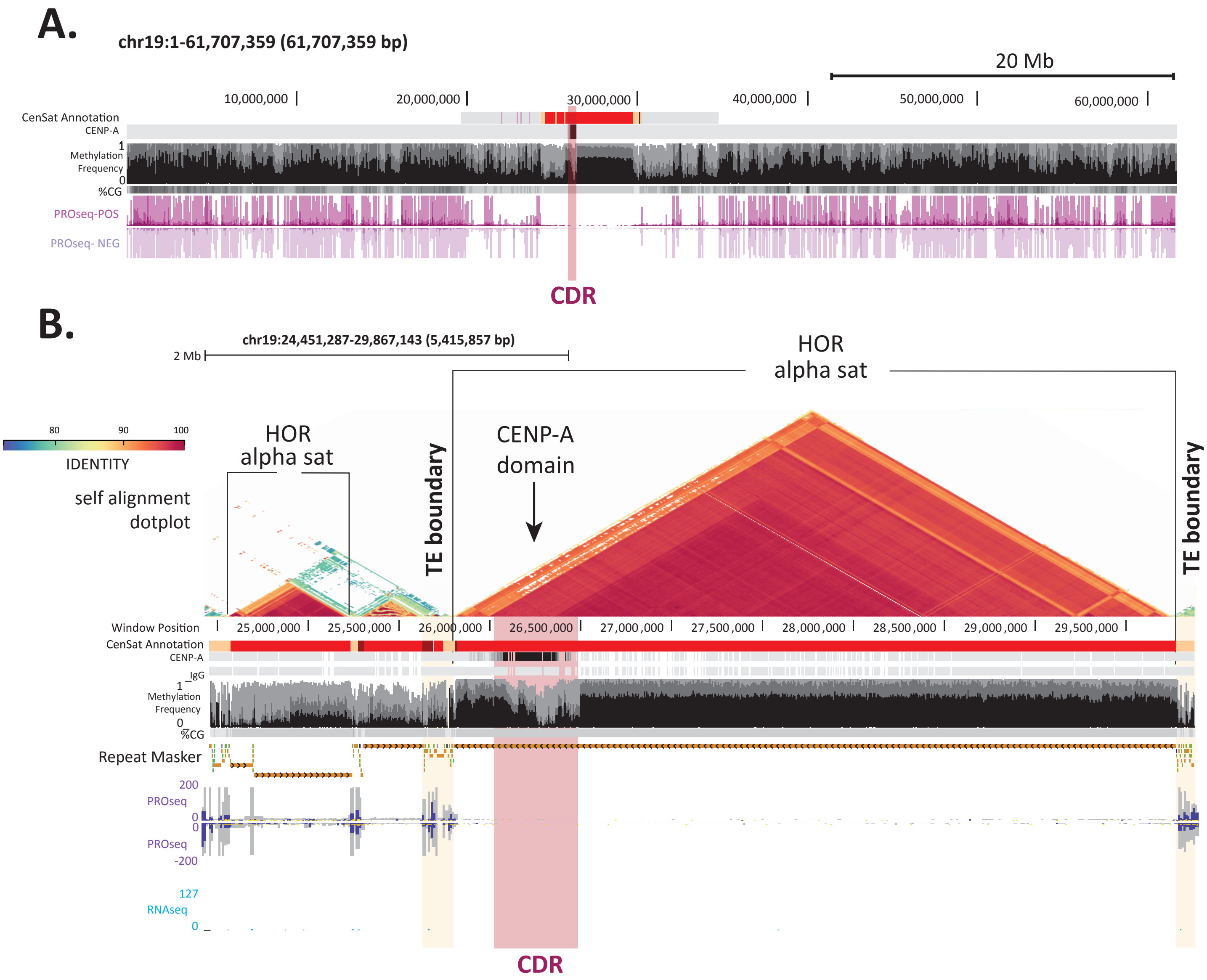


Fig. S40. Sites of engaged RNA polymerase across Chromosome 19. (A) Chromosome 19 from telomere to telomere. From top to bottom, cenSAT track (*12*), CENP-A CUT&RUN (*12*), methylation frequency (24), %CG, positive (purple) and negative PRO-seq signal (pink) from Bowtie2 default "best match". The CDR (*21*)centromere dip region) that coincides with CENP-A (*12*) is indicated in pale red. **(B)** Zoom of the centromere region of Chromosome 19. Tracks from top to bottom: self-alignment dot plot (scale of % identity is shown), cenSAT track, CENP-A CUT&RUN, IgG, methylation frequency, repeat masker v2 track, PRO-seq tracks show Bowtie2 "best match" mapping (yellow), k-100 over-fit mapping (grey), and single (blue) and dual filtered (red) k-100 as per Fig. 4D-E, and RNA-seq mapped reads (Bowtie2 default, "single match") (blue). The CDR is indicated in pink, the TE boundaries are in tan.

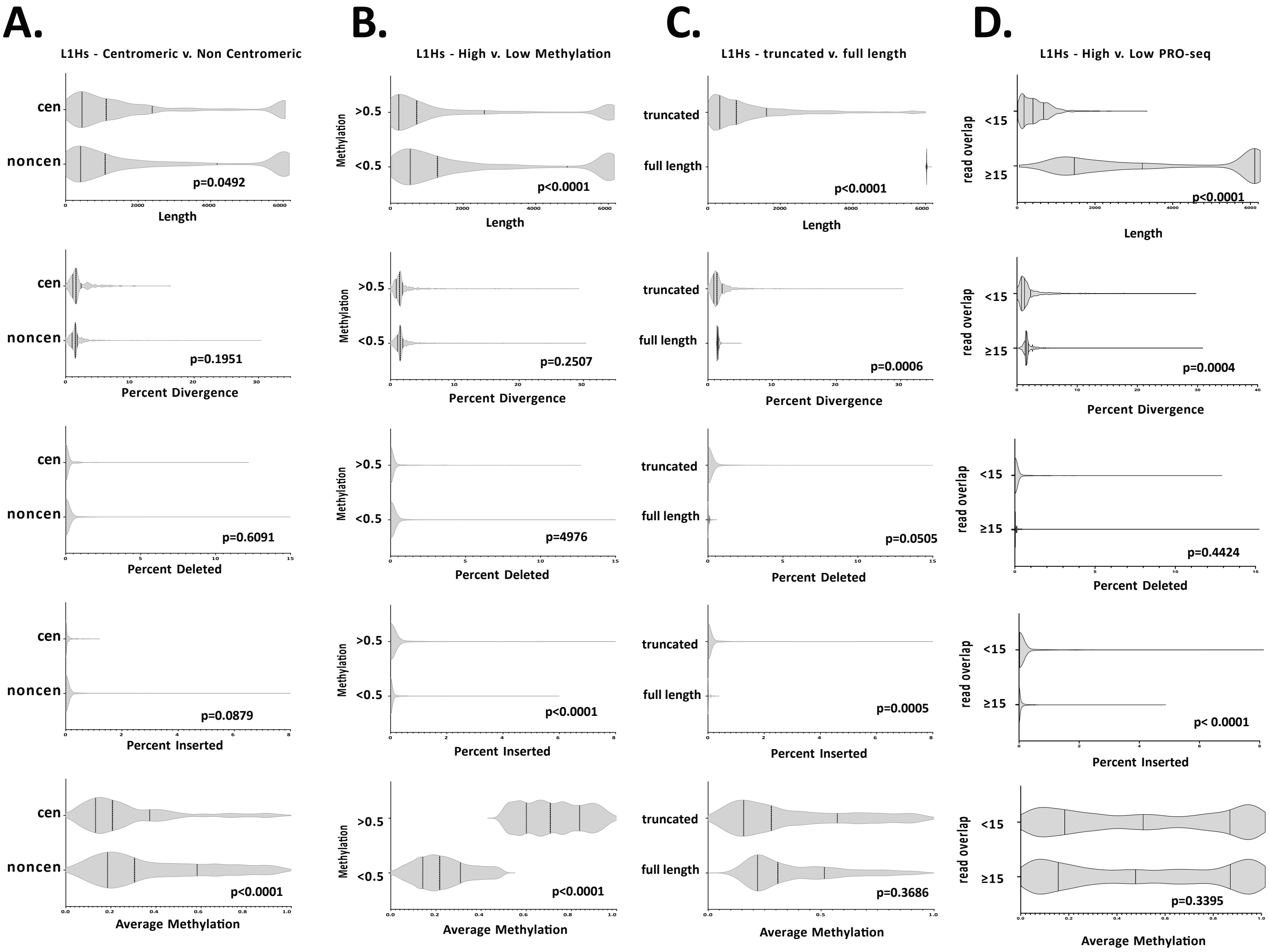
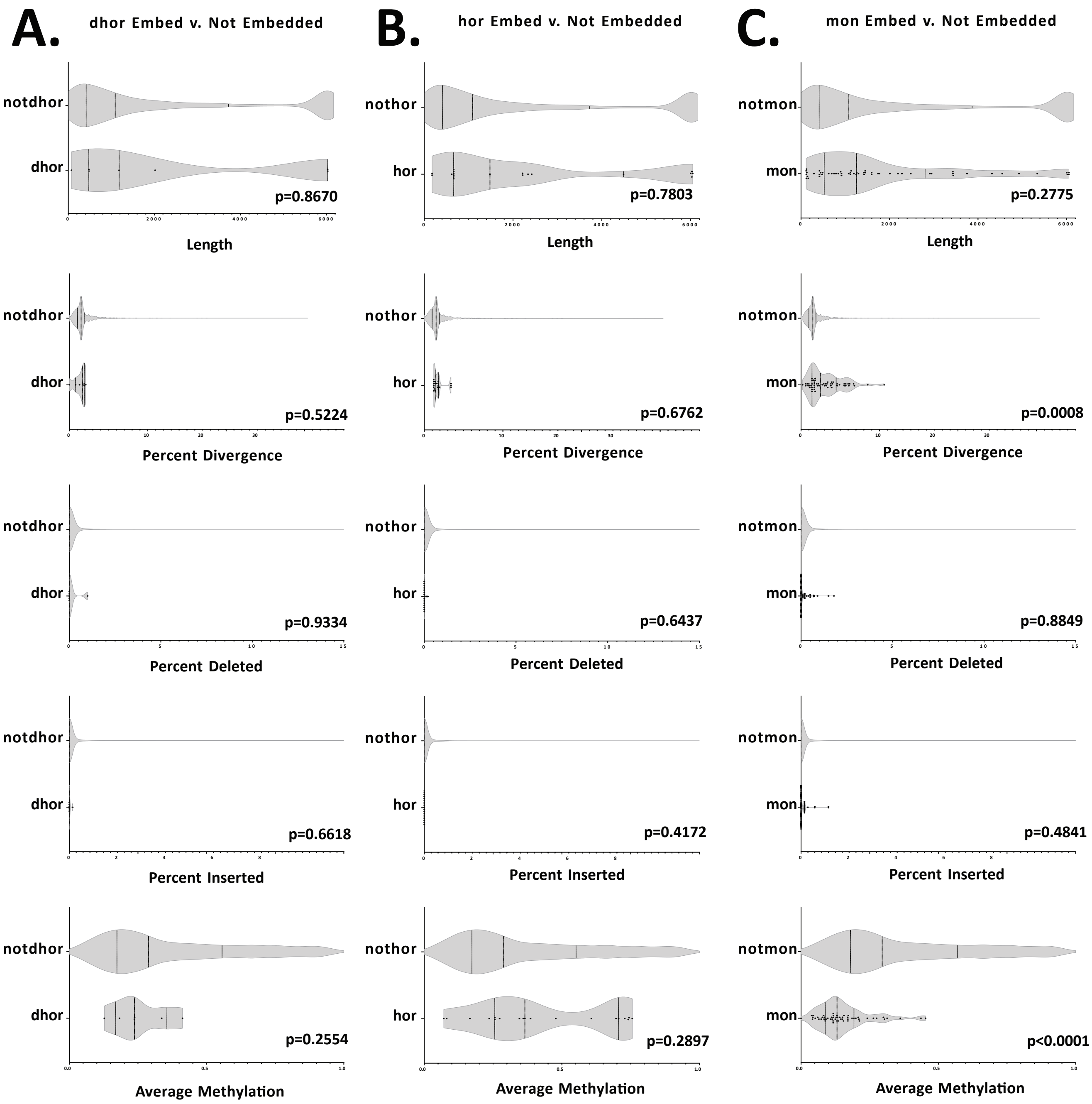


Fig. S41. Violin plots of L1Hs elements show differences in length, divergence, deletions, insertions, and methylation of repeats found in the centromere (considered as centromere and centromere transitions as defined in (*12*)). **(A)**, with varying methylation patterns **(B)**, length **(C)** and with varying PRO-seq expression levels **(D)**. Centromeric L1Hs are significantly shorter (p=0.0492) and have a lower average methylation (p<0.0001) than non-centromeric L1Hs. L1Hs with an average methylation < 0.5 are longer and have less insertions than L1Hs with an average methylation > 0.5 (p<0.0001). Transcribed L1Hs elements are longer, less diverged, and have less insertions than those predicted to be incapable of transposition (p=0.0005-<0.0001) based on truncation. Similarly, L1Hs with ≥15 overlapping PRO-seq reads are longer (p<0.0001) and less diverged (p=0.0004) from the consensus than L1Hs with <15 overlapping PRO-seq reads.


A.
dhor Embed v. Not Embedded
notdhor
dhor
p=0.8670
0
2000
4000
6000
Length
notdhor
dhor
p=0.5224
0
10
20
30
Percent Divergence
notdhor
dhor
p=0.9334
0
5
10
15
Percent Deleted
notdhor
dhor
p=0.6618
0
2
4
6
8
Percent Inserted
notdhor
dhor
p=0.2554
0.0
0.5
1.0
Average Methylation
B.
hor Embed v. Not Embedded
nothor
hor
p=0.7803
0
2000
4000
6000
Length
nothor
hor
p=0.6762
0
10
20
30
Percent Divergence
nothor
hor
p=0.6437
0
5
10
15
Percent Deleted
nothor
hor
p=0.4172
0
2
4
6
8
Percent Inserted
nothor
hor
p=0.2897
0.0
0.5
1.0
Average Methylation
C.
mon Embed v. Not Embedded
notmon
mon
p=0.2775
0
2000
4000
6000
Length
notmon
mon
p=0.0008
0
10
20
30
Percent Divergence
notmon
mon
p=0.8849
0
5
10
15
Percent Deleted
notmon
mon
p=0.4841
0
2
4
6
8
Percent Inserted
notmon
mon
p<0.0001
0.0
0.5
1.0
Average Methylation

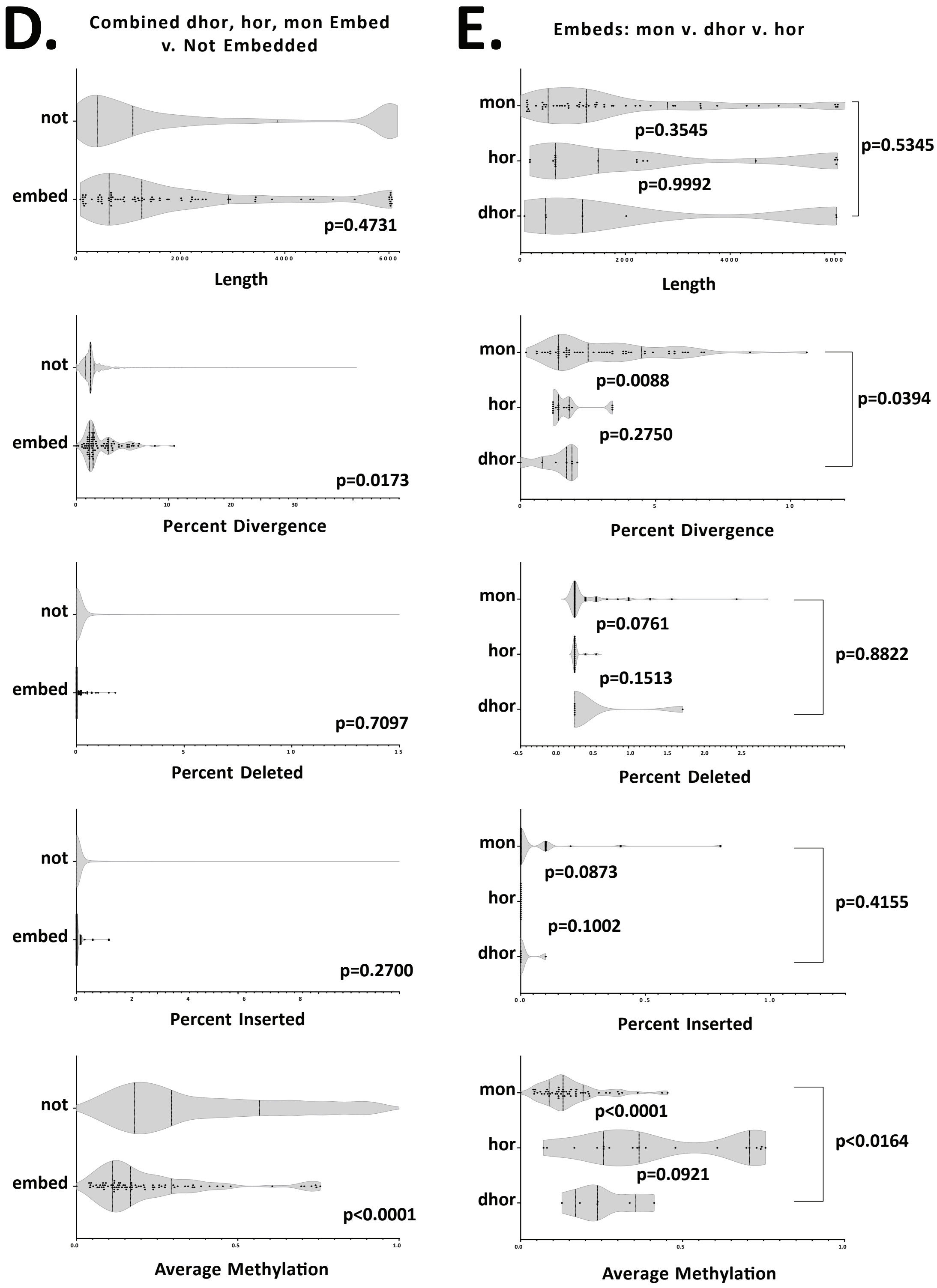
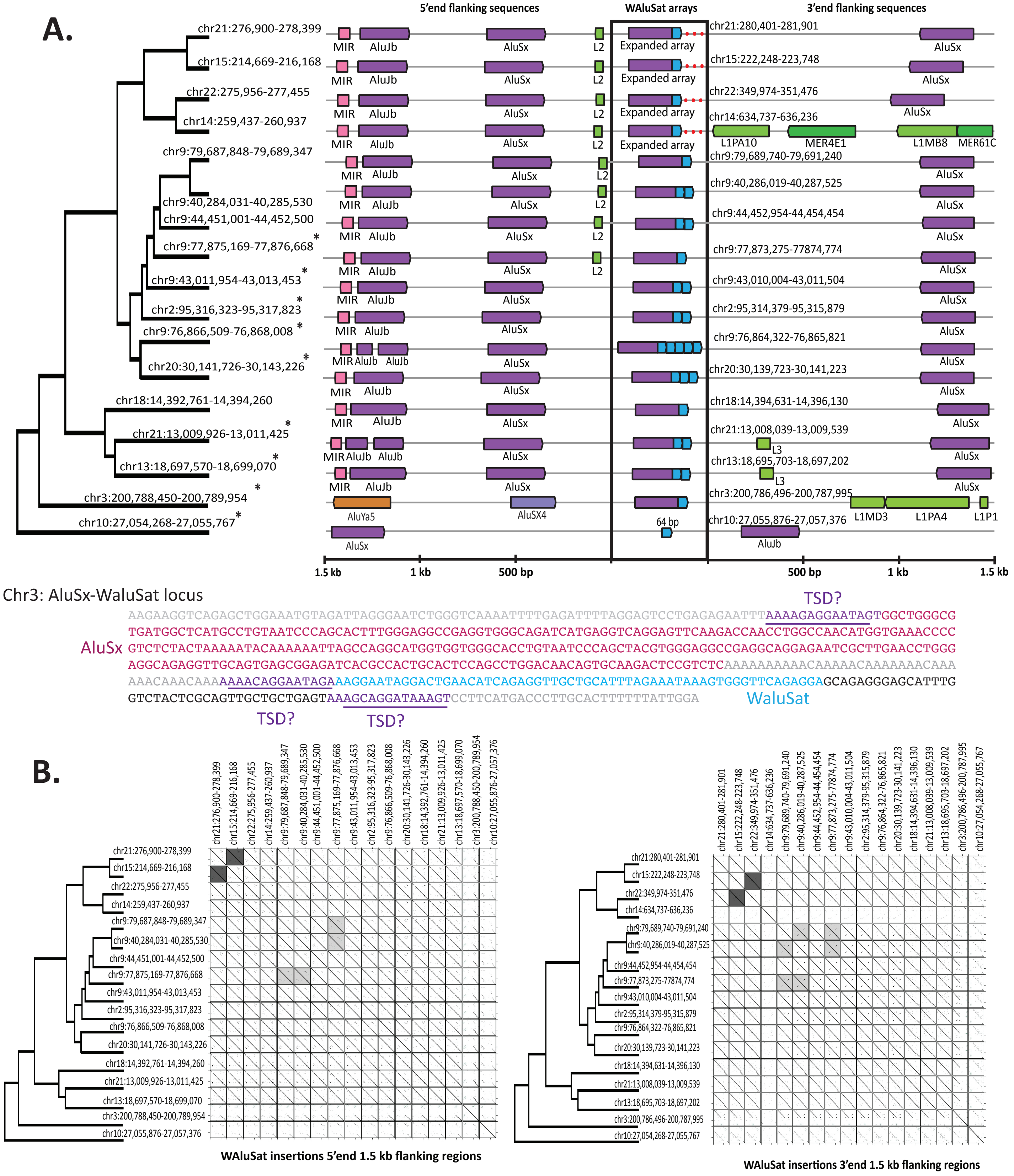


Fig. S42. Violin plots of L1Hs elements embedded in dHORs (A), HORs (B), and alpha satellite monomers (C) reveal few significant variations in length, divergence, deletions, insertions, and methylation of repeats. L1Hs embedded within alpha satellite monomers are both more diverged and less methylated than those not embedded in monomers (p=0.0008-<0.0001), accounting for the similar pattern in variation identified when comparing L1Hs embedded in these combined regions to unembedded repeats **(D)**. Comparing embeds in each classification **(E)** reveals that L1Hs embedded in monomers are more diverged and less methylated than those embedded in HORs (p=0.0088; p<0.0001, respectively) and dHORs (p=0.0394; p=0.0164, respectively), while those embedded in HORs and dHORs do not differ significantly.


A.
5'end flanking sequences
WAluSat arrays
3'end flanking sequences
chr21:276,900-278,399
chr15:214,669-216,168
chr22:275,956-277,455
chr14:259,437-260,937
chr9:79,687,848-79,689,347
chr9:40,284,031-40,285,530
chr9:44,451,001-44,452,500
chr9:77,875,169-77,876,668 *
chr9:43,011,954-43,013,453 *
chr2:95,316,323-95,317,823 *
chr9:76,866,509-76,868,008 *
chr20:30,141,726-30,143,226 *
chr18:14,392,761-14,394,260
chr21:13,009,926-13,011,425 *
chr13:18,697,570-18,699,070 *
chr3:200,788,450-200,789,954 *
chr10:27,054,268-27,055,767 *
MIR
AluJb
AluSx
L2
Expanded array
AluYa5
AluSX4
64 bp
chr21:280,401-281,901
chr15:222,248-223,748
chr22:349,974-351,476
chr14:634,737-636,236
chr9:79,689,740-79,691,240
chr9:40,286,019-40,287,525
chr9:44,452,954-44,454,454
chr9:77,873,275-77874,774
chr9:43,010,004-43,011,504
chr2:95,314,379-95,315,879
chr9:76,864,322-76,865,821
chr20:30,139,723-30,141,223
chr18:14,394,631-14,396,130
chr21:13,008,039-13,009,539
chr13:18,695,703-18,697,202
chr3:200,786,496-200,787,995
chr10:27,055,876-27,057,376
L1PA10
MER4E1
L1MB8
MER61C
L3
L1MD3
L1PA4
L1P1
1.5 kb
1 kb
500 bp
500 bp
1 kb
1.5 kb
Chr3: AluSx-WaluSat locus
TSD?
AluSx
WaluSat
TSD?
TSD?
B.
WAluSat insertions 5'end 1.5 kb flanking regions
WAluSat insertions 3'end 1.5 kb flanking regions

```
AAGAAGGTCAGAGCTGGAAATGTAGATTAGGGAATCTGGGTCAAAATTTTGAGATTTTAGGAGTCCTGAGAGAATTTAAAAGAGGAATAGTGGCTGGGCG
TGATGGCTCATGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGTGGGCAGATCATGAGGTCAGGAGTTCAAGACCAACCTGGCCAACATGGTGAAACCCC
GTCTCTACTAAAAATACAAAAAATTAGCCAGGCATGGTGGTGGGCACCTGTAATCCCAGCTACGTGGGAGGCCGAGGCAGGAGAATCGCTTGAACCTGGG
AGGCAGAGGTTGCAGTGAGCGGAGATCACGCCACTGCACTCCAGCCTGGACAACAGTGCAAGACTCCGTCTCAAAAAAAAAACAAAAACAAAAAAAACAAA
AAACAAACAAAAAAACAGGAATAGAAAGGAATAGGACTGAACATCAGAGGTTGCTGCATTTAGAAATAAAGTGGGTTCAGAGGAGCAGAGGGAGCATTTG
GTCTACTCGCAGTTGCTGCTGAGTAAAGCAGGATAAAGTCCTTCATGACCCTTGCACTTTTTTATTGGA
```

Fig. S43. WaluSat arrays flanking regions in the T2T-CHM13 genome. (A) Schematic representation of WaluSat arrays and TE insertions and 1.5 kb flanking regions annotated in T2T-CHM13. All flanking regions are organized in the same orientation for sake of analysis and schematic representations are scaled while WaluSat array representations are not to scale. Asterisks indicate sequences in inverted orientation for sake of analysis. Red dots indicate the presence of variable expanded WaluSat arrays. Sequence of Chromosome 3 AluSx-WaluSat is shown with putative TSDs. **(B)** Dotplot comparisons of 1.5 kb of sequence from both and 3' flanking regions show an absence of sequence identity in the most ancient WaluSat insertions on Chromosome 10 and Chromosome 3 positions. Darker shading indicates higher similarity shared along subsequences. Phylogenetic reconstruction used is based on Figure 5A.

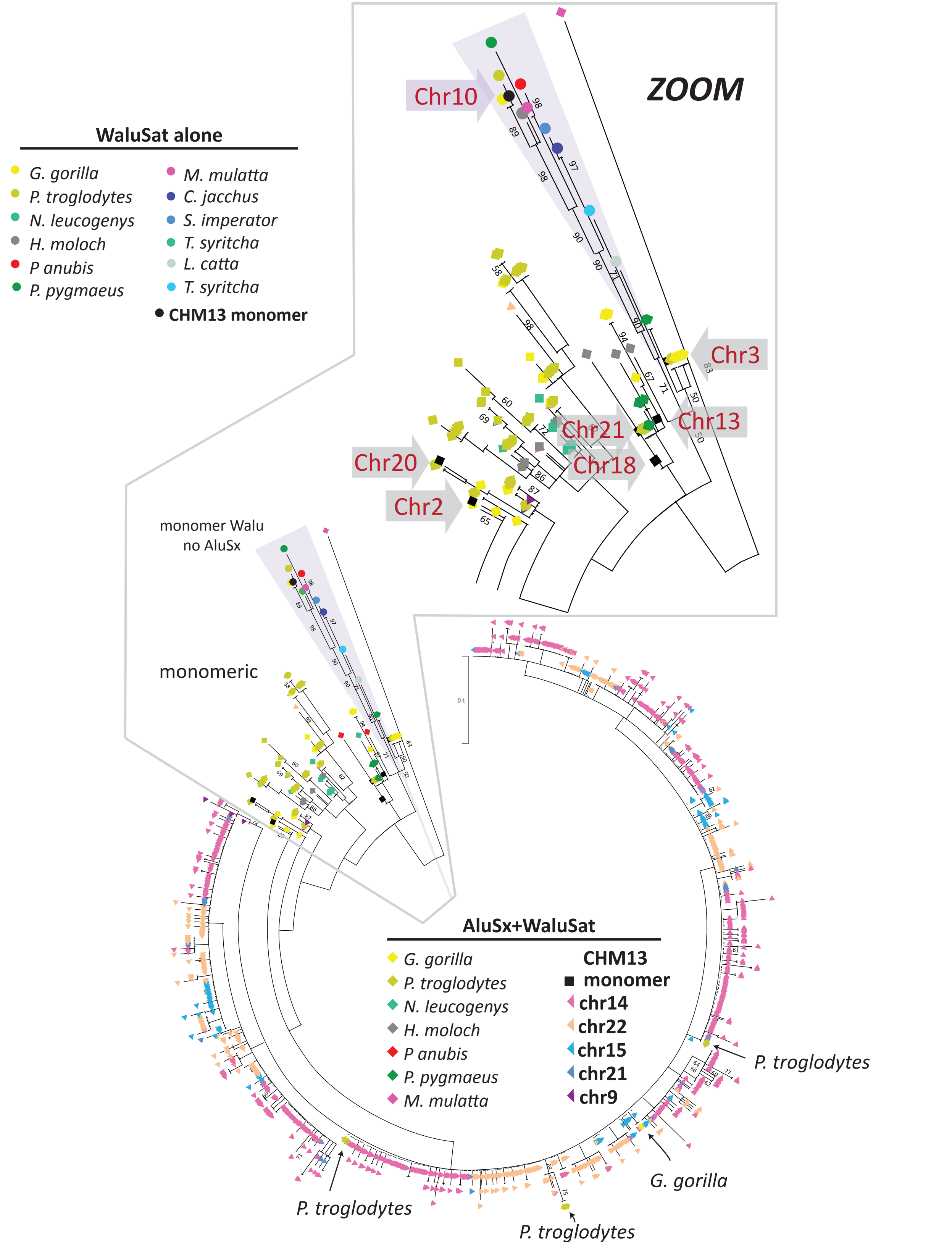


Fig S44. Phylogenetic analysis of the WaluSat monomer across primate lineages. ML analyses show the WaluSat sequences transduced by AluSx cluster together (main circle), as do the WaluSat monomers (Chromosomes 2, 3, 10, 13, 18, 21) monomers (boxed), which are found in Catarrhini and Hominoidea primates. Only the WaluSat monomer on T2T-CHM13 Chromosome 10 clusters with Hominoidea, Catarrhini, Platyrrhini, and Prosimians (zoom), indicating this locus is the progenitor of the satellite repeat.

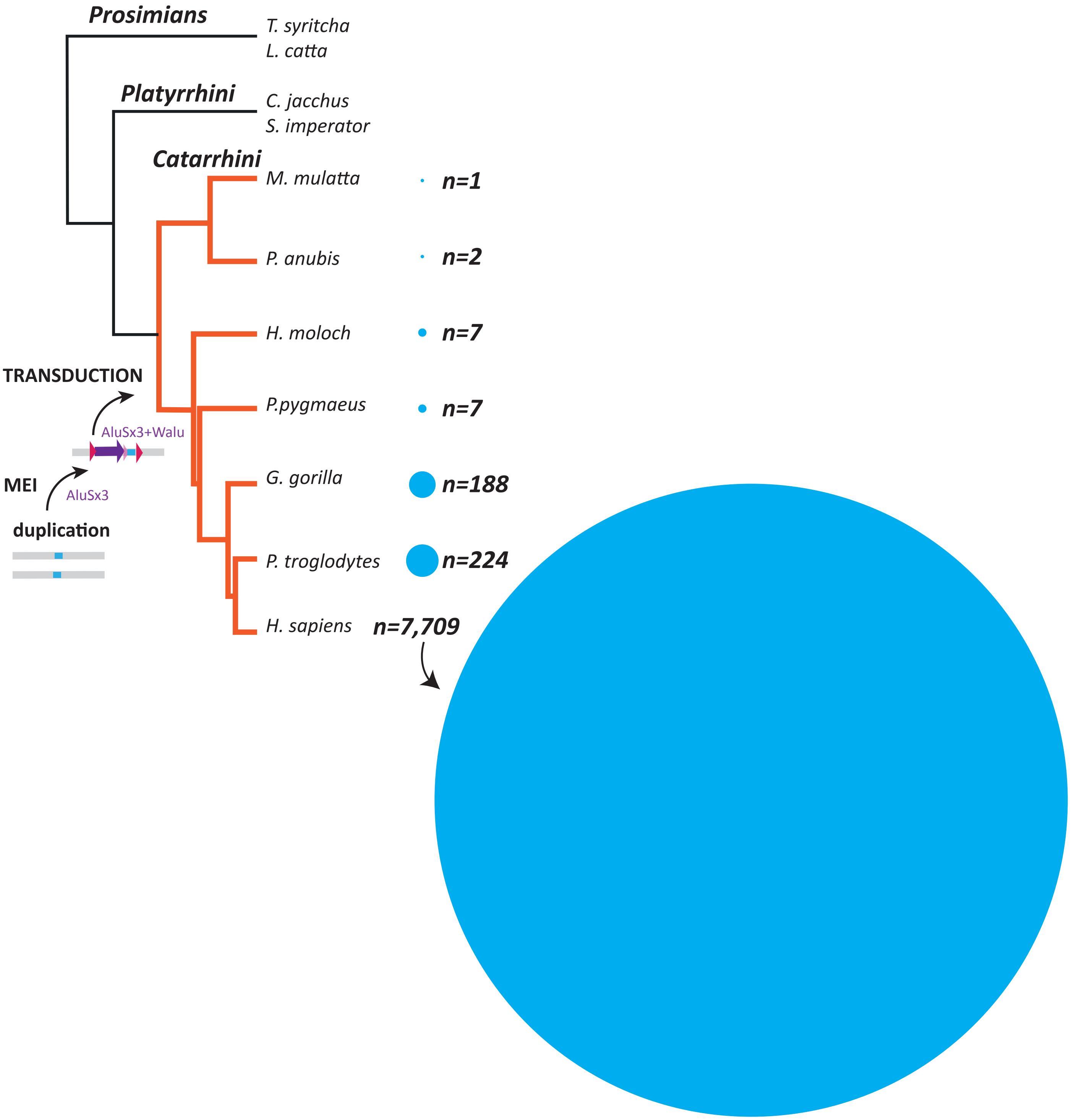
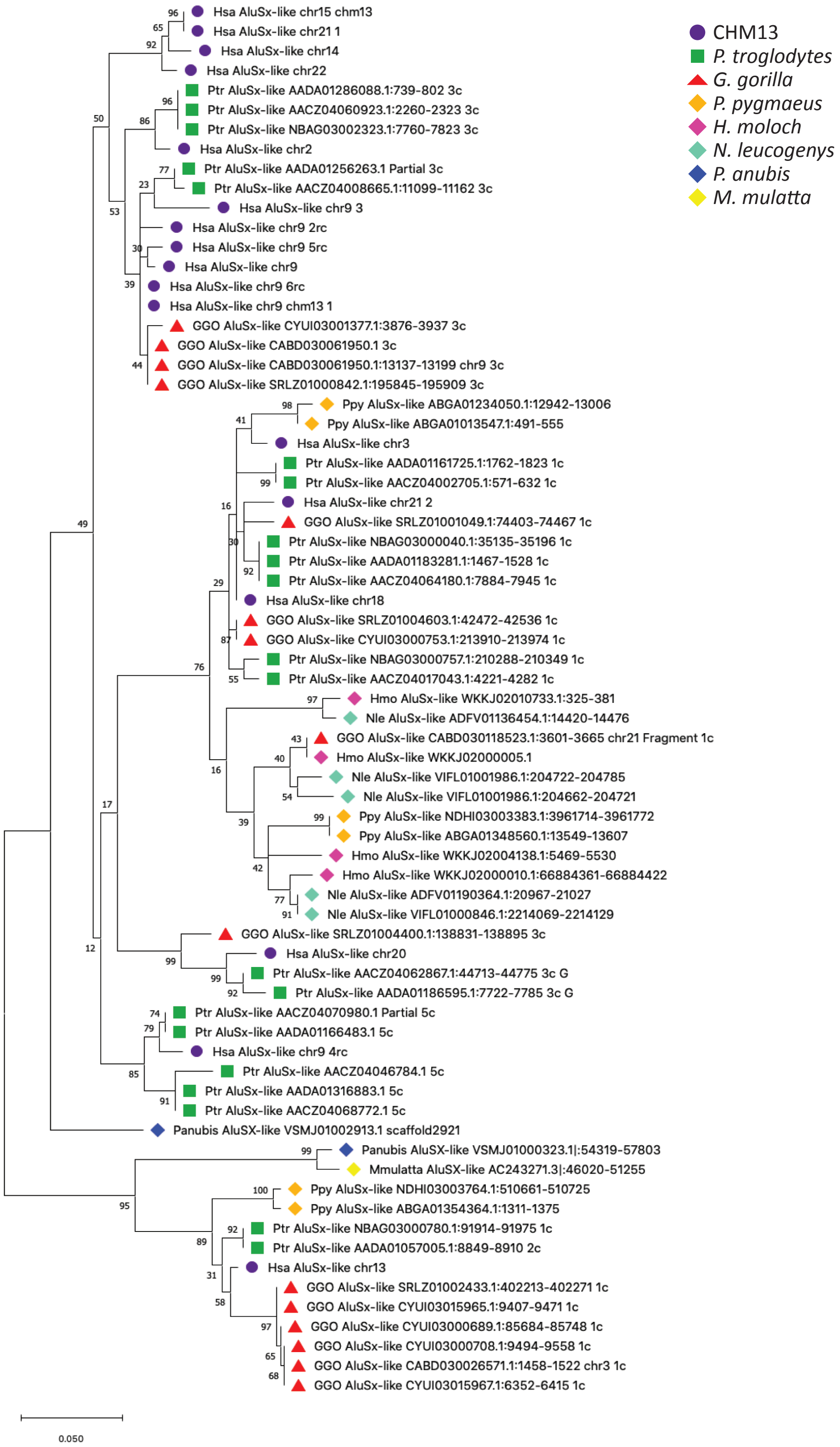


Fig. S45. Phylogenetic analyses identify the transduction of WaluSat by an *Alu*Sx element in the last shared common ancestor with Hominoidea and Catarrhini. Phylogenetic tree showing relationships among primates (Hominoidea, Catarrhini, Platyrrhini, and Prosimians). All primates shown carry the WaluSat novel satellite sequence as a solo locus. Catarrhini and Hominoidea (orange branches) show evidence of transduction of WaluSat by an *Alu*Sx3 element. Copy numbers of the WaluSat are indicated by a proportional blue circle to the right of the branch (n = number of WaluSats in a tandem array). This provides evidence for a human specific amplification through the evolution of hominids.



● CHM13
🟩 P. troglodytes
▲ G. gorilla
◆ P. pygmaeus
◆ H. moloch
◆ N. leucogenys
◆ P. anubis
◆ M. mulatta
96
Hsa AluSx-like chr15 chm13
65
Hsa AluSx-like chr21 1
92
Hsa AluSx-like chr14
Hsa AluSx-like chr22
Ptr AluSx-like AADA01286088.1:739-802 3c
96
Ptr AluSx-like AACZ04060923.1:2260-2323 3c
50
86
Ptr AluSx-like NBAG03002323.1:7760-7823 3c
Hsa AluSx-like chr2
77
Ptr AluSx-like AADA01256263.1 Partial 3c
23
Ptr AluSx-like AACZ04008665.1:11099-11162 3c
Hsa AluSx-like chr9 3
53
Hsa AluSx-like chr9 2rc
30
Hsa AluSx-like chr9 5rc
Hsa AluSx-like chr9
39
Hsa AluSx-like chr9 6rc
Hsa AluSx-like chr9 chm13 1
GGO AluSx-like CYUI03001377.1:3876-3937 3c
GGO AluSx-like CABD030061950.1 3c
44
GGO AluSx-like CABD030061950.1:13137-13199 chr9 3c
GGO AluSx-like SRLZ01000842.1:195845-195909 3c
98
Ppy AluSx-like ABGA01234050.1:12942-13006
41
Ppy AluSx-like ABGA01013547.1:491-555
Hsa AluSx-like chr3
Ptr AluSx-like AADA01161725.1:1762-1823 1c
99
Ptr AluSx-like AACZ04002705.1:571-632 1c
16
Hsa AluSx-like chr21 2
GGO AluSx-like SRLZ01001049.1:74403-74467 1c
49
30
Ptr AluSx-like NBAG03000040.1:35135-35196 1c
Ptr AluSx-like AADA01183281.1:1467-1528 1c
92
Ptr AluSx-like AACZ04064180.1:7884-7945 1c
29
Hsa AluSx-like chr18
GGO AluSx-like SRLZ01004603.1:42472-42536 1c
87
GGO AluSx-like CYUI03000753.1:213910-213974 1c
Ptr AluSx-like NBAG03000757.1:210288-210349 1c
76
55
Ptr AluSx-like AACZ04017043.1:4221-4282 1c
97
Hmo AluSx-like WKKJ02010733.1:325-381
Nle AluSx-like ADFV01136454.1:14420-14476
43
GGO AluSx-like CABD030118523.1:3601-3665 chr21 Fragment 1c
40
Hmo AluSx-like WKKJ02000005.1
16
Nle AluSx-like VIFL01001986.1:204722-204785
54
Nle AluSx-like VIFL01001986.1:204662-204721
17
99
Ppy AluSx-like NDHI03003383.1:3961714-3961772
39
Ppy AluSx-like ABGA01348560.1:13549-13607
Hmo AluSx-like WKKJ02004138.1:5469-5530
42
Hmo AluSx-like WKKJ02000010.1:66884361-66884422
77
Nle AluSx-like ADFV01190364.1:20967-21027
91
Nle AluSx-like VIFL01000846.1:2214069-2214129
GGO AluSx-like SRLZ01004400.1:138831-138895 3c
12
Hsa AluSx-like chr20
99
99
Ptr AluSx-like AACZ04062867.1:44713-44775 3c G
92
Ptr AluSx-like AADA01186595.1:7722-7785 3c G
74
Ptr AluSx-like AACZ04070980.1 Partial 5c
79
Ptr AluSx-like AADA01166483.1 5c
Hsa AluSx-like chr9 4rc
Ptr AluSx-like AACZ04046784.1 5c
85
Ptr AluSx-like AADA01316883.1 5c
91
Ptr AluSx-like AACZ04068772.1 5c
Panubis AluSX-like VSMJ01002913.1 scaffold2921
99
Panubis AluSX-like VSMJ01000323.1|:54319-57803
Mmulatta AluSX-like AC243271.3|:46020-51255
100
Ppy AluSx-like NDHI03003764.1:510661-510725
95
Ppy AluSx-like ABGA01354364.1:1311-1375
92
Ptr AluSx-like NBAG03000780.1:91914-91975 1c
89
Ptr AluSx-like AADA01057005.1:8849-8910 2c
Hsa AluSx-like chr13
31
GGO AluSx-like SRLZ01002433.1:402213-402271 1c
58
GGO AluSx-like CYUI03015965.1:9407-9471 1c
97
GGO AluSx-like CYUI03000689.1:85684-85748 1c
GGO AluSx-like CYUI03000708.1:9494-9558 1c
65
GGO AluSx-like CABD030026571.1:1458-1522 chr3 1c
68
GGO AluSx-like CYUI03015967.1:6352-6415 1c
0.050

Fig. S46. Maximum-Likelihood tree derived from 71 *Alu*Sx elements associated with WaluSat monomeric or expanded arrays in Catarrhini. Evolutionary origin and phylogenetic relationship of *Alu*Sx elements associated with WaluSat sequences are depicted as present in the Catarrhini common ancestor and having arisen at different times within Catarrhini lineage. The phylogenetic distribution of *Alu*Sx elements corroborates the hypothesis of an ancestral transduction event and recent expansion in human acrocentric chromosomes. The evolutionary history of *Alu*Sx elements was inferred by using the Maximum Likelihood method and T93 model. The tree with the highest log likelihood (-2155.25) is shown. All branches are labeled with the bootstrap values with n = 1000 replicates.

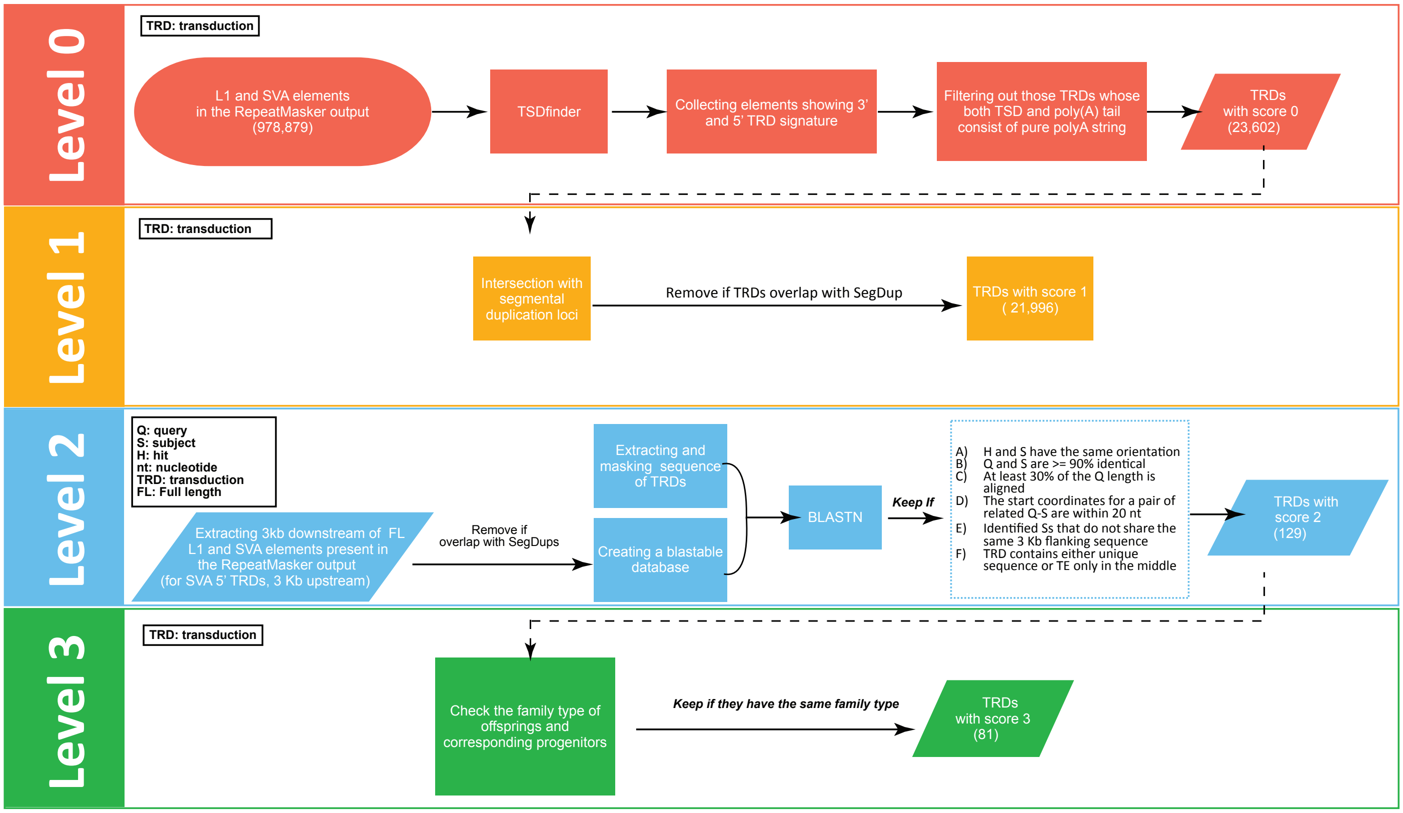

Level 0
TRD: transduction
L1 and SVA elements in the RepeatMasker output (978,879)
TSDfinder
Collecting elements showing 3' and 5' TRD signature
Filtering out those TRDs whose both TSD and poly(A) tail consist of pure polyA string
TRDs with score 0 (23,602)
Level 1
TRD: transduction
Intersection with segmental duplication loci
Remove if TRDs overlap with SegDup
TRDs with score 1 (21,996)
Level 2
Q: query
S: subject
H: hit
nt: nucleotide
TRD: transduction
FL: Full length
Extracting 3kb downstream of FL L1 and SVA elements present in the RepeatMasker output (for SVA 5' TRDs, 3 Kb upstream)
Remove if overlap with SegDups
Extracting and masking sequence of TRDs
Creating a blastable database
BLASTN
Keep If
A) H and S have the same orientation
B) Q and S are >= 90% identical
C) At least 30% of the Q length is aligned
D) The start coordinates for a pair of related Q-S are within 20 nt
E) Identified Ss that do not share the same 3 Kb flanking sequence
F) TRD contains either unique sequence or TE only in the middle
TRDs with score 2 (129)
Level 3
TRD: transduction
Check the family type of offsprings and corresponding progenitors
Keep if they have the same family type
TRDs with score 3 (81)

Fig. S47. Transduction analysis of T2T-CHM13 implemented four filtering steps to annotate high confidence events (Levels 1-3).

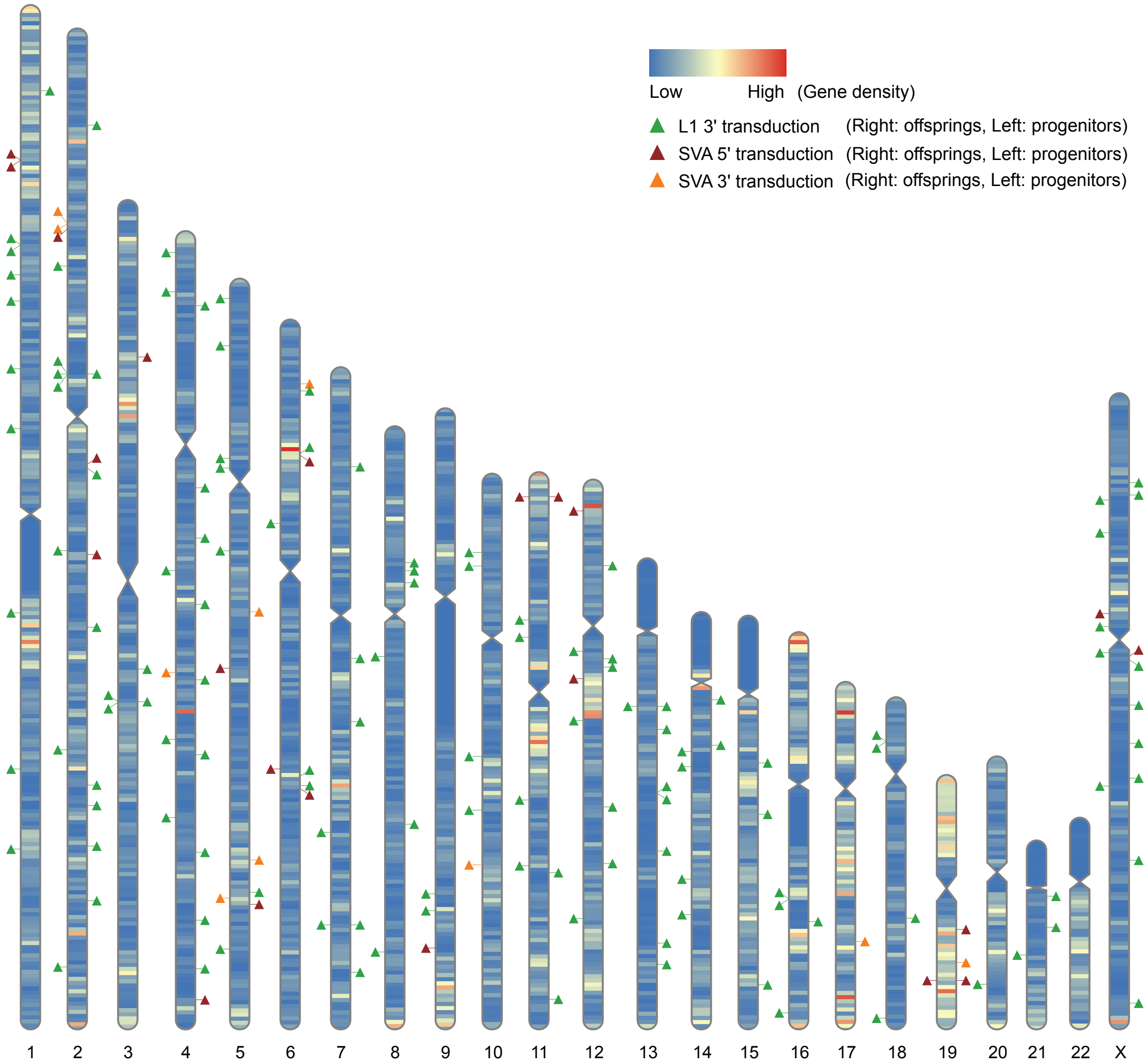


Figure S48. Transduction events are found genome wide in T2T-CHM13. Ideogram of CHM13 showing all transduction progenitors (left) and offspring (right) with respect to TE family (colored triangles) and gene density per 1Mbp bins.

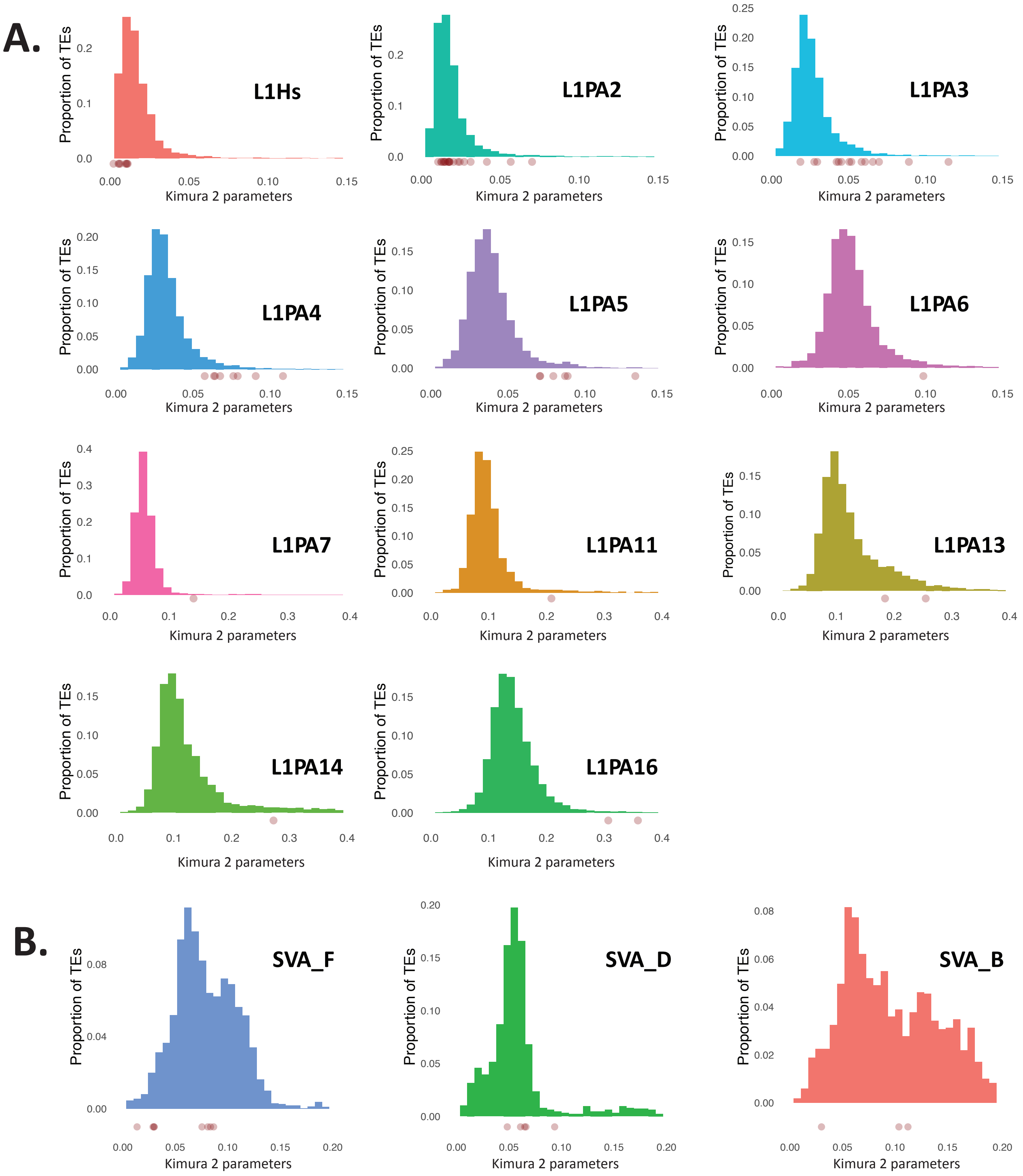


Fig. S49. Kimura 2-parameter distance for (A) L1 and (B) SVA from T2T-CHM13 (histogram) and between source and offspring TE (red dots).

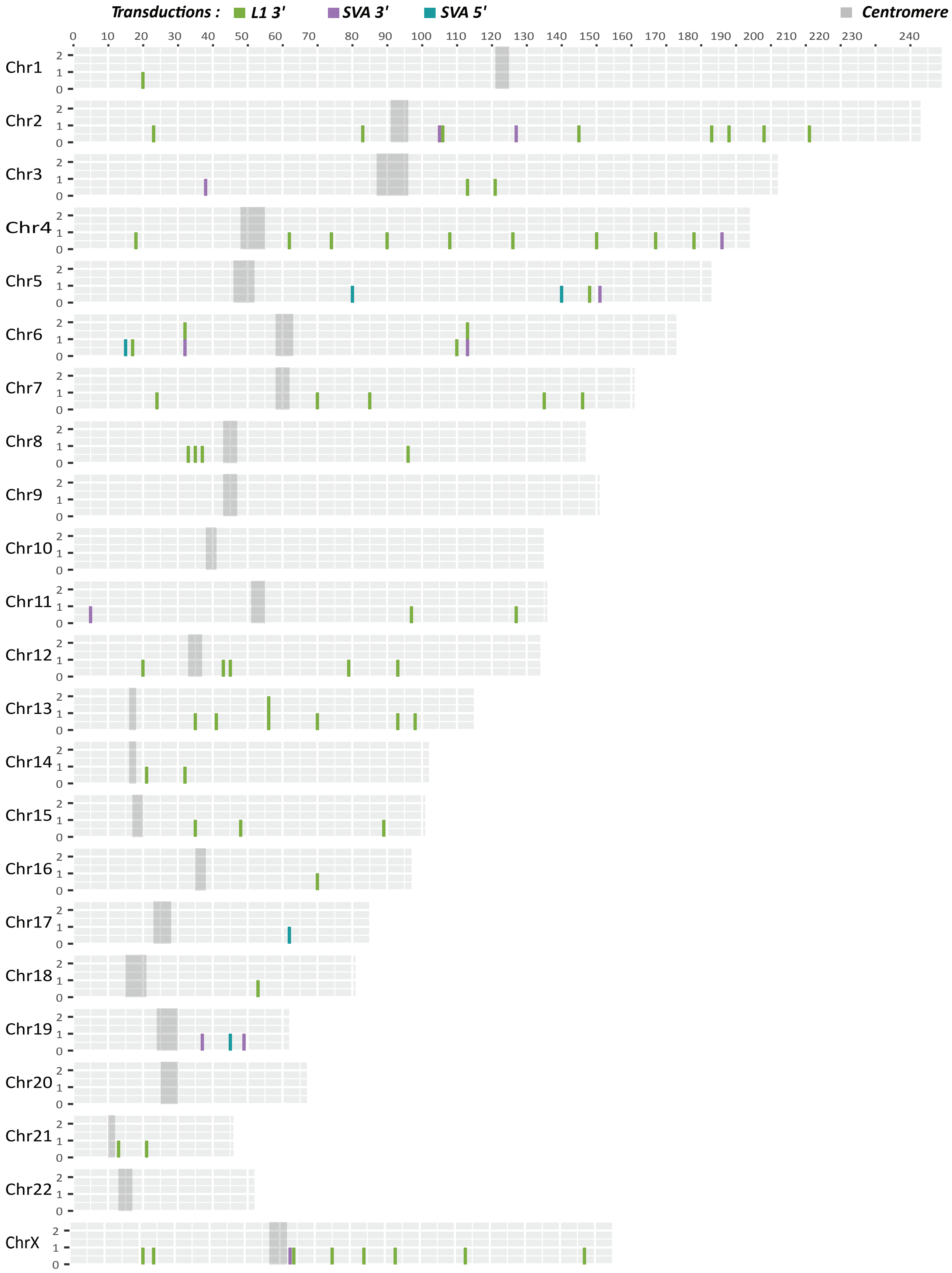


Fig. S50. Frequency of transduction events (offspring only) across each chromosome in T2T-CHM13. Element is indicated by color as per key at top.

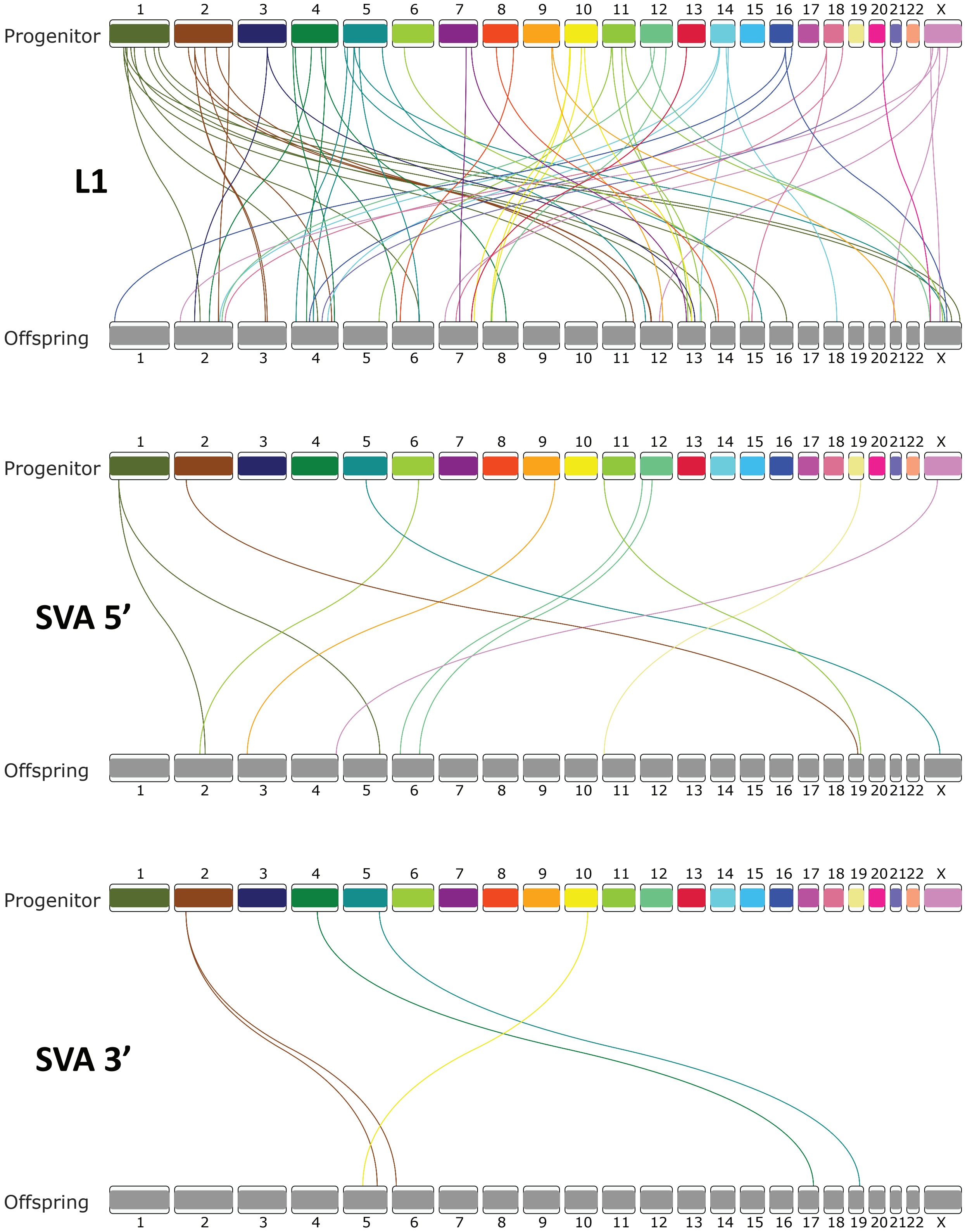


Fig. S51. Linkage graphs showing transduction progenitors and offspring across T2T-CHM13 for each repeat class (LINE, 5’ and 3’ SVA transduction events) as indicated.

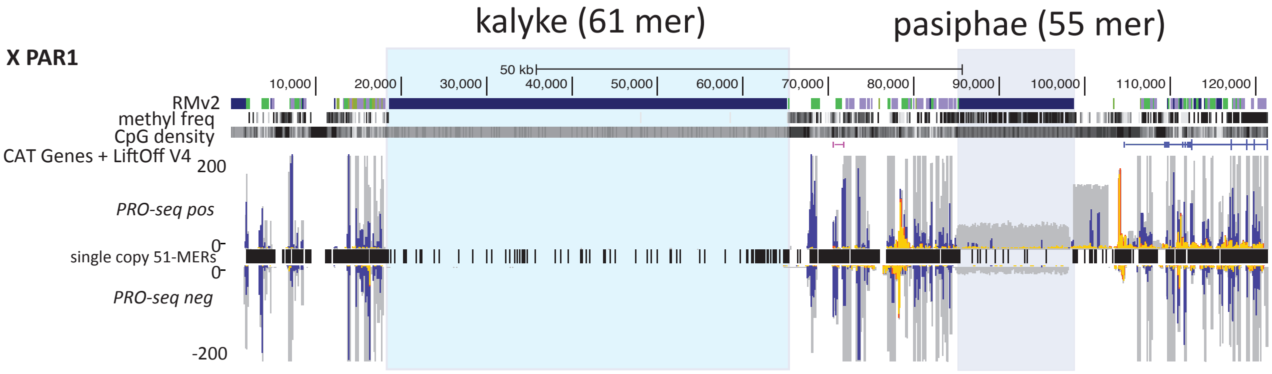
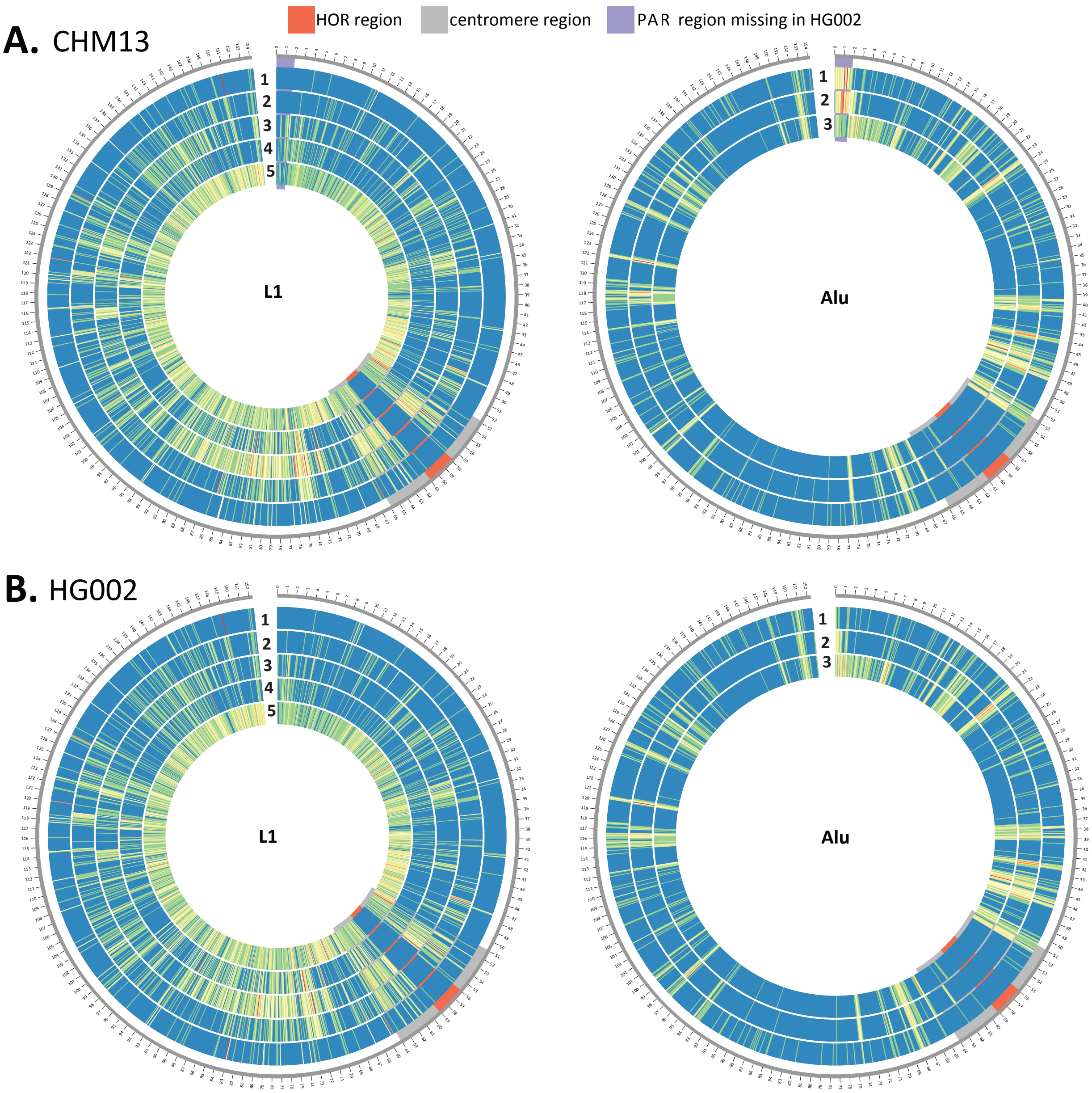


Fig. S52. Pseudoautosomal region (PAR) of the CHM13 X chromosome contains previously unknown repeat annotations. Included in these annotations are two tandem arrays, kalyke (61bp unit) and pasiphae (55bp unit). Kalyke has no detectable methylation while pasiphae has high levels of methylation coincident with some transcription (as detected by PRO-seq signal), indicating large arrays are differentially regulated on the X chromosome. PRO-seq tracks show Bowtie2 “best match” mapping (yellow), k-100 over-fit mapping (grey), and single (blue) and dual filtered (red) k-100 as per Fig. 4D-E.



Max number repeats per 100kb bin

young → old

	Track 1 L1HS	Track 2 L1PA2-3, L1P1	Track 3 L1PA4-17, L1P2-4, L1PREC2	Track 4 L1PB1-4, L1P5, L1PBa-b, L1P, L1PB	Track 5 L1M
🟦	0.6	1.9	6.3	2.9	16.3
🟩	1.1	3.7	12.6	5.7	32.6
🟨	1.7	5.6	18.9	8.6	48.9
🟨	2.3	7.4	25.1	11.4	65.1
🟨	2.9	9.3	31.4	14.3	81.4
🟧	3.4	11.1	37.7	17.1	97.7
🟥	4.0	13.0	44.0	20.0	114.0

Max number repeats per 100kb bin

young → old

	Track 1 AluY	Track 2 AluS	Track 3 AluJ
🟦	5.4	20.6	10.0
🟩	10.9	41.1	20.0
🟨	16.3	61.7	30.0
🟨	21.7	82.3	40.0
🟨	27.1	102.9	50.0
🟧	32.6	123.4	60.0
🟥	38.0	144.0	70.0

Fig. S53. ChrX density of L1 and Alu subfamilies across CHM13 and HG002. Subfamilies of L1s (LINE) and *Alu*s (SINE) were grouped into general evolutionary age groups from youngest, including potentially mobile ones (L1HS, *Alu*Y), to oldest (L1M, *Alu*J). Counts of these grouped subfamilies were binned into Kbp windows across chrX in **(A)** T2T-CHM13 and **(B)** HG002 and are shown as Circos heat-maps. Centromere blocks (including centromere transition regions) are denoted by grey bars, HORs are denoted by orange bars, and the portion of the PAR1 that remains unassembled in HG002 discordantis denoted in purple, all of which span all tracks. Tracks are numbered (1, 2, …) starting from the outer ring as indicated. In order to accurately compare density between the X's, each L1 or Alu subfamily track is shown on the same scale (i.e. T2T-CHM13 L1Hs and HG002 L1Hs) with the scales for each subfamily group located below each set. At this resolution, Alus are depleted in centromeric regions, possibly due to their enrichment in genic regions (*153*), but are found enriched in the PAR1 region with increased *Alu*Y density closer to the telomere. In contrast, L1s have some of the highest density peaks at and around the centromere.

References and Notes

1. E. B. Chuong, N. C. Elde, C. Feschotte, Regulatory activities of transposable elements: From conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017). doi:10.1038/nrg.2016.139 Medline

2. R. Cordaux, S. Udit, M. A. Batzer, C. Feschotte, Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 8101–8106 (2006). doi:10.1073/pnas.0601161103 Medline

3. E. V. Koonin, Viruses and mobile elements as drivers of evolutionary transitions. *Philos. Trans. R. Soc. London Ser. B* **371**, 20150442 (2016). doi:10.1098/rstb.2015.0442 Medline

4. A. Koga, H. Tanabe, Y. Hirai, H. Imai, M. Imamura, T. Oishi, R. Stanyon, H. Hirai, Co-opted megasatellite DNA drives evolution of secondary night vision in Azara's owl monkey. *Genome Biol. Evol.* **9**, 1963–1970 (2017). doi:10.1093/gbe/evx142 Medline

5. D. C. Hancks, H. H. Kazazian Jr., Roles for retrotransposon insertions in human disease. *Mob. DNA* **7**, 9 (2016). doi:10.1186/s13100-016-0065-9 Medline

6. J. Storer, R. Hubley, J. Rosen, T. J. Wheeler, A. F. Smit, The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 2 (2021). doi:10.1186/s13100-020-00230-y Medline

7. J. D. Fernandes, A. Zamudio-Hurtado, H. Clawson, W. J. Kent, D. Haussler, S. R. Salama, M. Haeussler, The UCSC repeat browser allows discovery and visualization of evolutionary conflict across repeat families. *Mob. DNA* **11**, 13 (2020). doi:10.1186/s13100-020-00208-w Medline

8. J. M. Flynn, R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, A. F. Smit, RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9451–9457 (2020). doi:10.1073/pnas.1921046117 Medline

9. D. Olson, T. Wheeler, ULTRA: A model based tool to detect tandem repeats. *ACM BCB* **2018**, 37–46 (2018). doi:10.1145/3233547.3233604 Medline

10. M. J. P. Chaisson, J. Huddleston, M. Y. Dennis, P. H. Sudmant, M. Malig, F. Hormozdiari, F. Antonacci, U. Surti, R. Sandstrom, M. Boitano, J. M. Landolin, J. A. Stamatoyannopoulos, M. W. Hunkapiller, J. Korlach, E. E. Eichler, Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015). doi:10.1038/nature13907 Medline

11. S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, S. Aganezov, S. J. Hoyt, M. Diekhans, G. A. Logsdon, M. Alonge, S. E. Antonarakis, M. Borchers, G. G. Bouffard, S. Y. Brooks, G. V. Caldas, H. Cheng, C.-S. Chin, W. Chow, L. G. de Lima, P. C. Dishuck, R. Durbin, T. Dvorkina, I. T. Fiddes, G. Formenti, R. S. Fulton, A. Fungtammasan, E. Garrison, P. G. S. Grady, T. A. Graves-Lindsay, I. M. Hall, N. F. Hansen, G. A. Hartley, M. Haukness, K. Howe, M. W. Hunkapiller, C. Jain, M. Jain, E. D. Jarvis, P. Kerpedjiev, M. Kirsche, M. Kolmogorov, J. Korlach, M. Kremitzki, H. Li, V. V. Maduro, T. Marschall, A. M. McCartney, J. McDaniel, D. E. Miller, J. C. Mullikin, E. W. Myers, N. D. Olson, B.

Paten, P. Peluso, P. A. Pevzner, D. Porubsky, T. Potapova, E. I. Rogaev, J. A. Rosenfeld, S. L. Salzberg, V. A. Schneider, F. J. Sedlazeck, K. Shafin, C. J. Shew, A. Shumate, Y. Sims, A. F. A. Smit, D. C. Soto, I. Sović, J. M. Storer, A. Streets, B. A. Sullivan, F. Thibaud-Nissen, J. Torrance, J. Wagner, B. P. Walenz, A. Wenger, J. M. D. Wood, C. Xiao, S. M. Yan, A. C. Young, S. Zarate, U. Surti, R. C. McCoy, M. Y. Dennis, I. A. Alexandrov, J. L. Gerton, R. J. O'Neill, W. Timp, J. M. Zook, M. C. Schatz, E. E. Eichler, K. H. Miga, A. M. Phillippy, The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

12. N. Altemose, G. A. Logsdon, A. V. Bzikadze, P. Sidhwani, S. A. Langley, G. V. Caldas, S. J. Hoyt, L. Uralsky, F. D. Ryabov, C. J. Shew, M. E. G. Sauria, M. Borchers, A. Gershman, A. Mikheenko, V. A. Shepelev, T. Dvorkina, O. Kunyavskaya, M. R. Vollger, A. Rhie, A. M. McCartney, M. Asri, R. Lorig-Roach, K. Shafin, S. Aganezov, D. Olson, L. Gomes de Lima, T. Potapova, G. A. Hartley, M. Haukness, P. Kerpedjiev, F. Gusev, K. Tigyi, S. Brooks, A. Young, S. Nurk, S. Koren, S. R. Salama, B. Paten, E. I. Rogaev, A. Streets, G. H. Karpen, A. F. Dernburg, B. A. Sullivan, A. F. Straight, T. J. Wheeler, J. L. Gerton, E. E. Eichler, A. M. Phillippy, W. Timp, M. Y. Dennis, R. J. O'Neill, J. M. Zook, M. C. Schatz, P. A. Pevzner, M. Diekhans, C. H. Langley, I. A. Alexandrov, K. H. Miga, Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).

13. Dfam, Transposable Element Classification; www.dfam.org/classification/tree.

14. B. Piégu, S. Bire, P. Arensburger, Y. Bigot, A survey of transposable element classification systems—A call for a fundamental update to meet the challenge of their diversity and complexity. *Mol. Phylogenet. Evol.* **86**, 90–109 (2015). doi:10.1016/j.ympev.2015.03.009 Medline

15. T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, A. H. Schulman, A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007). doi:10.1038/nrg2165 Medline

16. V. V. Kapitonov, J. Jurka, A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.* **9**, 411–412 (2008). doi:10.1038/nrg2165-c1 Medline

17. International Wheat Genome Sequencing Consortium (IWGSC), Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191 (2018). doi:10.1126/science.aar7191 Medline

18. See supplementary materials and methods.

19. K. M. Carey, G. Patterson, T. J. Wheeler, Transposable element subfamily annotation has a reproducibility problem. *Mob. DNA* **12**, 4 (2021). doi:10.1186/s13100-021-00232-4 Medline

20. P. Pajic, P. Pavlidis, K. Dean, L. Neznanova, R.-A. Romano, D. Garneau, E. Daugherity, A. Globig, S. Ruhl, O. Gokcumen, Independent amylase gene copy number bursts correlate with dietary preferences in mammals. *eLife* **8**, e44628 (2019). doi:10.7554/eLife.44628 Medline

21. A. Gershman, M. E.G. Sauria, P. W. Hook, S. J. Hoyt, R. Razaghi, S. Koren, N. Altemose, G. V. Caldas, M. R. Vollger, G. A. Logsdon, A. Rhie, E. E. Eichler, M. C. Schatz, R. J. O'Neill, A. M. Phillippy, K. H. Miga, W. Timp, Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).

22. D. B. Mahat, H. Kwak, G. T. Booth, I. H. Jonkers, C. G. Danko, R. K. Patel, C. T. Waters, K. Munson, L. J. Core, J. T. Lis, Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11**, 1455–1476 (2016). doi:10.1038/nprot.2016.086 Medline

23. H. Kwak, N. J. Fuda, L. J. Core, J. T. Lis, Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013). doi:10.1126/science.1229386 Medline

24. E. M. Wissink, A. Vihervaara, N. D. Tippens, J. T. Lis, Nascent RNA analyses: Tracking transcription and its regulation. *Nat. Rev. Genet.* **20**, 705–723 (2019). doi:10.1038/s41576-019-0159-6 Medline

25. J. O. Kriegs, G. Churakov, J. Jurka, J. Brosius, J. Schmitz, Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends Genet.* **23**, 158–161 (2007). doi:10.1016/j.tig.2007.02.002 Medline

26. P. J. Thompson, T. S. Macfarlan, M. C. Lorincz, Long terminal repeats: From parasitic elements to building blocks of the transcriptional regulatory repertoire. *Mol. Cell* **62**, 766–776 (2016). doi:10.1016/j.molcel.2016.03.029 Medline

27. J. Raiz, A. Damert, S. Chira, U. Held, S. Klawitter, M. Hamdorf, J. Löwer, W. H. Strätling, R. Löwer, G. G. Schumann, The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res.* **40**, 1666–1683 (2012). doi:10.1093/nar/gkr863 Medline

28. D. C. Hancks, J. L. Goodier, P. K. Mandal, L. E. Cheung, H. H. Kazazian Jr., Retrotransposition of marked SVA elements by human L1s in cultured cells. *Hum. Mol. Genet.* **20**, 3386–3400 (2011). doi:10.1093/hmg/ddr245 Medline

29. D. C. Hancks, H. H. Kazazian Jr., SVA retrotransposons: Evolution and genetic instability. *Semin. Cancer Biol.* **20**, 234–245 (2010). doi:10.1016/j.semcancer.2010.04.001 Medline

30. J. Pontis, E. Planet, S. Offner, P. Turelli, J. Duc, A. Coudray, T. W. Theunissen, R. Jaenisch, D. Trono, Hominoid-specific transposable elements and KZFPs facilitate human embryonic genome activation and control transcription in naive human ESCs. *Cell Stem Cell* **24**, 724–735.e5 (2019). doi:10.1016/j.stem.2019.03.012 Medline

31. C. R. Beck, J. L. Garcia-Perez, R. M. Badge, J. V. Moran, LINE-1 elements in structural variation and disease. *Annu. Rev. Genomics Hum. Genet.* **12**, 187–215 (2011). doi:10.1146/annurev-genom-082509-141802 Medline

32. G. D. Swergold, Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol. Cell. Biol.* **10**, 6718–6729 (1990). Medline

33. M. Sokolowski, M. Chynces, D. deHaro, C. M. Christian, V. P. Belancio, Truncated ORF1 proteins can suppress LINE-1 retrotransposition in *trans*. *Nucleic Acids Res.* **45**, 5294–5308 (2017). doi:10.1093/nar/gkx211 Medline

34. B. Brouha, J. Schustak, R. M. Badge, S. Lutz-Prigge, A. H. Farley, J. V. Moran, H. H. Kazazian Jr., Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 5280–5285 (2003). doi:10.1073/pnas.0831042100 Medline

35. K. Fatyol, K. Illies, A. A. Szalay, D. C. Diamond, C. Janish, Mer22-related sequence elements form pericentric repetitive DNA families in primates. *Mol. Gen. Genet.* **262**, 931–939 (2000). doi:10.1007/PL00008661 Medline

36. R. Nishiyama, L. Qi, K. Tsumagari, K. Weissbecker, L. Dubeau, M. Champagne, S. Sikka, H. Nagai, M. Ehrlich, A DNA repeat, NBL2, is hypermethylated in some cancers but hypomethylated in others. *Cancer Biol. Ther.* **4**, 446–454 (2005). doi:10.4161/cbt.4.4.1622 Medline

37. D. Thoraval, J. Asakawa, K. Wimmer, R. Kuick, B. Lamb, B. Richardson, P. Ambros, T. Glover, S. Hanash, Demethylation of repetitive DNA sequences in neuroblastoma. *Genes Chromosomes Cancer* **17**, 234–244 (1996). doi:10.1002/(SICI)1098-2264(199612)17:4<234:AID-GCC5>3.0.CO;2-4 Medline

38. D. C. Tremblay, G. Alexander Jr., S. Moseley, B. P. Chadwick, Expression, tandem repeat copy number variation and stability of four macrosatellite arrays in the human genome. *BMC Genomics* **11**, 632 (2010). doi:10.1186/1471-2164-11-632 Medline

39. J. K. Samuelsson, G. Dumbovic, C. Polo, C. Moreta, A. Alibés, T. Ruiz-Larroya, P. Giménez-Bonafé, S. Alonso, S.-V. Forcales, P. Manuel, Helicase lymphoid-specific enzyme contributes to the maintenance of methylation of SST1 pericentromeric repeats that are frequently demethylated in colon cancer and associate with genomic damage. *Epigenomes* **1**, 2 (2017). doi:10.3390/epigenomes1010002 Medline

40. S. Igarashi, H. Suzuki, T. Niinuma, H. Shimizu, M. Nojima, H. Iwaki, T. Nobuoka, T. Nishida, Y. Miyazaki, H. Takamaru, E. Yamamoto, H. Yamamoto, T. Tokino, T. Hasegawa, K. Hirata, K. Imai, M. Toyota, Y. Shinomura, A novel correlation between LINE-1 hypomethylation and the malignancy of gastrointestinal stromal tumors. *Clin. Cancer Res.* **16**, 5114–5123 (2010). doi:10.1158/1078-0432.CCR-10-0581 Medline

41. K. Suzuki, I. Suzuki, A. Leodolter, S. Alonso, S. Horiuchi, K. Yamashita, M. Perucho, Global DNA demethylation in gastrointestinal cancer is age dependent and precedes genomic damage. *Cancer Cell* **9**, 199–207 (2006). doi:10.1016/j.ccr.2006.02.016 Medline

42. H. Nagai, Y. S. Kim, T. Yasuda, Y. Ohmachi, H. Yokouchi, M. Monden, M. Emi, N. Konishi, M. Nogami, K. Okumura, K. Matsubara, A novel sperm-specific hypomethylation sequence is a demethylation hotspot in human hepatocellular carcinomas. *Gene* **237**, 15–20 (1999). doi:10.1016/S0378-1119(99)00322-4 Medline

43. G. A. Hartley, M. Okhovat, R. J. O'Neill, L. Carbone, Comparative analyses of gibbon centromeres reveal dynamic genus-specific shifts in repeat composition. *Mol. Biol. Evol.* **38**, 3972–3992 (2021). doi:10.1093/molbev/msab148 Medline

44. A. P. Bird, Gene number, noise reduction and biological complexity. *Trends Genet.* **11**, 94–100 (1995). doi:10.1016/S0168-9525(00)89009-5 Medline

45. J. A. Yoder, C. P. Walsh, T. H. Bestor, Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13**, 335–340 (1997). doi:10.1016/S0168-9525(97)01181-5 Medline

46. M. P. Ball, J. B. Li, Y. Gao, J.-H. Lee, E. M. LeProust, I.-H. Park, B. Xie, G. Q. Daley, G. M. Church, Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.* **27**, 361–368 (2009). doi:10.1038/nbt.1533 Medline

47. R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q.-M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, J. R. Ecker, Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009). doi:10.1038/nature08514 Medline

48. G. Dumbović, J. Biayna, J. Banús, J. Samuelsson, A. Roth, S. Diederichs, S. Alonso, M. Buschbeck, M. Perucho, S.-V. Forcales, A novel long non-coding RNA from NBL2 pericentromeric macrosatellite forms a perinucleolar aggregate structure in colon cancer. *Nucleic Acids Res.* **46**, 5504–5524 (2018). doi:10.1093/nar/gky263 Medline

49. J. Carlevaro-Fita, T. Polidori, M. Das, C. Navarro, T. I. Zoller, R. Johnson, Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. *Genome Res.* **29**, 208–222 (2019). doi:10.1101/gr.229922.117 Medline

50. B. González, M. Navarro-Jiménez, M. J. Alonso-De Gennaro, S. M. Jansen, I. Granada, M. Perucho, S. Alonso, Somatic hypomethylation of pericentromeric SST1 repeats and tetraploidization in human colorectal cancer cells. *Cancers* **13**, 5353 (2021). doi:10.3390/cancers13215353 Medline

51. G. O. M. Bobkov, N. Gilbert, P. Heun, Centromere transcription allows CENP-A to transit from chromatin association to stable incorporation. *J. Cell Biol.* **217**, 1957–1972 (2018). doi:10.1083/jcb.201611087 Medline

52. S. Rošić, F. Köhler, S. Erhardt, Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. *J. Cell Biol.* **207**, 335–349 (2014). doi:10.1083/jcb.201404097 Medline

53. D. M. Carone, C. Zhang, L. E. Hall, C. Obergfell, B. R. Carone, M. J. O'Neill, R. J. O'Neill, Hypermorphic expression of centromeric retroelement-encoded small RNAs impairs CENP-A loading. *Chromosome Res.* **21**, 49–62 (2013). doi:10.1007/s10577-013-9337-0 Medline

54. S. M. McNulty, L. L. Sullivan, B. A. Sullivan, Human centromeres produce chromosome-specific and array-specific alpha satellite transcripts that are complexed with CENP-A and CENP-C. *Dev. Cell* **42**, 226–240.e6 (2017). doi:10.1016/j.devcel.2017.07.001 Medline

55. S. Catania, A. L. Pidoux, R. C. Allshire, Sequence features and transcriptional stalling within centromere DNA promote establishment of CENP-A chromatin. *PLOS Genet.* **11**, e1004986 (2015). doi:10.1371/journal.pgen.1004986 Medline

56. C. C. Chen, S. Bowers, Z. Lipinszki, J. Palladino, S. Trusiak, E. Bettini, L. Rosin, M. R. Przewloka, D. M. Glover, R. J. O'Neill, B. G. Mellone, Establishment of centromeric

chromatin by the CENP-A assembly factor CAL1 requires FACT-mediated transcription. *Dev. Cell* **34**, 73–84 (2015). doi:10.1016/j.devcel.2015.05.012 Medline

57. A. C. Chueh, E. L. Northrop, K. H. Brettingham-Moore, K. H. A. Choo, L. H. Wong, LINE retrotransposon RNA is an essential structural and functional epigenetic component of a core neocentromeric chromatin. *PLOS Genet.* **5**, e1000354 (2009). doi:10.1371/journal.pgen.1000354 Medline

58. G. A. Logsdon, M. R. Vollger, P. Hsieh, Y. Mao, M. A. Liskovykh, S. Koren, S. Nurk, L. Mercuri, P. C. Dishuck, A. Rhie, L. G. de Lima, T. Dvorkina, D. Porubsky, W. T. Harvey, A. Mikheenko, A. V. Bzikadze, M. Kremitzki, T. A. Graves-Lindsay, C. Jain, K. Hoekzema, S. C. Murali, K. M. Munson, C. Baker, M. Sorensen, A. M. Lewis, U. Surti, J. L. Gerton, V. Larionov, M. Ventura, K. H. Miga, A. M. Phillippy, E. E. Eichler, The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021). doi:10.1038/s41586-021-03420-7 Medline

59. K. H. Miga, S. Koren, A. Rhie, M. R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G. A. Logsdon, V. A. Schneider, T. Potapova, J. Wood, W. Chow, J. Armstrong, J. Fredrickson, E. Pak, K. Tigyi, M. Kremitzki, C. Markovic, V. Maduro, A. Dutra, G. G. Bouffard, A. M. Chang, N. F. Hansen, A. B. Wilfert, F. Thibaud-Nissen, A. D. Schmitt, J.-M. Belton, S. Selvaraj, M. Y. Dennis, D. C. Soto, R. Sahasrabudhe, G. Kaya, J. Quick, N. J. Loman, N. Holmes, M. Loose, U. Surti, R. A. Risques, T. A. Graves Lindsay, R. Fulton, I. Hall, B. Paten, K. Howe, W. Timp, A. Young, J. C. Mullikin, P. A. Pevzner, J. L. Gerton, B. A. Sullivan, E. E. Eichler, A. M. Phillippy, Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020). doi:10.1038/s41586-020-2547-7 Medline

60. L. E. T. Jansen, B. E. Black, D. R. Foltz, D. W. Cleveland, Propagation of centromeric chromatin requires exit from mitosis. *J. Cell Biol.* **176**, 795–805 (2007). doi:10.1083/jcb.200701066 Medline

61. W. L. Johnson, W. T. Yewdell, J. C. Bell, S. M. McNulty, Z. Duda, R. J. O’Neill, B. A. Sullivan, A. F. Straight, RNA-dependent stabilization of SUV39H1 at constitutive heterochromatin. *eLife* **6**, e25299 (2017). doi:10.7554/eLife.25299 Medline

62. A. C. Chueh, L. H. Wong, N. Wong, K. H. A. Choo, Variable and hierarchical size distribution of L1-retroelement-enriched CENP-A clusters within a functional human neocentromere. *Hum. Mol. Genet.* **14**, 85–93 (2005). doi:10.1093/hmg/ddi008 Medline

63. R. J. O’Neill, M. J. O’Neill, J. A. Graves, Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* **393**, 68–72 (1998). doi:10.1038/29985 Medline

64. S. J. Klein, R. J. O’Neill, Transposable elements: Genome innovation, chromosome diversity, and centromere conflict. *Chromosome Res.* **26**, 5–23 (2018). doi:10.1007/s10577-017-9569-5 Medline

65. M. R. Vollger, X. Guitart, P. C. Dishuck, L. Mercuri, W. T. Harvey, A. Gershman, M. Diekhans, A. Sulovari, K. M. Munson, A. M. Lewis, K. Hoekzema, D. Porubsky, R. Li, S. Nurk, S. Koren, K. H. Miga, A. M. Phillippy, W. Timp, M. Ventura, E. E. Eichler, Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).

66. A. Rizzo, E. Salvati, M. Porru, C. D'Angelo, M. F. Stevens, M. D'Incalci, C. Leonetti, E. Gilson, G. Zupi, A. Biroccio, Stabilization of quadruplex DNA perturbs telomere replication leading to the activation of an ATR-dependent ATM signaling pathway. *Nucleic Acids Res.* **37**, 5353–5364 (2009). doi:10.1093/nar/gkp582 Medline

67. S. T. Szak, O. K. Pickeral, D. Landsman, J. D. Boeke, Identifying related L1 retrotransposons by analyzing 3′ transduced sequences. *Genome Biol.* **4**, R30 (2003). doi:10.1186/gb-2003-4-5-r30 Medline

68. J. L. Goodier, E. M. Ostertag, H. H. Kazazian Jr., Transduction of 3′-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9**, 653–657 (2000). doi:10.1093/hmg/9.4.653 Medline

69. O. K. Pickeral, W. Makałowski, M. S. Boguski, J. D. Boeke, Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res.* **10**, 411–415 (2000). doi:10.1101/gr.10.4.411 Medline

70. A. Damert, J. Raiz, A. V. Horn, J. Löwer, H. Wang, J. Xing, M. A. Batzer, R. Löwer, G. G. Schumann, 5′-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. *Genome Res.* **19**, 1992–2008 (2009). doi:10.1101/gr.093435.109 Medline

71. J. Xing, H. Wang, V. P. Belancio, R. Cordaux, P. L. Deininger, M. A. Batzer, Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 17608–17613 (2006). doi:10.1073/pnas.0603224103 Medline

72. E. J. Gardner, V. K. Lam, D. N. Harris, N. T. Chuang, E. C. Scott, W. S. Pittard, R. E. Mills, S. E. Devine; 1000 Genomes Project Consortium, The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017). doi:10.1101/gr.218032.116 Medline

73. P. Ebert, P. A. Audano, Q. Zhu, B. Rodriguez-Martin, D. Porubsky, M. J. Bonder, A. Sulovari, J. Ebler, W. Zhou, R. Serra Mari, F. Yilmaz, X. Zhao, P. Hsieh, J. Lee, S. Kumar, J. Lin, T. Rausch, Y. Chen, J. Ren, M. Santamarina, W. Höps, H. Ashraf, N. T. Chuang, X. Yang, K. M. Munson, A. P. Lewis, S. Fairley, L. J. Tallon, W. E. Clarke, A. O. Basile, M. Byrska-Bishop, A. Corvelo, U. S. Evani, T.-Y. Lu, M. J. P. Chaisson, J. Chen, C. Li, H. Brand, A. M. Wenger, M. Ghareghani, W. T. Harvey, B. Raeder, P. Hasenfeld, A. A. Regier, H. J. Abel, I. M. Hall, P. Flicek, O. Stegle, M. B. Gerstein, J. M. C. Tubio, Z. Mu, Y. I. Li, X. Shi, A. R. Hastie, K. Ye, Z. Chong, A. D. Sanders, M. C. Zody, M. E. Talkowski, R. E. Mills, S. E. Devine, C. Lee, J. O. Korbel, T. Marschall, E. E. Eichler, Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021). doi:10.1126/science.abf7117 Medline

74. J. M. C. Tubio, Y. Li, Y. S. Ju, I. Martincorena, S. L. Cooke, M. Tojo, G. Gundem, C. P. Pipinikas, J. Zamora, K. Raine, A. Menzies, P. Roman-Garcia, A. Fullam, M. Gerstung, A. Shlien, P. S. Tarpey, E. Papaemmanuil, S. Knappskog, P. Van Loo, M. Ramakrishna, H. R. Davies, J. Marshall, D. C. Wedge, J. W. Teague, A. P. Butler, S. Nik-Zainal, L. Alexandrov, S. Behjati, L. R. Yates, N. Bolli, L. Mudie, C. Hardy, S. Martin, S. McLaren, S. O'Meara, E. Anderson, M. Maddison, S. Gamble, C. Foster, A. Y. Warren, H. Whitaker, D. Brewer, R. Eeles, C. Cooper, D. Neal, A. G. Lynch, T. Visakorpi, W. B. Isaacs, L. V. Veer, C. Caldas, C. Desmedt, C. Sotiriou, S. Aparicio, J. A. Foekens, J. E.

Eyfjörd, S. R. Lakhani, G. Thomas, O. Myklebost, P. N. Span, A.-L. Børresen-Dale, A. L. Richardson, M. Van de Vijver, A. Vincent-Salomon, G. G. Van den Eynden, A. M. Flanagan, P. A. Futreal, S. M. Janes, G. S. Bova, M. R. Stratton, U. McDermott, P. J. Campbell; ICGC Breast Cancer Group; ICGC Bone Cancer Group; ICGC Prostate Cancer Group, Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014). doi:10.1126/science.1251343 Medline

75. B. Pradhan, T. Cajuso, R. Katainen, P. Sulo, T. Tanskanen, O. Kilpivaara, E. Pitkänen, L. A. Aaltonen, L. Kauppi, K. Palin, Detection of subclonal L1 transductions in colorectal cancer by long-distance inverse-PCR and Nanopore sequencing. *Sci. Rep.* **7**, 14521 (2017). doi:10.1038/s41598-017-15076-3 Medline

76. J. V. Moran, R. J. DeBerardinis, H. H. Kazazian Jr., Exon shuffling by L1 retrotransposition. *Science* **283**, 1530–1534 (1999). doi:10.1126/science.283.5407.1530 Medline

77. J. C. Chow, C. Ciaudo, M. J. Fazzari, N. Mise, N. Servant, J. L. Glass, M. Attreed, P. Avner, A. Wutz, E. Barillot, J. M. Greally, O. Voinnet, E. Heard, LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell* **141**, 956–969 (2010). doi:10.1016/j.cell.2010.04.042 Medline

78. M. F. Lyon, X-chromosome inactivation: A repeat hypothesis. *Cytogenet. Cell Genet.* **80**, 133–137 (1998). doi:10.1159/000014969 Medline

79. J. Chaumeil, P. Le Baccon, A. Wutz, E. Heard, A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes Dev.* **20**, 2223–2237 (2006). doi:10.1101/gad.380906 Medline

80. N. Stavropoulos, N. Lu, J. T. Lee, A functional role for *Tsix* transcription in blocking *Xist* RNA accumulation but not in X-chromosome choice. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 10232–10237 (2001). doi:10.1073/pnas.171243598 Medline

81. E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Blüthgen, J. Dekker, E. Heard, Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012). doi:10.1038/nature11049 Medline

82. J. G. van Bemmel, R. Galupa, C. Gard, N. Servant, C. Picard, J. Davies, A. J. Szempruch, Y. Zhan, J. J. Żylicz, E. P. Nora, S. Lameiras, E. de Wit, D. Gentien, S. Baulande, L. Giorgetti, M. Guttman, J. R. Hughes, D. R. Higgs, J. Gribnau, E. Heard, The bipartite TAD organization of the X-inactivation center ensures opposing developmental regulation of *Tsix* and *Xist*. *Nat. Genet.* **51**, 1024–1034 (2019). doi:10.1038/s41588-019-0412-0 Medline

83. X. Chen, Y. Ma, L. Wang, X. Zhang, Y. Yu, W. Lü, X. Xie, X. Cheng, Loss of X chromosome inactivation in androgenetic complete hydatidiform moles with 46, XX karyotype. *Int. J. Gynecol. Pathol.* **40**, 333–341 (2021). doi:10.1097/PGP.0000000000000697 Medline

84. Z. N. Kronenberg, I. T. Fiddes, D. Gordon, S. Murali, S. Cantsilieris, O. S. Meyerson, J. G. Underwood, B. J. Nelson, M. J. P. Chaisson, M. L. Dougherty, K. M. Munson, A. R. Hastie, M. Diekhans, F. Hormozdiari, N. Lorusso, K. Hoekzema, R. Qiu, K. Clark, A.

Raja, A. E. Welch, M. Sorensen, C. Baker, R. S. Fulton, J. Armstrong, T. A. Graves-Lindsay, A. M. Denli, E. R. Hoppe, P. Hsieh, C. M. Hill, A. W. C. Pang, J. Lee, E. T. Lam, S. K. Dutcher, F. H. Gage, W. C. Warren, J. Shendure, D. Haussler, V. A. Schneider, H. Cao, M. Ventura, R. K. Wilson, B. Paten, A. Pollen, E. E. Eichler, High-resolution comparative analysis of great ape genomes. *Science* **360**, eaar6343 (2018). doi:10.1126/science.aar6343 Medline

85. H. A. Lewin, G. E. Robinson, W. J. Kress, W. J. Baker, J. Coddington, K. A. Crandall, R. Durbin, S. V. Edwards, F. Forest, M. T. P. Gilbert, M. M. Goldstein, I. V. Grigoriev, K. J. Hackett, D. Haussler, E. D. Jarvis, W. E. Johnson, A. Patrinos, S. Richards, J. C. Castilla-Rubio, M.-A. van Sluys, P. S. Soltis, X. Xu, H. Yang, G. Zhang, Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 4325–4333 (2018). doi:10.1073/pnas.1720115115 Medline

86. W. M. Guiblet, M. DeGiorgio, X. Cheng, F. Chiaromonte, K. A. Eckert, Y.-F. Huang, K. D. Makova, Selection and thermostability suggest G-quadruplexes are novel functional elements of the human genome. *Genome Res.* **31**, 1136–1149 (2021). doi:10.1101/gr.269589.120 Medline

87. C. Roden, A. S. Gladfelter, RNA contributions to the form and function of biomolecular condensates. *Nat. Rev. Mol. Cell Biol.* **22**, 183–195 (2021). doi:10.1038/s41580-020-0264-6 Medline

88. A. F. A. Smit, R. Hubley, P. Green, RepeatMasker Open-4.0, 2013–2015; www.repeatmasker.org.

89. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). doi:10.1093/bioinformatics/btq033 Medline

90. L. Zhang, H. H. S. Lu, W.-Y. Chung, J. Yang, W.-H. Li, Patterns of segmental duplication in the human genome. *Mol. Biol. Evol.* **22**, 135–141 (2005). doi:10.1093/molbev/msh262 Medline

91. H. Pagès, BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs, R package version 1.62.0; https://bioconductor.org/packages/BSgenome.

92. L. Scrucca, M. Fop, T. B. Murphy, A. E. Raftery, Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **8**, 289–317 (2016). doi:10.32614/RJ-2016-021 Medline

93. J. Xing, Y. Zhang, K. Han, A. H. Salem, S. K. Sen, C. D. Huff, Q. Zhou, E. F. Kirkness, S. Levy, M. A. Batzer, L. B. Jorde, Mobile elements create structural variation: Analysis of a complete human genome. *Genome Res.* **19**, 1516–1526 (2009). doi:10.1101/gr.091827.109 Medline

94. H. Khan, A. Smit, S. Boissinot, Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res.* **16**, 78–87 (2006). doi:10.1101/gr.4001406 Medline

95. A. F. Smit, G. Tóth, A. D. Riggs, J. Jurka, Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.* **246**, 401–417 (1995). doi:10.1006/jmbi.1994.0095 Medline

96. A. V. Furano, D. D. Duvernell, S. Boissinot, L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. *Trends Genet.* **20**, 9–14 (2004). doi:10.1016/j.tig.2003.11.006 Medline

97. M. K. Konkel, J. A. Walker, M. A. Batzer, LINEs and SINEs of primate evolution. *Evol. Anthropol.* **19**, 236–249 (2010). doi:10.1002/evan.20283 Medline

98. H. Wang, J. Xing, D. Grover, D. J. Hedges, K. Han, J. A. Walker, M. A. Batzer, SVA elements: A hominid-specific retroposon family. *J. Mol. Biol.* **354**, 994–1007 (2005). doi:10.1016/j.jmb.2005.09.085 Medline

99. G. L. Freimanis, thesis, University of Wolverhampton (2008).

100. J. Judd, L. A. Wojenski, L. M. Wainman, N. D. Tippens, E. J. Rice, A. Dziubek, G. J. Villafano, E. M. Wissink, P. Versluis, L. Bagepalli, S. R. Shah, D. B. Mahat, J. M. Tome, C. G. Danko, J. T. Lis, L. J. Core, A rapid, sensitive, scalable method for Precision Run-On sequencing (PRO-seq). bioRxiv 2020.05.18.102277 [Preprint] (2020). https://doi.org/10.1101/2020.05.18.102277.

101. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011). doi:10.14806/ej.17.1.200

102. Fastx-toolkit, FASTQ/A short-reads preprocessing tools; http://hannonlab. cshl. edu/fastx_toolkit.

103. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012). doi:10.1038/nmeth.1923 Medline

104. Z. Hao, D. Lv, Y. Ge, J. Shi, D. Weijers, G. Yu, J. Chen, *RIdeogram*: Drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput. Sci.* **6**, e251 (2020). doi:10.7717/peerj-cs.251 Medline

105. M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, M. A. Marra, Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009). doi:10.1101/gr.092759.109 Medline

106. F. Ramírez, D. P. Ryan, B. Grüning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dündar, T. Manke, deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016). doi:10.1093/nar/gkw257 Medline

107. K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002). doi:10.1093/nar/gkf436 Medline

108. A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019). doi:10.1093/bioinformatics/btz305 Medline

109. M. Nei, S. Kumar, *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, 2000).

110. D. Darriba, G. L. Taboada, R. Doallo, D. Posada, jModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* **9**, 772 (2012). doi:10.1038/nmeth.2109 Medline

111. A. Frattini, M. Fabbri, R. Valli, E. De Paoli, G. Montalbano, L. Gribaldo, F. Pasquali, E. Maserati, High variability of genomic instability and gene expression profiling in different HeLa clones. *Sci. Rep.* **5**, 15377 (2015). doi:10.1038/srep15377 Medline

112. M. L. Whitfield, G. Sherlock, A. J. Saldanha, J. I. Murray, C. A. Ball, K. E. Alexander, J. C. Matese, C. M. Perou, M. M. Hurt, P. O. Brown, D. Botstein, Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13**, 1977–2000 (2002). doi:10.1091/mbc.02-02-0030 Medline

113. Y. H. Hung, S. Huang, M. K. Dame, Q. Yu, Q. C. Yu, Y. A. Zeng, J. G. Camp, J. R. Spence, P. Sethupathy, Chromatin regulatory dynamics of early human small intestinal development using a directed differentiation model. *Nucleic Acids Res.* **49**, 726–744 (2021). doi:10.1093/nar/gkaa1204 Medline

114. A. M. M. Cartney, K. Shafin, M. Alonge, A. V. Bzikadze, G. Formenti, A. Fungtammasan, K. Howe, C. Jain, S. Koren, G. A. Logsdon, K. H. Miga, A. Mikheenko, B. Paten, A. Shumate, D. C. Soto, I. Sović, J. M. D. Wood, J. M. Zook, A. M. Phillippy, A. Rhie, Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. bioRxiv 2021.07.02.450803 [Preprint] (2021). https://doi.org/10.1101/2021.07.02.450803.

115. V. Brázda, J. Kolomazník, J. Lýsek, M. Bartas, M. Fojta, J. Šťastný, J.-L. Mergny, G4Hunter web application: A web server for G-quadruplex prediction. *Bioinformatics* **35**, 3493–3495 (2019). doi:10.1093/bioinformatics/btz087 Medline

116. D. Gordon, J. Huddleston, M. J. P. Chaisson, C. M. Hill, Z. N. Kronenberg, K. M. Munson, M. Malig, A. Raja, I. Fiddes, L. W. Hillier, C. Dunn, C. Baker, J. Armstrong, M. Diekhans, B. Paten, J. Shendure, R. K. Wilson, D. Haussler, C.-S. Chin, E. E. Eichler, Long-read sequence assembly of the gorilla genome. *Science* **352**, aae0344 (2016). doi:10.1126/science.aae0344 Medline

117. Y. He, X. Luo, B. Zhou, T. Hu, X. Meng, P. A. Audano, Z. N. Kronenberg, E. E. Eichler, J. Jin, Y. Guo, Y. Yang, X. Qi, B. Su, Long-read assembly of the Chinese rhesus macaque genome and identification of ape-specific structural variants. *Nat. Commun.* **10**, 4233 (2019). doi:10.1038/s41467-019-12174-w Medline

118. L. Wang, J. Wu, X. Liu, D. Di, Y. Liang, Y. Feng, S. Zhang, B. Li, X.-G. Qi, A high-quality genome assembly for the endangered golden snub-nosed monkey (*Rhinopithecus roxellana*). *Gigascience* **8**, giz098 (2019). doi:10.1093/gigascience/giz098 Medline

119. V. Jayakumar, H. Ishii, M. Seki, W. Kumita, T. Inoue, S. Hase, K. Sato, H. Okano, E. Sasaki, Y. Sakakibara, An improved de novo genome assembly of the common marmoset genome yields improved contiguity and increased mapping rates of sequence data. *BMC Genomics* **21** (suppl. 3), 243 (2020). doi:10.1186/s12864-020-6657-2 Medline

120. P. A. Larsen, R. A. Harris, Y. Liu, S. C. Murali, C. R. Campbell, A. D. Brown, B. A. Sullivan, J. Shelton, S. J. Brown, M. Raveendran, O. Dudchenko, I. Machol, N. C. Durand, M. S. Shamim, E. L. Aiden, D. M. Muzny, R. A. Gibbs, A. D. Yoder, J. Rogers, K. C. Worley, Hybrid de novo genome assembly and centromere characterization of the gray mouse lemur (*Microcebus murinus*). *BMC Biol.* **15**, 110 (2017). doi:10.1186/s12915-017-0439-6 Medline

121. R. Bandyopadhyay, C. McQuillan, S. L. Page, K. H. Choo, L. G. Shaffer, Identification and characterization of satellite III subfamilies to the acrocentric chromosomes. *Chromosome Res.* **9**, 223–233 (2001). doi:10.1023/A:1016648404388 Medline

122. S. J. Hoyt, J. M. Storer, G. A. Hartley, P. G. S. Grady, A. Gershman, C. Limouse, R. Halabian, L. Wojenski, R. J. O'Neill, From telomere to telomere: the transcriptional and epigenetic state of human repeat elements analysis code: T2T-CHM13, Zenodo (2022); https://zenodo.org/record/5537106.

123. D. Ellinghaus, S. Kurtz, U. Willhoeft, *LTRharvest*, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008). doi:10.1186/1471-2105-9-18 Medline

124. TransposonPSI: An Application of PSI-Blast to Mine (Retro-)Transposon ORF Homologies, http://transposonpsi.sourceforge.net.

125. G. Benson, Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999). doi:10.1093/nar/27.2.573 Medline

126. L. Noé, G. Kucherov, YASS: Enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res.* **33**, W540–W543 (2005). doi:10.1093/nar/gki478 Medline

127. T. J. Wheeler, J. Clements, S. R. Eddy, R. Hubley, T. A. Jones, J. Jurka, A. F. A. Smit, R. D. Finn, Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.* **41**, D70–D82 (2013). doi:10.1093/nar/gks1265 Medline

128. R. Hubley, R. D. Finn, J. Clements, S. R. Eddy, T. A. Jones, W. Bao, A. F. A. Smit, T. J. Wheeler, The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2016). doi:10.1093/nar/gkv1272 Medline

129. A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, W. J. Kent, The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006). doi:10.1093/nar/gkj144 Medline

130. Y. Gondo, T. Okada, N. Matsuyama, Y. Saitoh, Y. Yanagisawa, J. E. Ikeda, Human megasatellite DNA RS447: Copy-number polymorphisms and interspecies conservation. *Genomics* **54**, 39–49 (1998). doi:10.1006/geno.1998.5545 Medline

131. P. E. Warburton, D. Hasson, F. Guillem, C. Lescale, X. Jin, G. Abrusan, Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics* **9**, 533 (2008). doi:10.1186/1471-2164-9-533 Medline

132. S. Baaj, G. Rafidi, J. Krueger, E. Chan, E. Vicker, A. Elfar, J. L. Doering, Organization of human 6kb tandem repeat. *Repbase Rep.* **14**, 2086 (2014).

133. J. E. Hewitt, R. Lyle, L. N. Clark, E. M. Valleley, T. J. Wright, C. Wijmenga, J. C. T. van Deutekom, F. Francis, P. T. Sharpe, M. Hofker, R. R. Frants, R. Williamson, Analysis of the tandem repeat locus D4Z4 associated with facioscapulohumeral muscular dystrophy. *Hum. Mol. Genet.* **3**, 1287–1295 (1994). doi:10.1093/hmg/3.8.1287 Medline

134. R. Meneveri, A. Agresti, A. Marozzi, S. Saccone, M. Rocchi, N. Archidiacono, G. Corneo, G. Della Valle, E. Ginelli, Molecular organization and chromosomal location of human GC-rich heterochromatic blocks. *Gene* **123**, 227–234 (1993). doi:10.1016/0378-1119(93)90128-P Medline

135. C. Wijmenga, J. E. Hewitt, L. A. Sandkuijl, L. N. Clark, T. J. Wright, H. G. Dauwerse, A.-M. Gruter, M. H. Hofker, P. Moerer, R. Williamson, G.-J. B. van Ommen, G. W. Padberg, R. R. Frants, Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. *Nat. Genet.* **2**, 26–30 (1992). doi:10.1038/ng0992-26 Medline

136. R. C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004). doi:10.1186/1471-2105-5-113 Medline

137. S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018). doi:10.1093/molbev/msy096 Medline

138. G. Stecher, K. Tamura, S. Kumar, Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Mol. Biol. Evol.* **37**, 1237–1239 (2020). doi:10.1093/molbev/msz312 Medline

139. M. R. Vollger, P. Kerpedjiev, A. M. Phillippy, E. E. Eichler, StainedGlass: Interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics* 10.1093/bioinformatics/btac018 (2022). doi:10.1093/bioinformatics/btac018 Medline

140. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018). doi:10.1093/bioinformatics/bty191 Medline

141. M. R. Vollger, P. Kerpedjiev, mrvollger/StainedGlass, Zenodo (2022); https://zenodo.org/record/5576601.

142. SAS Institute Inc., JMP; www.jmp.com.

143. T. Chu, Z. Wang, S.-P. Chou, C. G. Danko, Discovering transcriptional regulatory elements from run-on and sequencing data using the web-based dREG gateway. *Curr. Protoc. Bioinformatics* **66**, e70 (2019). doi:10.1002/cpbi.70 Medline

144. A. Rhie, B. P. Walenz, S. Koren, A. M. Phillippy, Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020). doi:10.1186/s13059-020-02134-9 Medline

145. Y. Fofanov, Y. Luo, C. Katili, J. Wang, Y. Belosludtsev, T. Powdrill, C. Belapurkar, V. Fofanov, T.-B. Li, S. Chumakov, B. M. Pettitt, How independent are the appearances of *n*-mers in different genomes? *Bioinformatics* **20**, 2421–2428 (2004). doi:10.1093/bioinformatics/bth266 Medline

146. M. Kokot, M. Dlugosz, S. Deorowicz, KMC 3: Counting and manipulating *k*-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017). doi:10.1093/bioinformatics/btx304 Medline

147. K. M. Seibt, T. Schmidt, T. Heitkam, FlexiDot: Highly customizable, ambiguity-aware dotplots for visual sequence analyses. *Bioinformatics* **34**, 3575–3577 (2018). doi:10.1093/bioinformatics/bty395 Medline

148. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990). doi:10.1016/S0022-2836(05)80360-2 Medline

149. J. Tica, E. Lee, A. Untergasser, S. Meiers, D. A. Garfield, O. Gokcumen, E. E. M. Furlong, P. J. Park, A. M. Stütz, J. O. Korbel, Next-generation sequencing-based detection of germline L1-mediated transductions. *BMC Genomics* **17**, 342 (2016). doi:10.1186/s12864-016-2670-x Medline

150. M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980). doi:10.1007/BF01731581 Medline

151. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004). doi:10.1093/bioinformatics/btg412 Medline

152. P. Hallast, L. Kibena, M. Punab, E. Arciero, S. Rootsi, M. Grigorova, R. Flores, M. A. Jobling, O. Poolamets, K. Pomm, P. Korrovits, K. Rull, Y. Xue, C. Tyler-Smith, M. Laan, A common 1.6 mb Y-chromosomal inversion predisposes to subsequent deletions and severe spermatogenic failure in humans. *eLife* **10**, e65420 (2021). doi:10.7554/eLife.65420 Medline

153. L.-L. Chen, L. Yang, *ALU*ternative Regulation for Gene Expression. *Trends Cell Biol.* **27**, 480–490 (2017). doi:10.1016/j.tcb.2017.01.002 Medline

154. A. R. Jha, D. F. Nixon, M. G. Rosenberg, J. N. Martin, S. G. Deeks, R. R. Hudson, K. E. Garrison, S. K. Pillai, Human endogenous retrovirus K106 (HERV-K106) was infectious after the emergence of anatomically modern humans. *PLOS ONE* **6**, e20234 (2011). doi:10.1371/journal.pone.0020234 Medline

155. J. Lee, R. Cordaux, K. Han, J. Wang, D. J. Hedges, P. Liang, M. A. Batzer, Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons. *Gene* **390**, 18–27 (2007). doi:10.1016/j.gene.2006.08.029 Medline