Genomic and epigenomic maps of mouse centromeres and pericentromeres

Gitika Choudhary[1#], Jingyue Chen[1#], Lucy Snipes[1], Smriti Bahl[1], Jenika Packiaraj[1], Leo Lin[1], Jitendra Thakur[1]

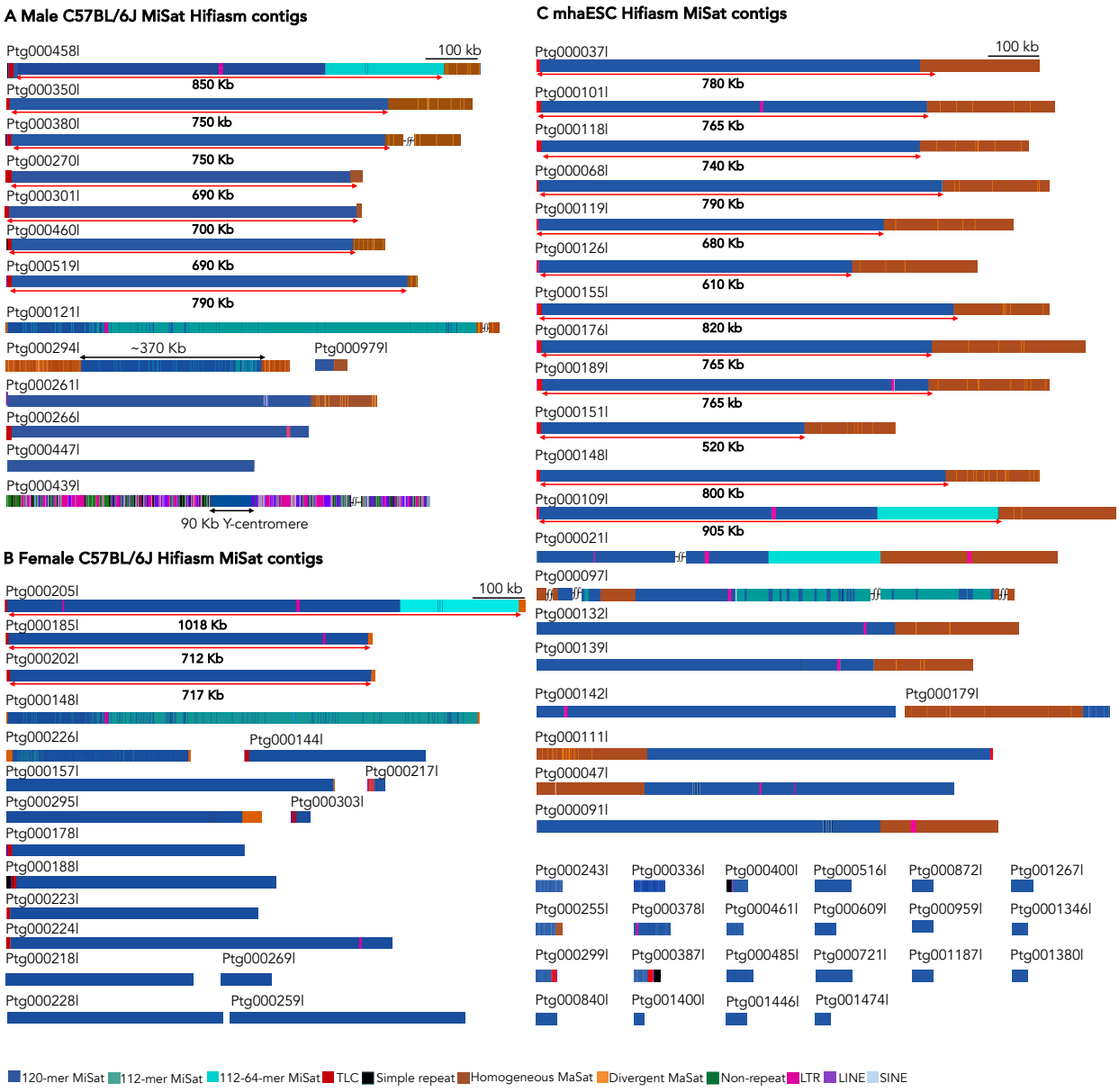
[#]These authors contributed equally.

Supplementary Information

Composition and density of non-satellite repeats at centromeric and pericentromeric regions.

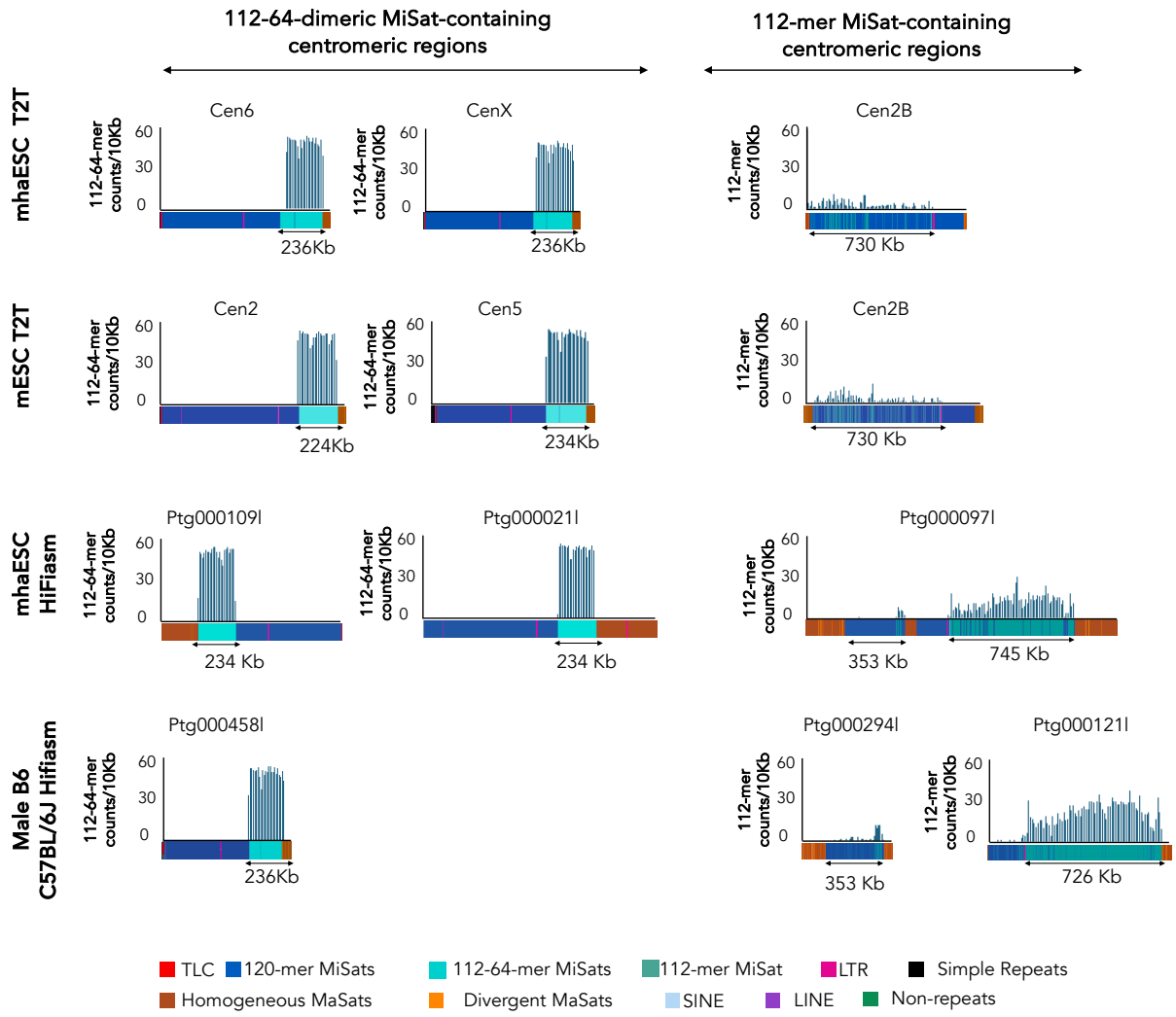
We determined the composition of non-satellite repeat elements at centromeres and pericentromeres. In the mhaESC T2T assembly, five centromeres (Cen2, Cen6, Cen7, Cen8, and CenX) each contained one, while one centromere (Cen16) contained two LTRs (Figures 1B and 4A–4D). In the male C57BL/6J Hifiasm assembly, three centromeric contigs contained LTRs (Supplementary Figures 1 and 10A). Centromeric regions were largely devoid of SINEs and LINEs (Figures 1B, 4A–4D, Supplementary Figures 1A–1C and 10A–10B). No LINE elements were detected within centromeres except at centromere-telomere junctions and interstitial telomeres in the mhaESC T2T assembly (Figures 4A-4B, Supplementary Figures 3 and 10A). Pericentric regions contained more abundant LTRs, SINEs, and LINEs as compared to centromeres (Figure 4B and Supplementary Figure 10A). Divergent MaSats at pericentric junctions proximal to chromosomal arms contained a higher density of LTR retrotransposons, LINEs, SINEs, and simple repeats than centromeric and pericentric contigs (Figure 4D and Supplementary Figure 10D). Furthermore, the density of all non-satellite repeats was highest in non-satellite islands within MaSats (Figure 4D). We also found ~1-2 copies of the tRNA gene tRNA-Lys-AAG present within pericentric non-satellite islands (Figure 4B). Such elements act as barrier elements that prevent the spreading of pericentric heterochromatin to adjacent centromeres in fission yeast [1,2], and they may have a similar function in mouse pericentric regions. Overall, the density of simple repeats, retrotransposons, LINEs, and SINEs increased gradually from centromeric regions to pericentric-chromosomal arm junctions.

Supplementary Figure 1

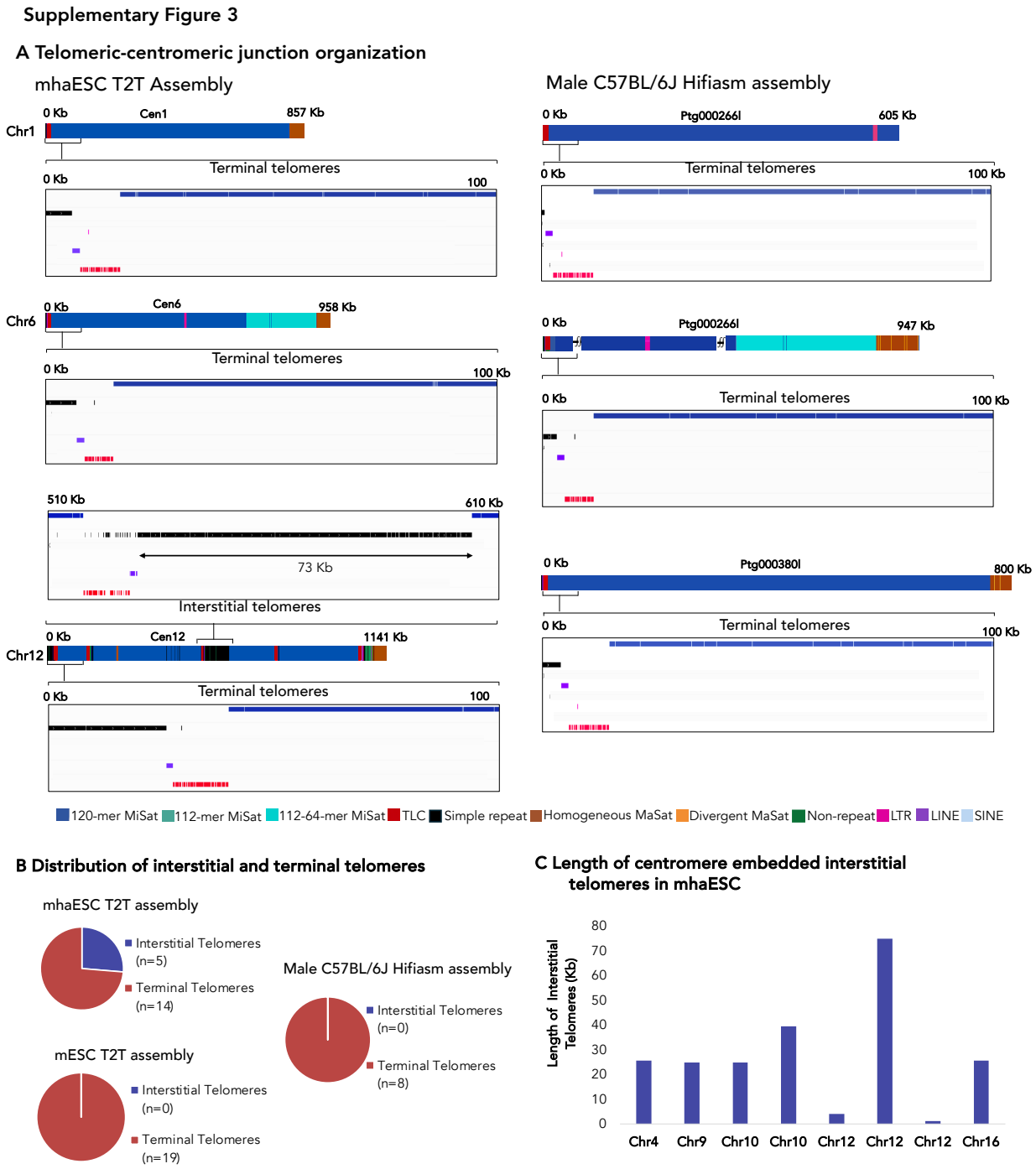


Supplementary Figure 1. Distribution of MiSat-containing contigs in Hifiasm genome assemblies. A) Hifiasm MiSat containing assemblies obtained from PacBio HiFi sequencing data derived from a C57BL/6J male (this study) **B)** Hifiasm MiSat containing assemblies obtained from PacBio HiFi sequencing data derived from a C57BL/6J [3]. **C)** Hifiasm MiSat containing assemblies obtained from PacBio HiFi sequencing data derived from mhaESC [4]. Each horizontal bar represents a contig; red bidirectional arrows mark regions spanning from the pericentric to telomeric ends of the contig. Color keys at the bottom denote specific repeat types.

Supplementary Figure 2



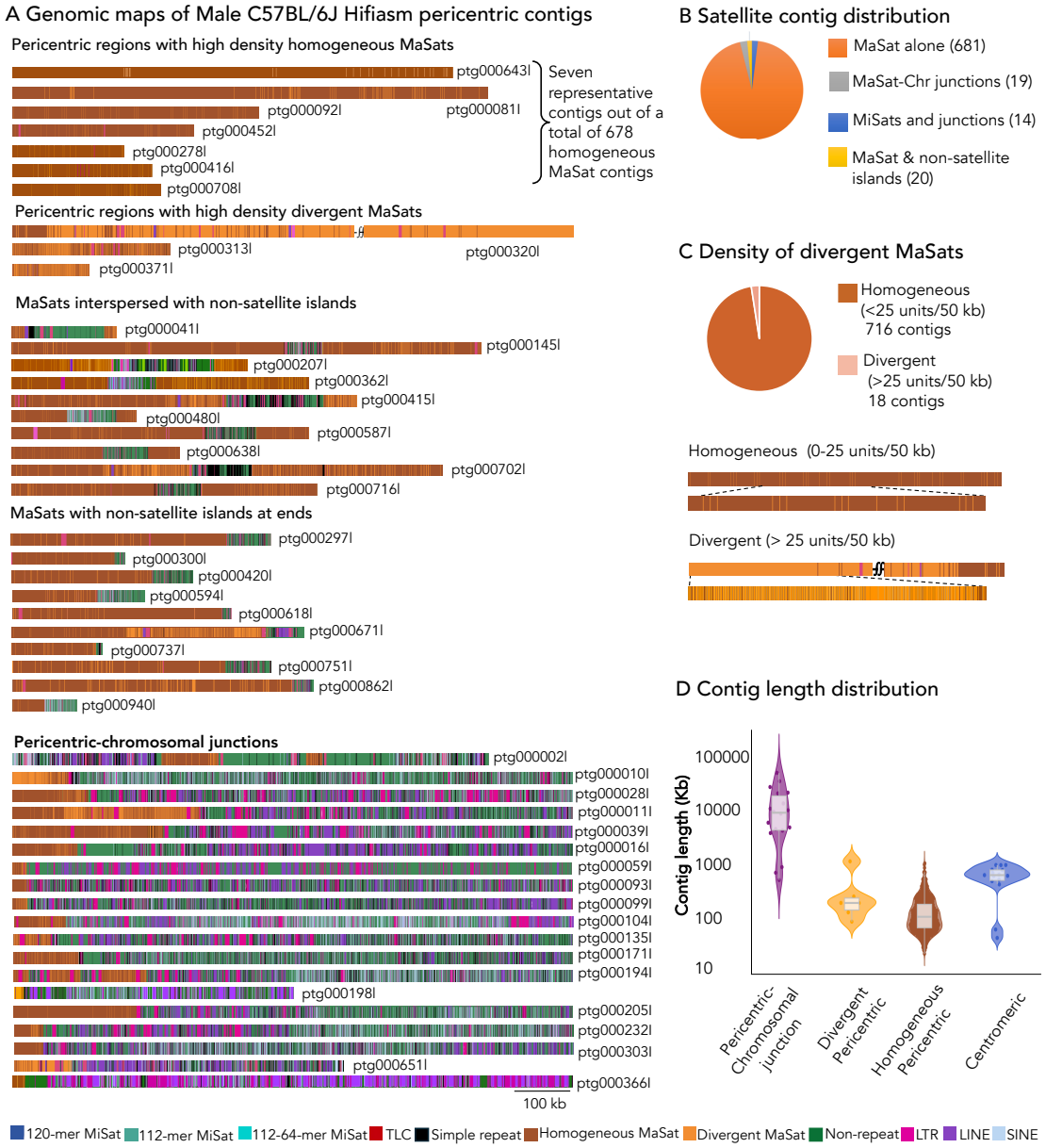
Supplementary Figure 2. Distribution of MiSat length variants across centromeres in T2T and Hifiasm assemblies. Density plots showing the genomic distribution of two types of MiSat monomers: 112-64-dimeric units (left panels) and 112-mer monomers (right panels) across centromeres in the indicated T2T and Hifiasm assemblies. Bar graphs above each annotated centromeric region represent local densities (read counts per 10 Kb bin), aligned to satellite composition tracks below.



Supplementary Figure 3. Telomeric-centromeric junction organization in mouse assemblies. A) Genomic organization of centromeric-telomeric junctions in representative chromosomes from the mhaESC T2T assembly (Left) and contigs from the male C57BL/6J Hifiasm assembly (Right). Tracks show the positions of TLC satellite arrays (red), L1 LINE elements (purple), telomeric simple repeats (black), and MiSats (blue). TLC-Sat arrays of some centromeric-telomeric junctions also contain LTR (magenta pink). Notably, interstitial telomeric repeat blocks (black) are found flanked by

TLC (red) and L1 elements (purple) within centromeric regions in the mhaESC T2T assembly but are absent in the Hifiasm contigs. Color keys at the bottom represent specific repeat types. **B)** Pie charts summarizing the distribution of telomeric arrays detected as either terminal or interstitial telomeres across the mhaESC T2T, mESC T2T, and C57BL/6J Hifiasm assemblies. **C)** Bar plot showing the lengths of interstitial telomeric blocks identified in the mhaESC T2T assembly across different chromosomes.

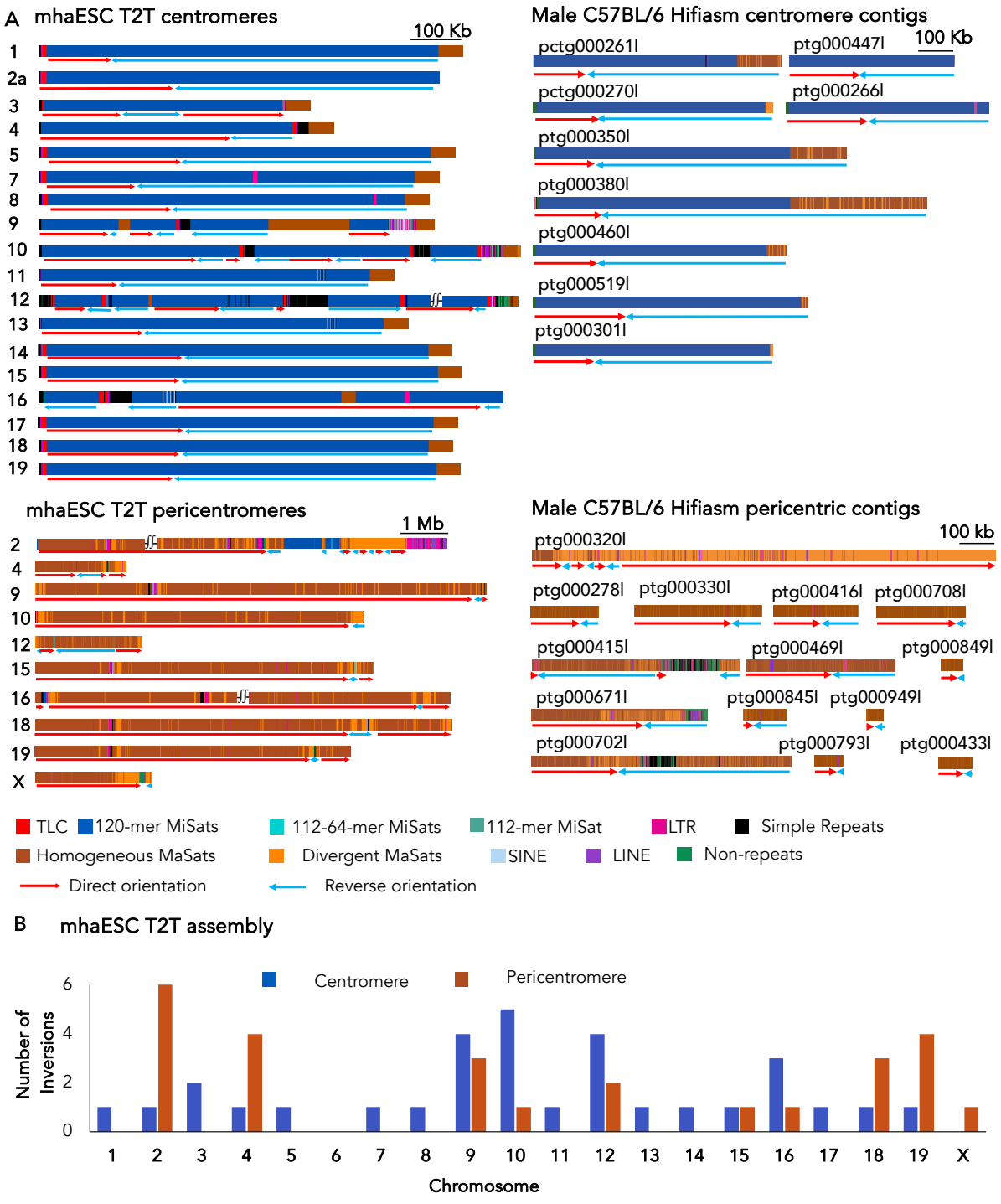
Supplementary Figure 4



Supplementary Figure 4. Pericentric MaSat containing contigs in the male C57BL/6J Hifiasm assembly. **A)** Hifiasm assemblies containing pericentric MaSats. Seven representative contigs out of a total of 678 homogeneous MaSat contigs are shown. For each contig spanning a pericentric–chromosomal arm junction, a 1 Mb region from the pericentric end is shown, except for Ptg000002l, Ptg000198l, and

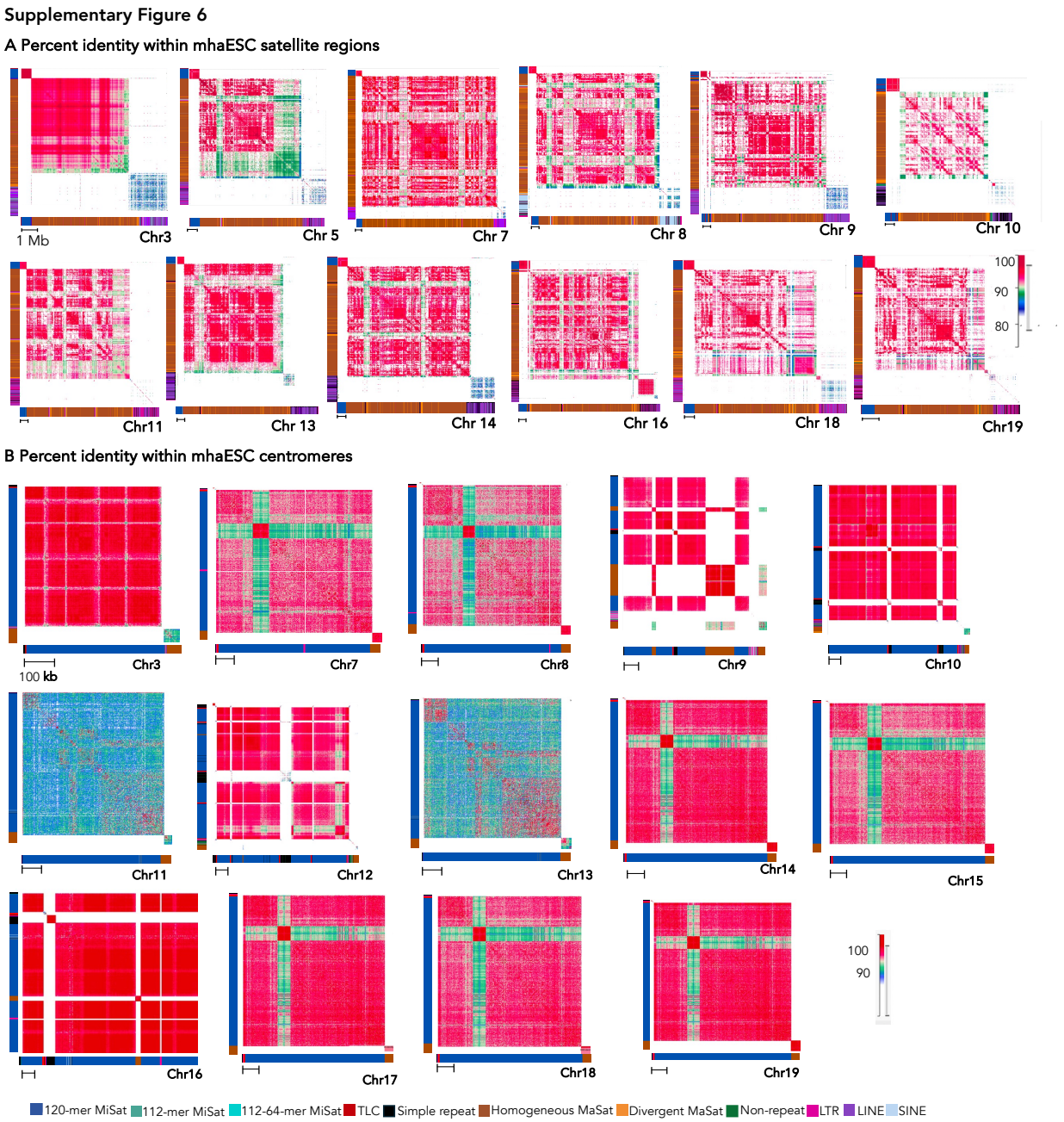
Ptg000651l, where the full contigs are displayed. **B)** Pie chart showing the relative distribution of contigs by MaSat content: homogeneous MaSat, MaSat with divergent sequence, and contigs containing both MaSat and MiSat. **C)** Classification of pericentric contigs based on divergent MaSat density. **D)** Violin plot showing the length distribution of contigs by satellite type category. Color keys at the bottom represent specific repeat types.

Supplementary Figure 5



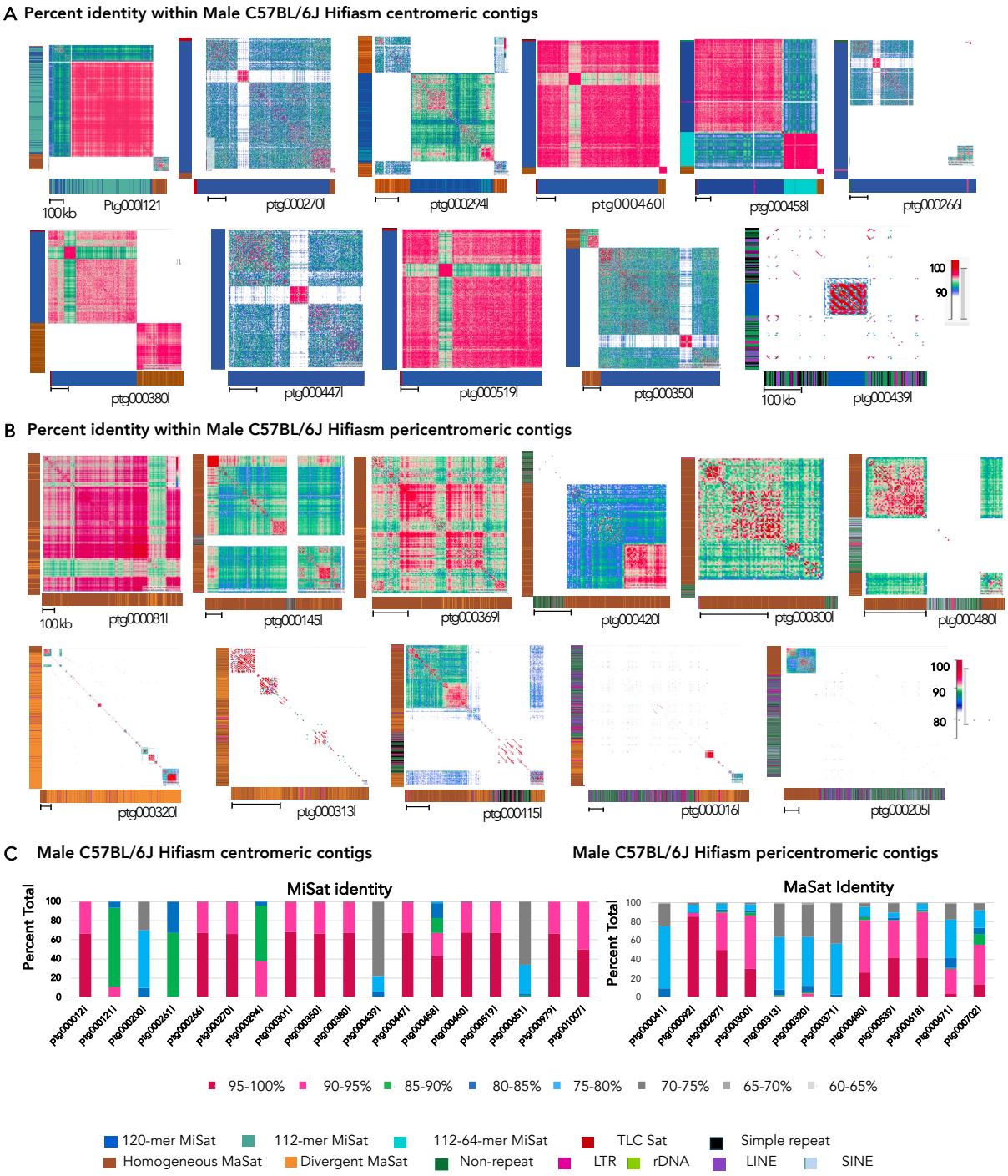
Supplementary Figure 5. Inversion events at centromeric and pericentromeric regions. A) Maps of all inversion events detected in centromeric, centromeric-pericentric junctions, and pericentric contigs in mhaESC T2T (Left) and male C57BL/6J assemblies (Right) are shown. Red and cyan arrows below each contig indicate forward and reverse orientation, respectively; a change in arrow direction denotes an inversion breakpoint. B) A bar graph showing the number of inversions per chromosome in the

mhaESC T2T assembly, separated into centromere (blue) and pericentromere (orange) categories.



Supplementary Figure 6. Sequence identity within and across centromeric and pericentric regions in mhaESC T2T chromosomes. **A)** StainedGlass heatmaps representing the percent identity of self-alignments within the centromeric, pericentric, and pericentric-chromosomal arm junctions of select mhaESC T2T chromosomes. Heatmap colors reflect sequence identity, with the scale bar indicating identity values. **B)** StainedGlass heatmaps illustrating identity across the centromeric regions of select mhaESC T2T chromosomes. Heatmap colors reflect sequence identity, with the scale bar indicating identity values.

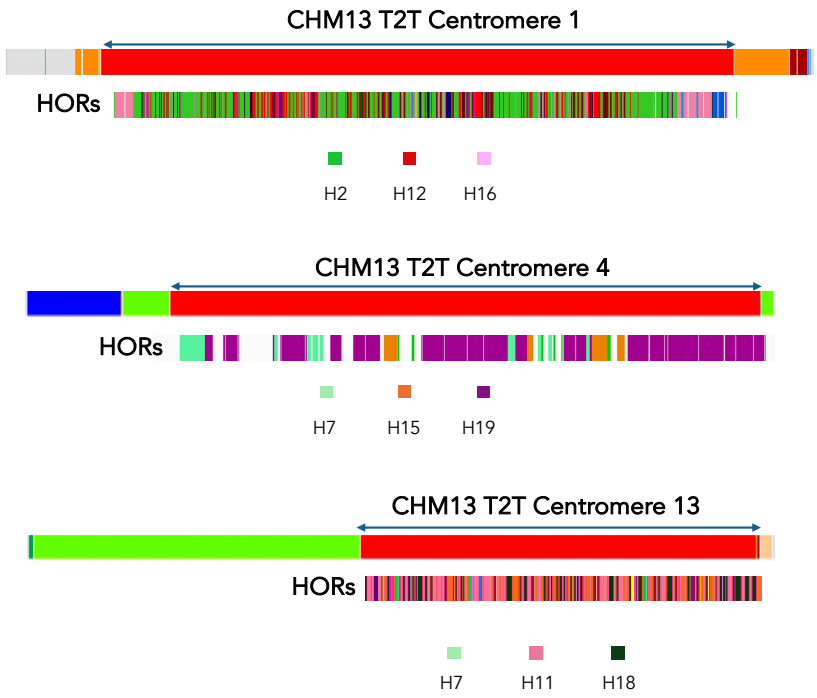
Supplementary Figure 7



Supplementary Figure 7. A) StainedGlass heatmaps representing the percent identity of self-alignments within centromeric contigs of the male C57BL/6J assembly. Heatmap colors reflect sequence identity, with the scale bar indicating identity values. **B)** StainedGlass heatmaps representing the percent identity of self-alignments within representative centromeric, pericentric, and pericentric-chromosomal arm junctions contigs of the male C57BL/6J assembly. Heatmap colors reflect sequence identity, with

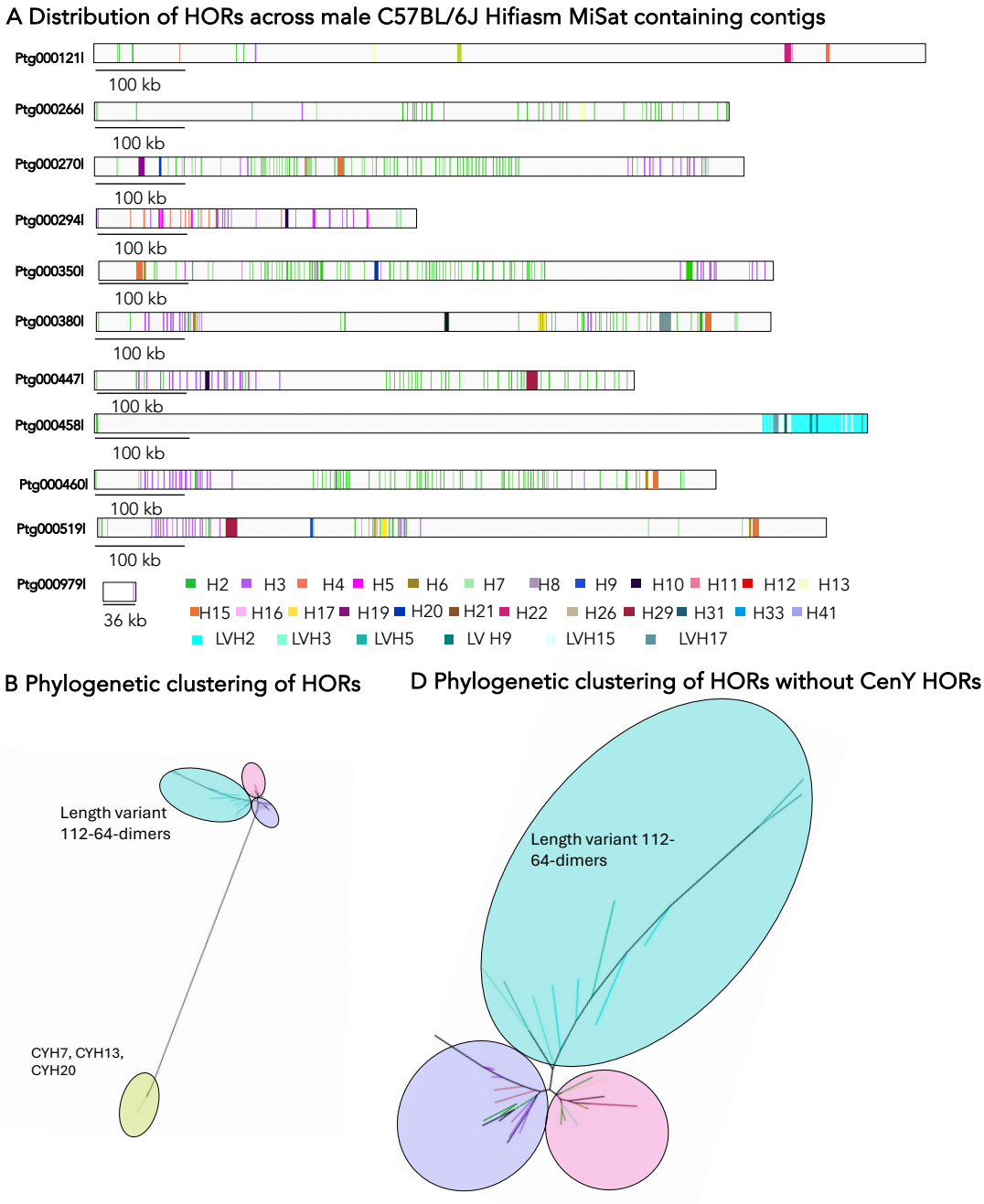
the scale bar indicating identity values. **C)** Percentage of sequence identity among MiSat units (Left) and MaSat units (Right) across the indicated contigs of the male C57BL/6J assembly.

Supplementary Figure 8



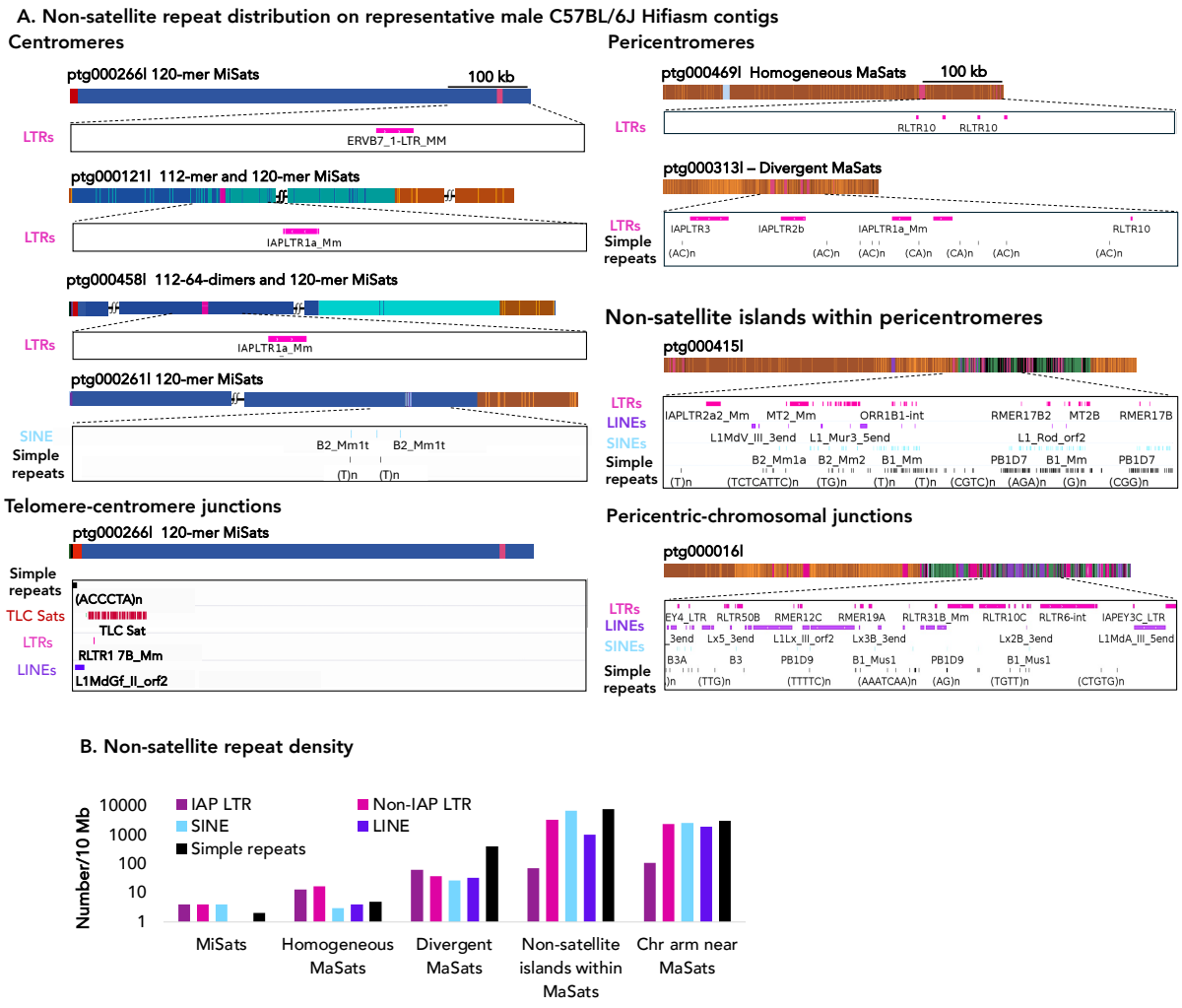
Supplementary Figure 8. Detection of well-characterized human centromeric HORs using CENdetectHOR. Detection of HORs on centromeres of chromosomes 1, 4, and 12 from the CHM13 T2T human genome assembly using CENdetectHOR. Each panel shows the centromeric region (Red) with HOR arrays detected by CENdetectHORs. Below each HOR track, color keys for the top three HORs are given. This figure serves as a validation benchmark, demonstrating that CENdetectHOR accurately detects canonical HORs in well-characterized human centromeres.

Supplementary Figure 9



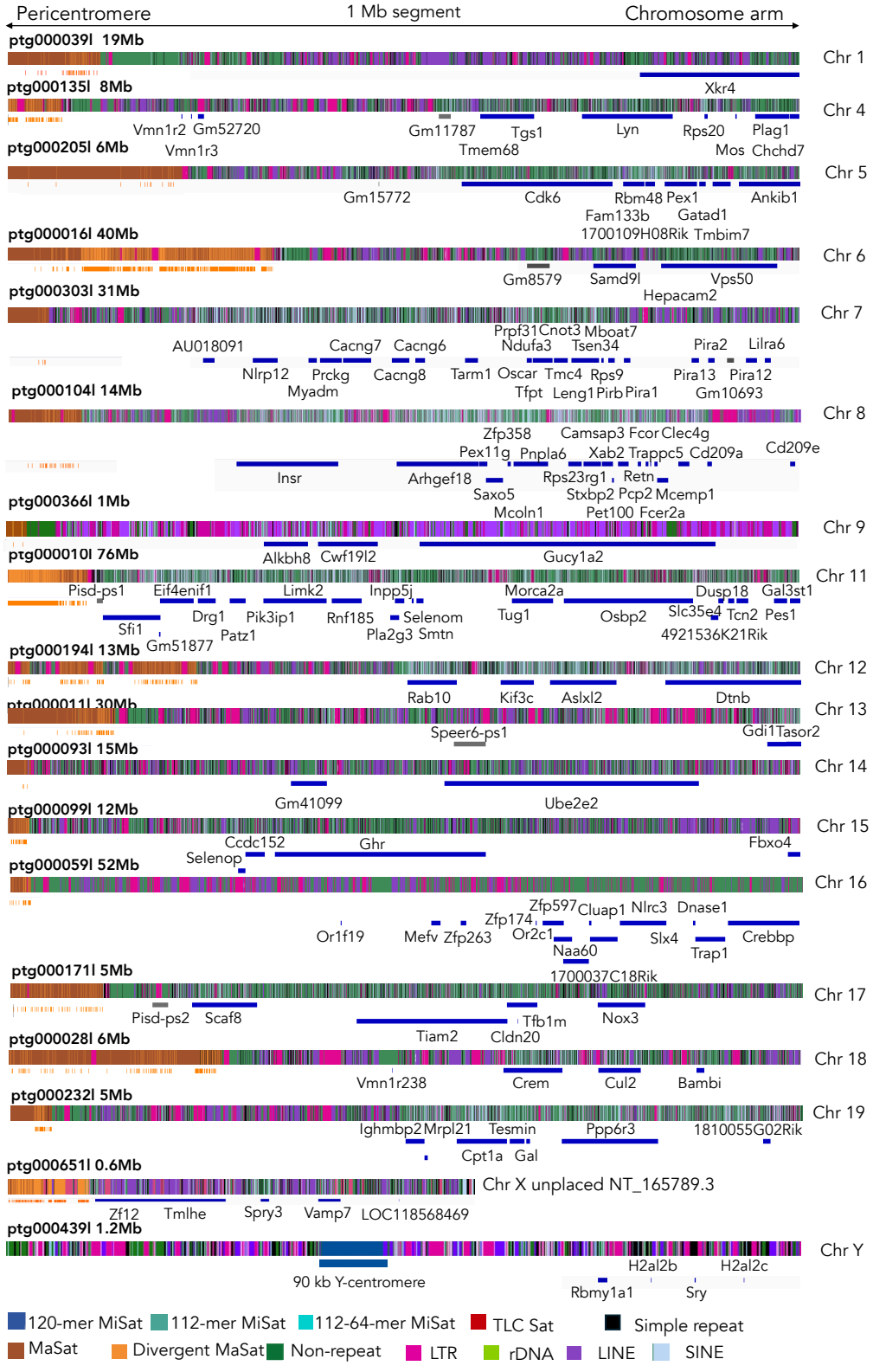
Supplementary Figure 9. Identification and characterization of HORs in mouse centromeres. A) Distribution of HORs detected by CENdetectHOR within MiSat-containing contigs from the male C57BL/6J Hifiasm assembly. Each horizontal bar represents a contig, with colored tick marks denoting different HOR variants. The Y-centromere HOR distribution is shown separately in Figure 3. **B)** Phylogenetic clustering of HORs detected across all MiSat contigs. Distinct clusters indicate divergence among HOR families, with a unique cluster formed by HORs from CYH7, CYH13, and CYH20. **C)** Phylogenetic relationships among HORs excluding the CenY.

Supplementary Figure 10



Supplementary Figure 10. Non-satellite repeat composition at centromeric and pericentric regions in male C57BL/6J assembly. A) Representative contigs showing the distribution of LTRs, LINEs, SINEs, and simple repeats on centromeric, pericentric, and junction regions. Contigs are drawn to scale. A zoomed-in view across a 100 Kb region is shown below each contig. **B)** The density of retrotransposons (IAP and non-IAP LTRs), SINEs, LINEs, and simple repeats across satellite regions.

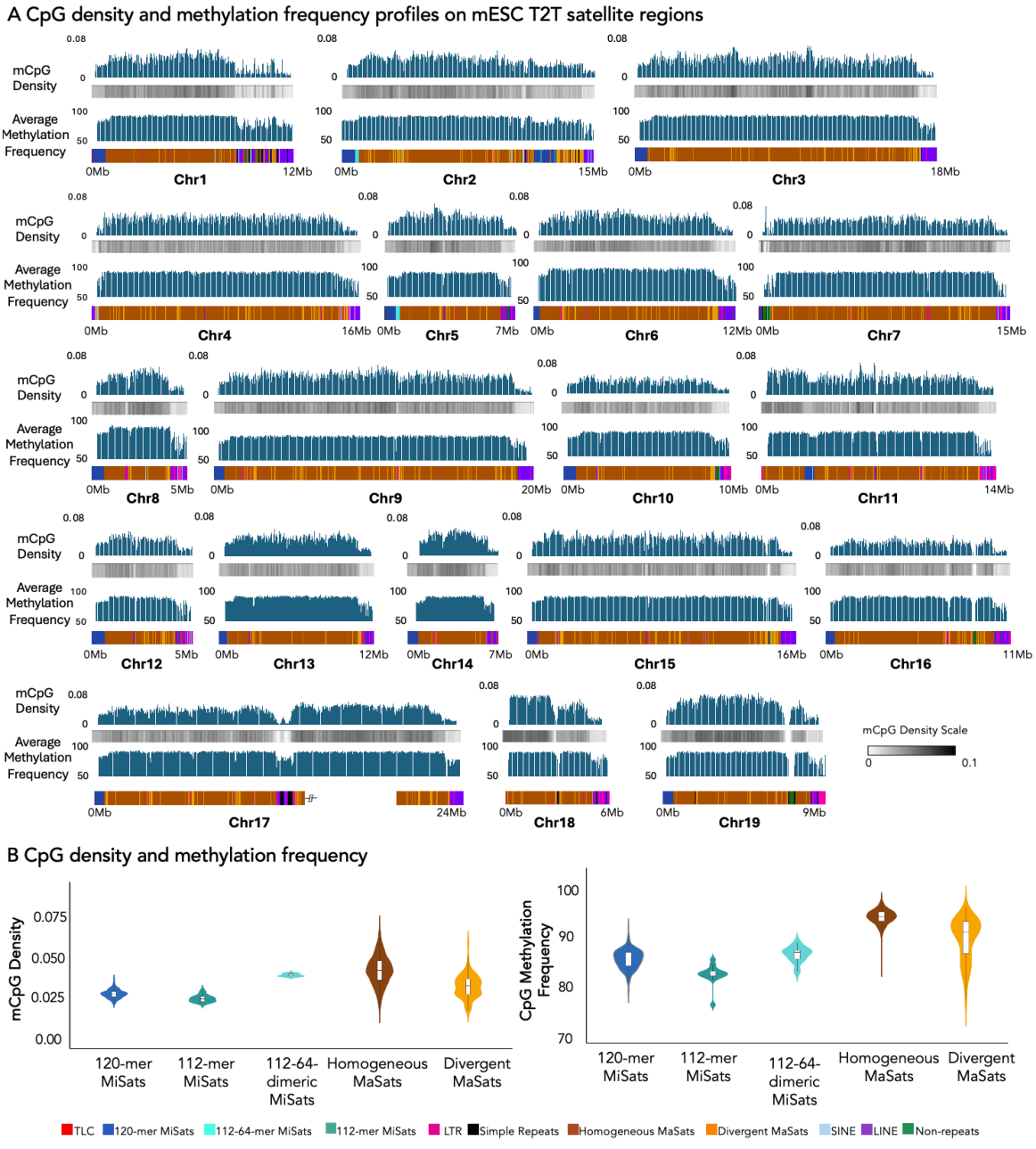
Supplementary Figure 11



Supplementary Figure 11. Assembly maps of male C57BL/6J Hifiasm contigs spanning pericentric-chromosomal arm junction. For each contig containing a

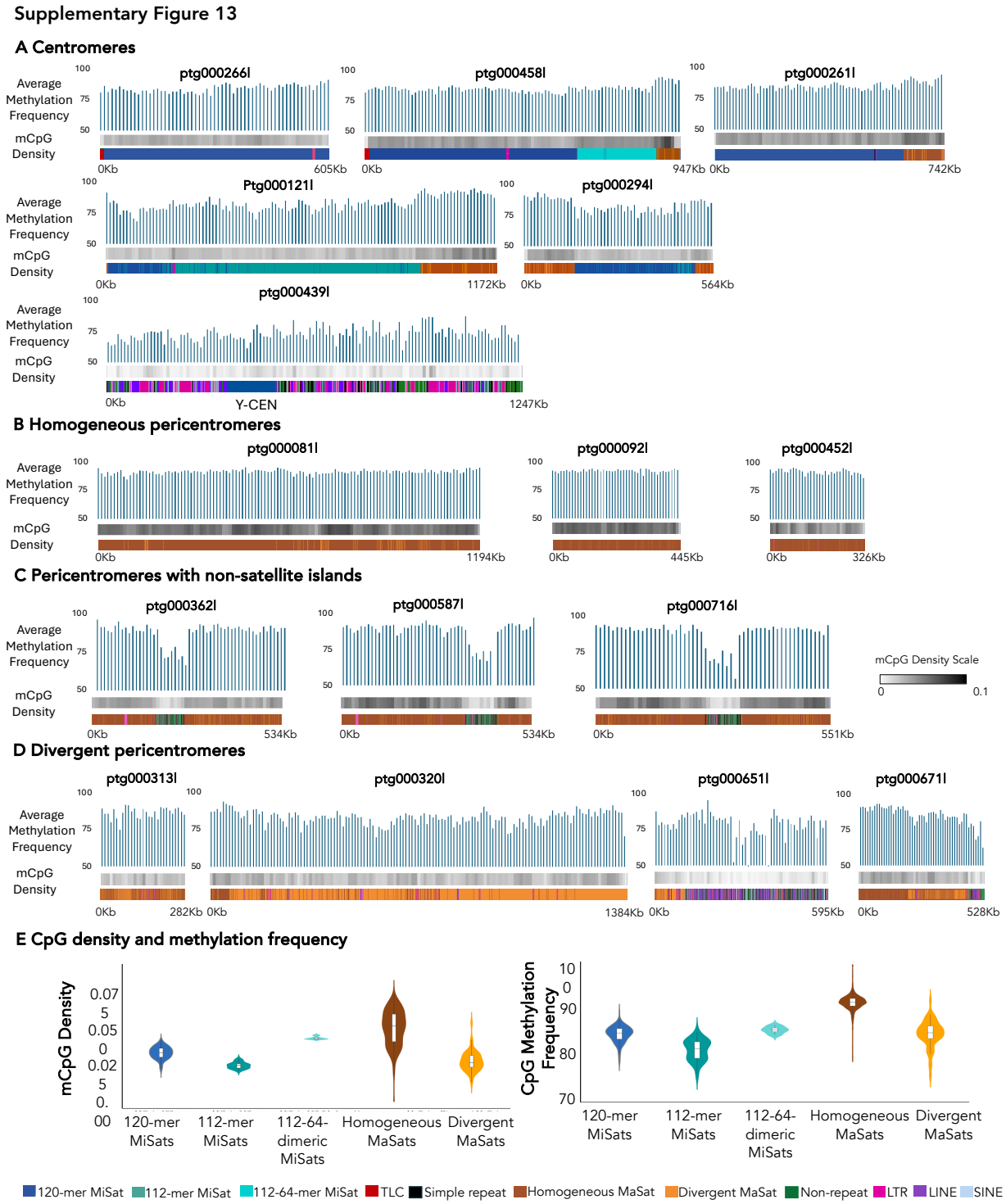
pericentric-chromosomal junction, a 1 Mb region encompassing the pericentric end is displayed, showing both the chromosome-specific genes that map to the GRCm39 genome assembly and the junction contig itself. Below each contig map, the divergent MaSat units are marked in orange, annotated genes are represented in blue, and pseudogenes are shown in grey. All chromosome-specific genes and pseudogenes we identified in these junctions are listed.

Supplementary Figure 12



Supplementary Figure 12. mCpG density and methylation frequency profiles on mESC T2T satellite regions. A) For each chromosome, mCpG site density per 50 Kb bin (Top) and DNA methylation frequency per 50 Kb bin (Bottom). **B)** Violin plots

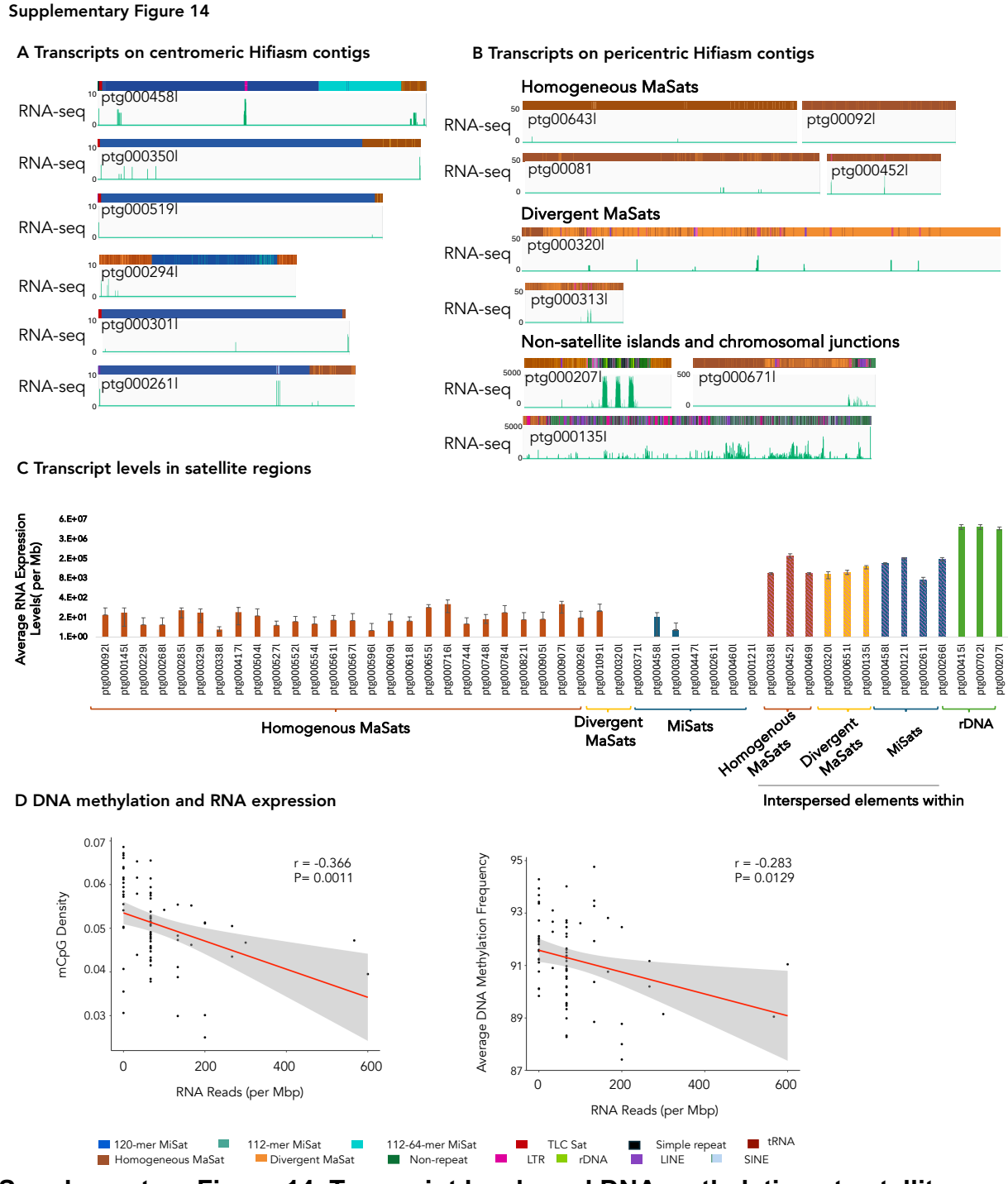
comparing overall mCpG density (Left) and CpG methylation frequency (Right) across different satellite repeat types.



Supplementary Figure 13. DNA methylation on centromeric and pericentric regions in the male C57BL/6J Hifiasm assembly. CpG methylation frequency and mCpG density across 10 Kb bins on **A)** centromeric, **B)** centromeric-pericentric junction,

C) homogeneous pericentric, and **D)** Divergent pericentric contigs are shown **E)** The mCpG density and DNA methylation frequency on individual contigs from divergent MaSat, homogeneous MaSat, and MiSat categories.

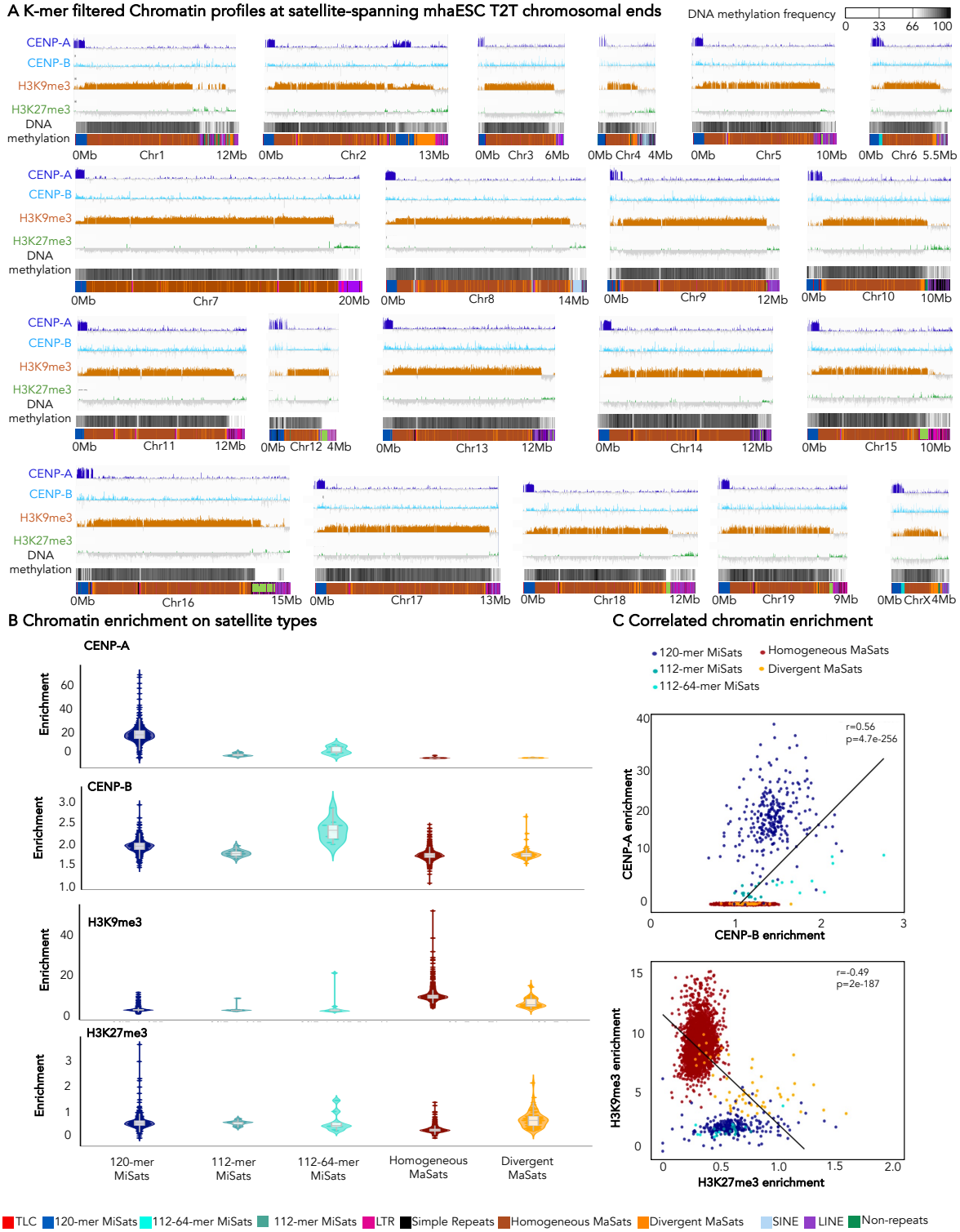
Supplementary Figure 14



Supplementary Figure 14. Transcript levels and DNA methylation at satellites and pericentric-chromosomal arm junctions in the male C57BL/6J Hifiasm assembly. **A)** RNA-seq signal profiles mapped to representative MiSat containing centromeric male C57BL/6J Hifiasm contigs. **B)** RNA-seq signal profiles mapped to representative

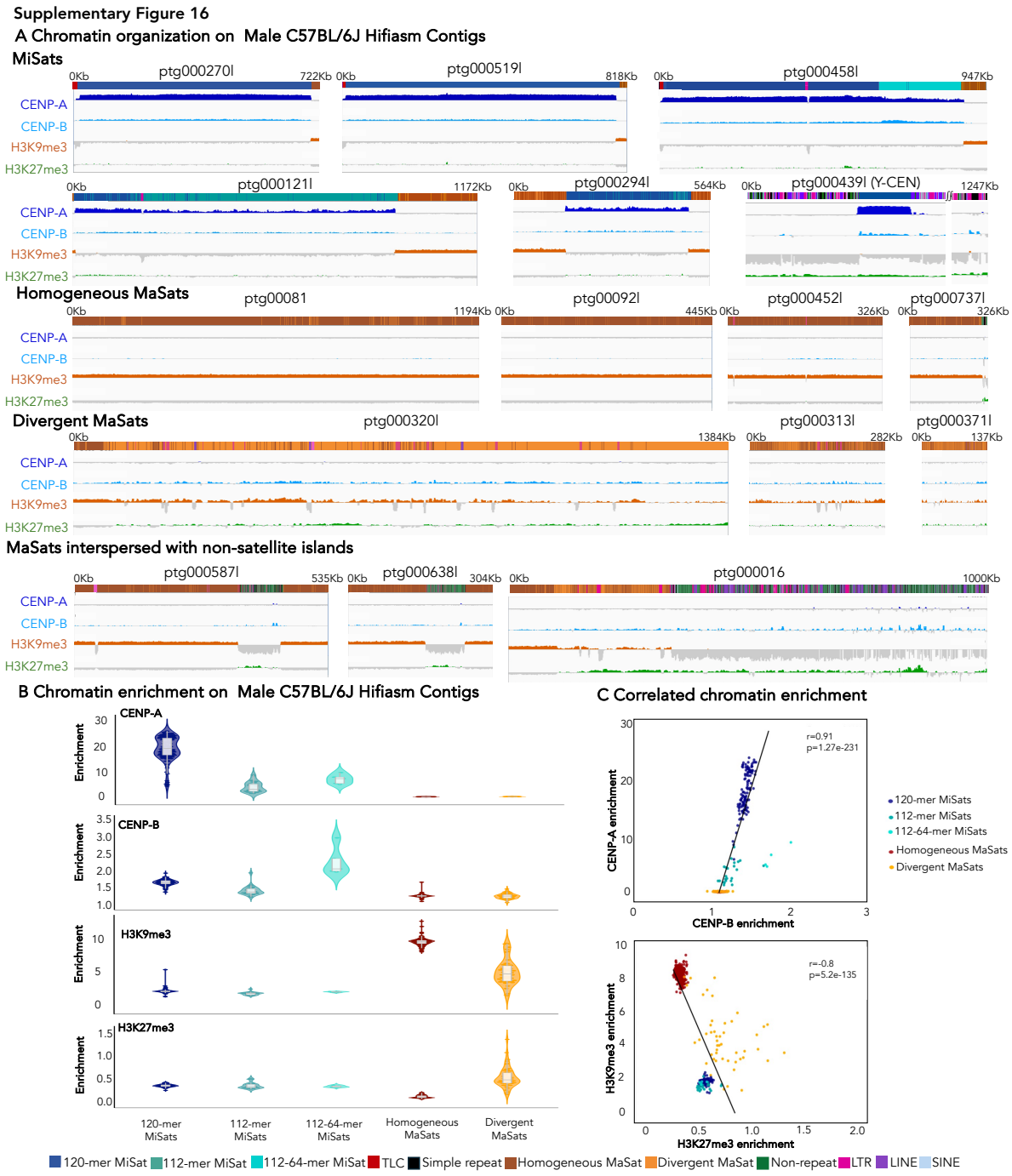
MaSat containing pericentric and pericentric-chromosomal arm junctions male C57BL/6J Hifiasm contigs. **C)** Quantification of average RNA-seq signal per contig, grouped by repeat context. Bars represent mean ± s.d. from $n = 3$ independent biological replicates. Each replicate corresponds to RNA-seq data from the liver tissue of an individual male C57BL/6J mouse. RNA was extracted, processed, and sequenced separately for each replicate without pooling. All replicates are biological; no technical replicates were included. **D)** Correlation between average RNA expression and mCpG density (Left) and DNA methylation frequency (Right Pearson correlation coefficients (r) were calculated using a two-sided test. Exact p-values and r values are shown on the plots. No multiple-comparison corrections were applied, as only pairwise correlations were tested.

Supplementary Figure 15



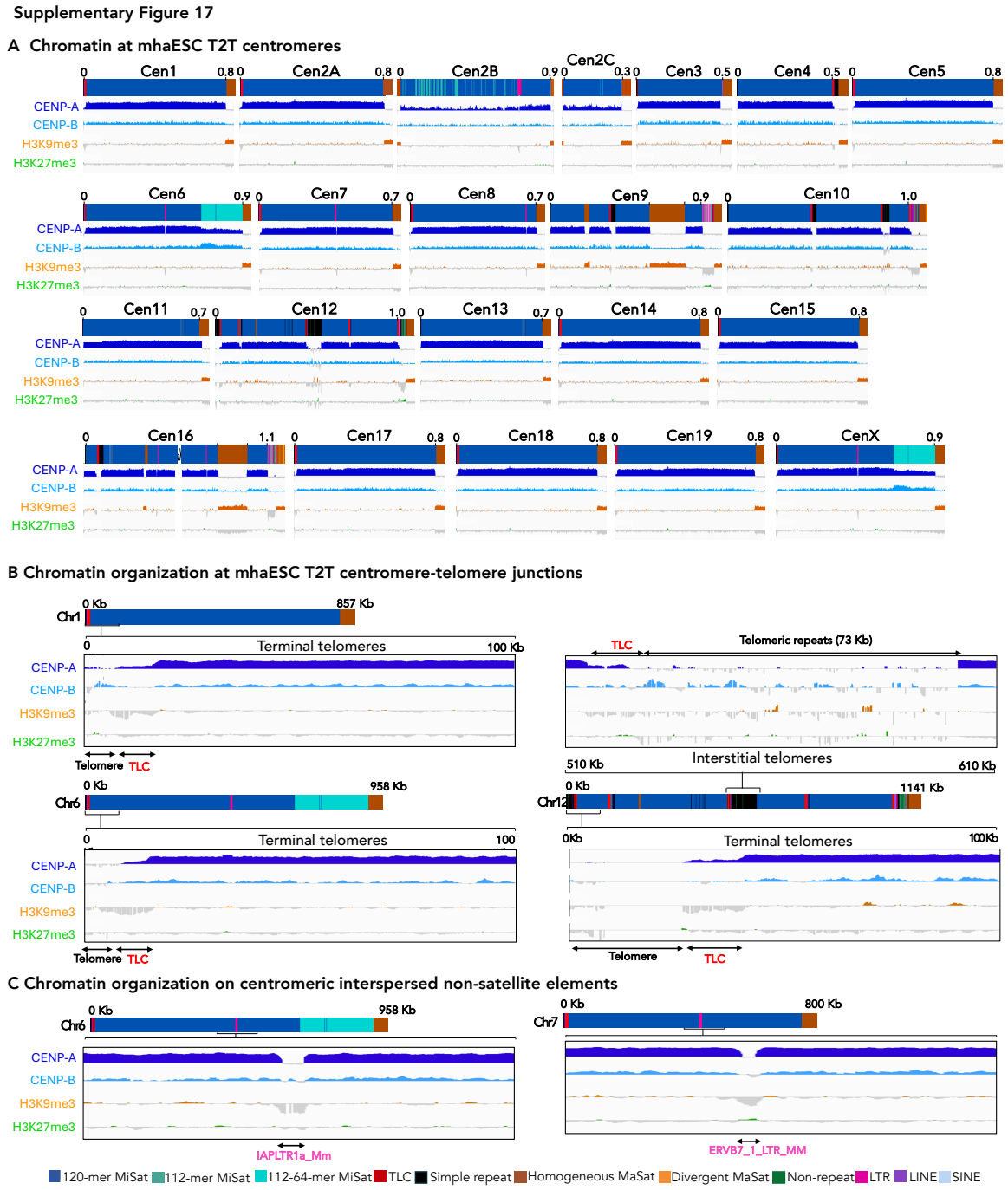
Supplementary Figure 15. Chromatin enrichment on uniquely mappable mhaESC T2T satellite regions using a k–mer filtering approach. A) CUT&RUN-Seq enrichment profiles for CENP-A, CENP-B, H3K9me3, and H3K27me3 across centromeric and pericentric satellite regions of the mhaESC T2T assembly. Read

alignments were performed using Bowtie2 following k-mer filtering to reduce mapping ambiguity. The Y-axis shows the $\log_2$ enrichment ratio relative to IgG controls, with fixed axis limits across a given chromatin mark: -4 to +10 for CENP-A, -3 to +2 for CENP-B, and -6 to +4 for H3K9me3 and H3K27me3. **B)** Violin plots showing normalized enrichment scores (signals relative to IgG) across 120-mer MiSats, 112-mer MiSats, 112-64-dimeric MiSats, homogeneous MaSats, and divergent MaSats.**C)** Scatter plots showing correlations between CENP-A and CENP-B enrichment (Top) and between H3K9me3 and H3K27me3 (Bottom) across annotated satellite types. Pearson correlation coefficients (r) and corresponding p-values are shown.



Supplementary Figure 16. Chromatin enrichment on satellite-containing regions in male C57BL/6J mice. A) Log_2 enrichment profiles of CENP-A, CENP-B, H3K9me3, and H3K27me3 relative to IgG control across representative centromeric and pericentric regions in the Hifiasm assembly of male C57BL/6J mice. The enrichment values are plotted with a fixed log_2 Y-axis range for a given chromatin mark: -4 to +10 for CENP-A, -3 to +2 for CENP-B, and -6 to +4 for both H3K9me3 and H3K27me3. **B)** Violin plots displaying normalized enrichment scores for each chromatin factor across major satellite classes: 120-mer MiSats, 112-mer MiSats, 112-64-dimeric MiSats, homogeneous MaSats, and divergent MaSats. **C)** Scatter plots showing correlations

between CENP-A and CENP-B (Top), and H3K9me3 and H3K27me3 (Bottom) enrichment signals across satellite categories. Pearson correlation coefficients (r) and p-values are indicated.

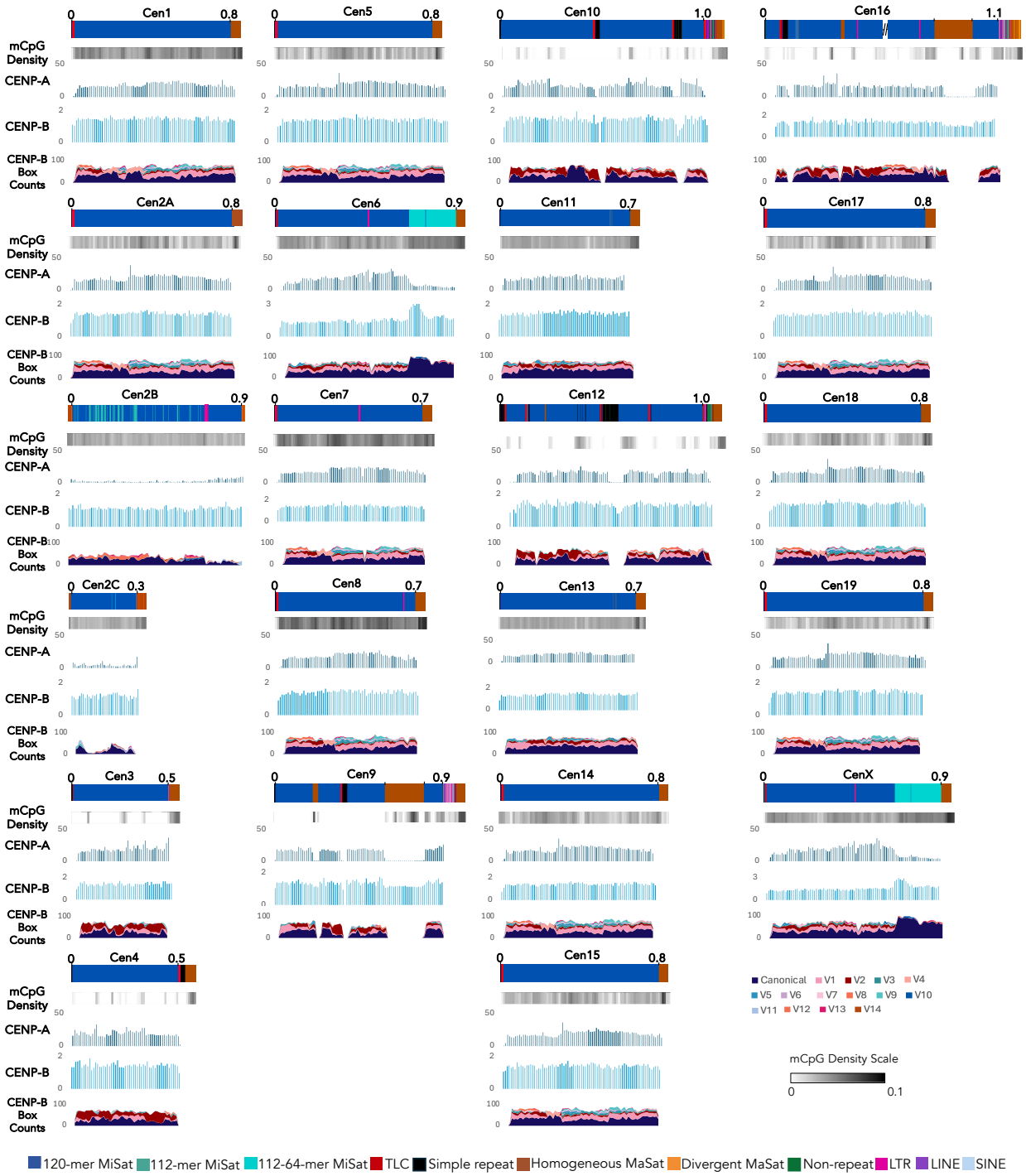


Supplementary Figure 17. Chromatin profiles on mESC T2T centromeres, centromere–telomere junctions, and interspersed non-satellite repeats. A) CUT&RUN-seq enrichment profiles (log_2 signal over IgG control) for CENP-A, CENP-B, H3K9me3, and H3K27me3 across satellite-rich centromeric regions of the mESC T2T genome assembly. Numbers above each centromere indicate the start and end

coordinates (in Mb) of the annotated centromeric regions. Chromatin profiles across **B)** centromere–telomere junctions. **C)** representative contigs containing interspersed non-satellite repeats located within centromeric regions, revealing chromatin signatures associated with these atypical centromeric segments. Each track shows enrichment with fixed Y-axis scaling for a given chromatin mark: -4 to +10 for CENP-A, -3 to +2 for CENP-B, and -6 to +4 for H3K9me3 and H3K27me3.

Supplementary Figure 18

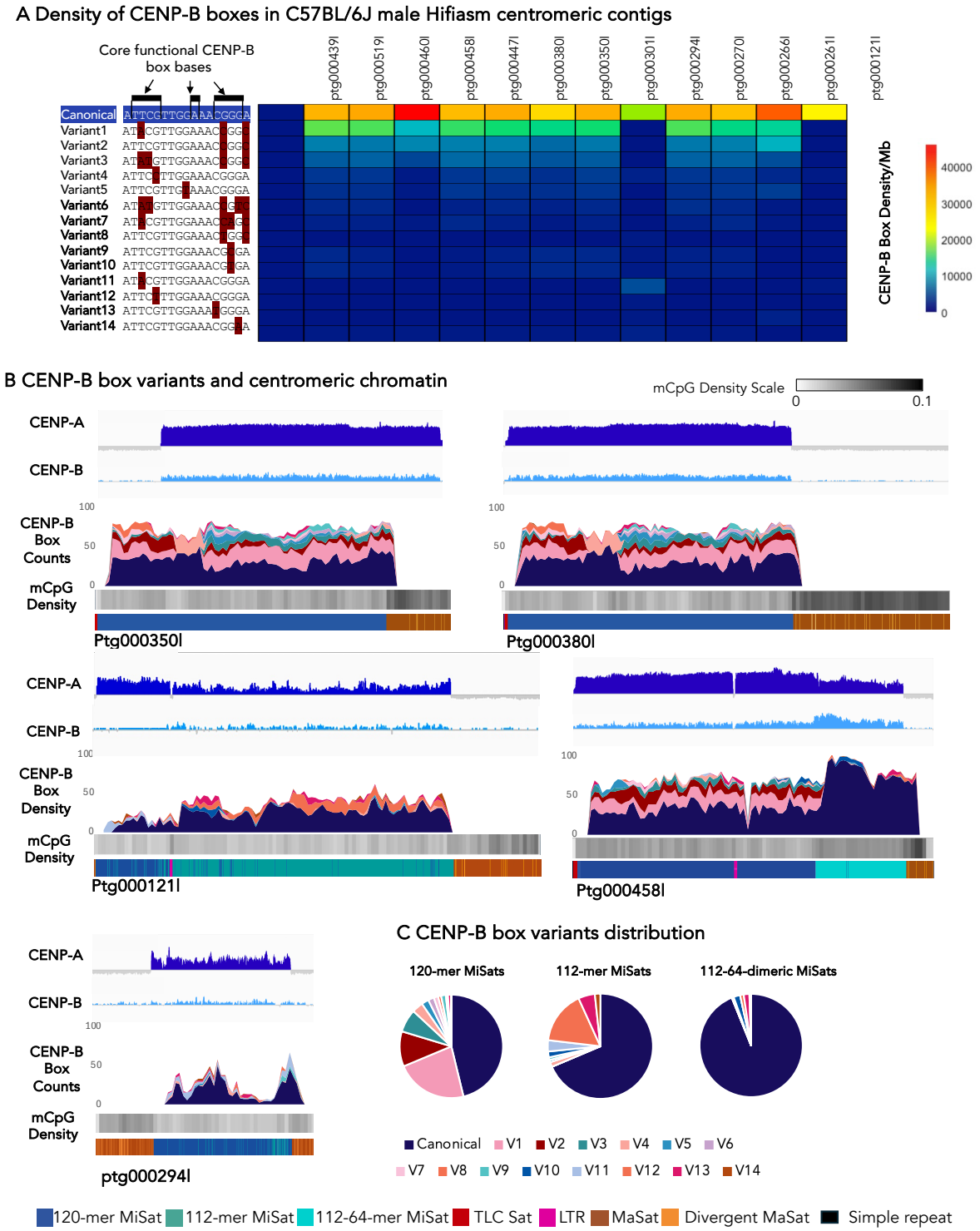
CENP-B box variants, Chromatin enrichment and DNA methylation on mhaESC T2T centromeres



Supplementary Figure 18. CENP-B box composition, chromatin enrichment and DNA methylation across all centromeres in the mhaESC T2T assembly. From top to bottom, the tracks include satellite repeat annotations, mCpG density, enrichment of CENP-A and CENP-B relative to the IgG control, and stacked bar plots depicting the relative abundance of canonical and variant CENP-B box motifs (V1–V14). Notably,

PacBio HiFi long CCS reads did not align well to the mhaESC T2T centromeres containing interstitial telomeres (Centromere 4, 9, 10, 12, and 16). Conversely, PacBio HiFi CCS reads aligned well with all mESC T2T centromeres (Supplementary Figure 12) and C57BL/6J Hifiasm contigs (Supplementary Figure 13). This indicates that the assembly of interstitial telomere-containing mhaESC T2T centromeres may need additional validation or refinement.

Supplementary Figure 19



Supplementary Figure 19. Sequence motifs and chromatin enrichment at centromeric and pericentric regions. A) Heatmap showing the density of canonical and variant CENP-B box motifs across MiSat containing C57BL/6J Hifiasm contigs. The consensus canonical CENP-B box sequence is shown at the top, and single-nucleotide deviations observed in variant CENP-B boxes are indicated in maroon. The color scale

indicating motif density per Mb is shown below the heatmap. **B)** Log_2 CENP-A and CENP-B enrichment profiles (Top), stacked bar plots of CENP-B box variant composition (middle) and DNA methylation profiles (Bottom) on contigs containing different classes of MiSats (120-mers, 112-mers, and 112-64-dimers). The chromatin enrichment values are plotted with a fixed Y-axis range for a given chromatin mark: -2 to +7 for CENP-A and -2 to +2 for CENP-B. **C)** Pie charts showing the distribution of CENP-B box variants on contigs containing 120-mer, 112-64-dimeric, and 112-mer MiSats.

References

1 Kuhn, R. M., Clarke, L. & Carbon, J. Clustered tRNA genes in Schizosaccharomyces pombe centromeric DNA sequence repeats. *Proc Natl Acad Sci U S A* **88**, 1306-1310 (1991).

2 Scott, K. C., Merrett, S. L. & Willard, H. F. A heterochromatin barrier partitions the fission yeast centromere into discrete chromatin domains. *Curr Biol* **16**, 119-129 (2006). https://doi.org:10.1016/j.cub.2005.11.065

3 Hon, T. *et al.* Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci Data* **7**, 399 (2020). https://doi.org:10.1038/s41597-020-00743-4

4 Liu, J. *et al.* The complete telomere-to-telomere sequence of a mouse genome. *Science* **386**, 1141-1146 (2024). https://doi.org:10.1126/science.adq8191