

# Supplementary Information

Architectural and evolutionary features of TE-derived TSSs shape tissue-specific promoter activity in the human genome

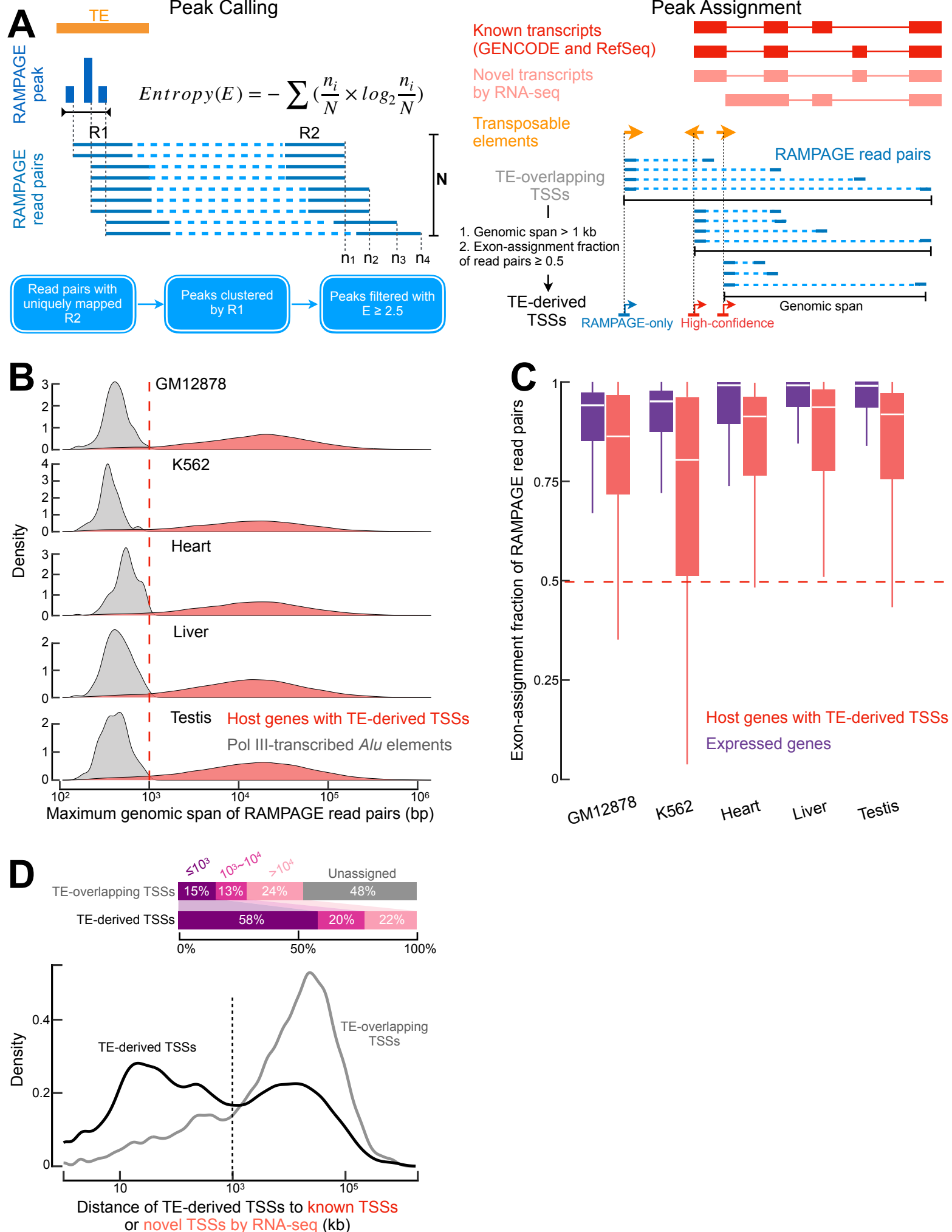
Ya Zhang<sup>1,#</sup>, Yahan Fan<sup>2</sup>, Huang Wu<sup>2</sup>, Xiao-Ou Zhang<sup>1,#</sup>

<sup>1</sup>Shanghai Key Laboratory of Maternal and Fetal Medicine, Clinical and Translational Research Center of Shanghai First Maternity and Infant Hospital, Frontier Science Center for Stem Cell Research, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China

<sup>2</sup>Department of Transfusion Medicine, Daping Hospital, State Key Laboratory of Trauma and Chemical Poisoning, Army Medical University, Chongqing 400042, China

#Corresponding authors: zhangaya1225@tongji.edu.cn, zhangxiaou@tongji.edu.cn

# Supplementary Figure 1



**Supplementary Figure 1.** Computational pipeline for identifying TE-derived TSSs genome-wide using RAMPAGE data

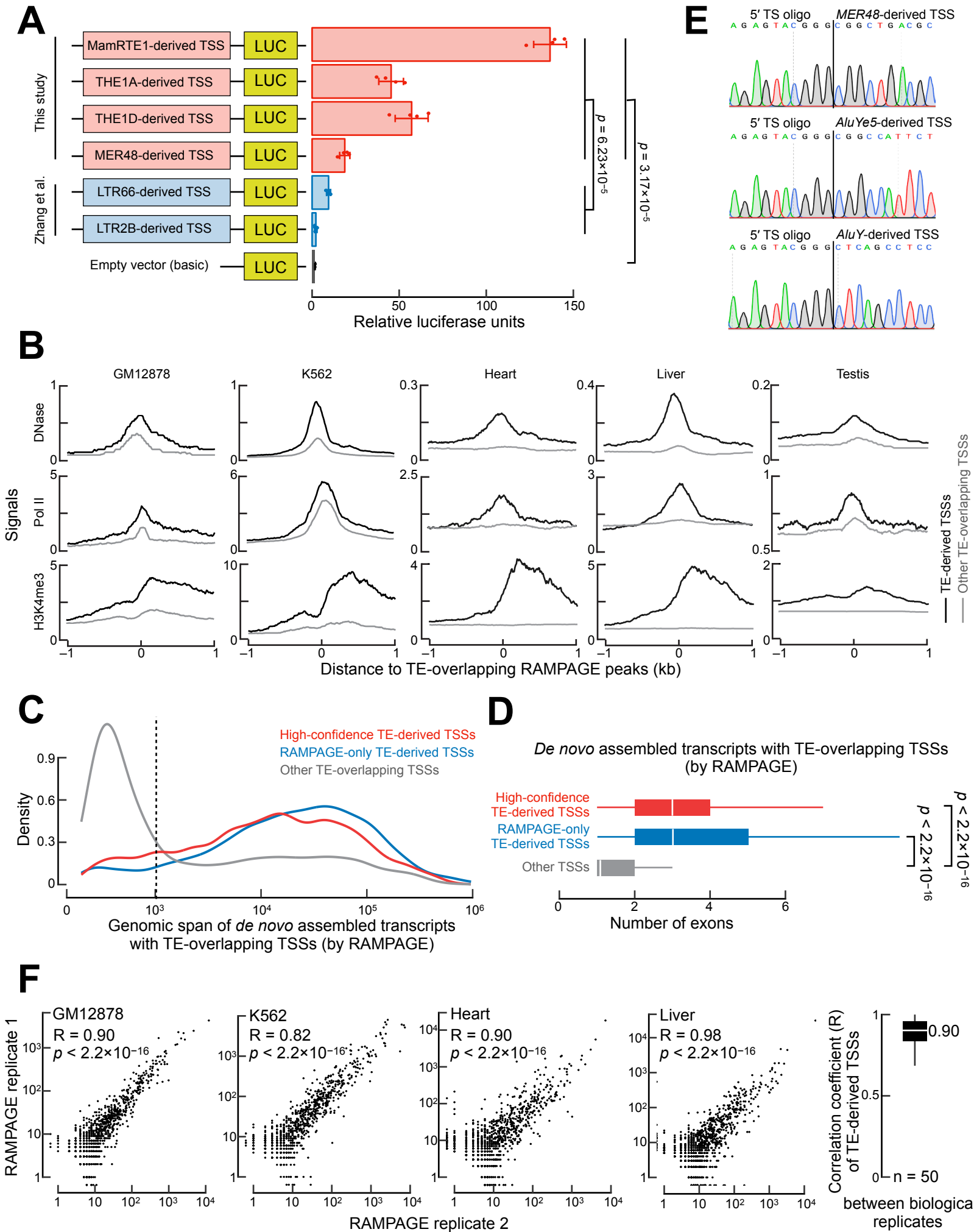
**A.** A computational pipeline for systematically identifying TE-derived TSSs. In the peak calling step (left panel), paired-end RAMPAGE reads with uniquely mapping R2 reads were first clustered to call peaks at the 5'-end of R1, and the resulting peaks (blue bars) were further filtered with entropy (E) of reads that overlapped TEs (orange bar). In the peak assignment step (right panel), TE-overlapping RAMPAGE peaks with 1 kb or longer genomic spans and with more than half of their read pairs linked to a downstream known exon or a novel exon assembled by the RNA-seq data in the corresponding biosample were defined as TE-derived TSSs. A subset of TE-derived TSSs was deemed of high confidence—they were within 1 kb of known TSSs (annotated as TSSs by GENCODE or RefSeq) or novel TSSs (transcripts assembled using RNA-seq data in the matched biosample). See panel D for more details on the 1 kb cutoff.

**B.** Length distributions of the genomic spans of RAMPAGE read pairs for host genes of TE-derived TSSs (red) and Pol III-transcribed Alu elements (grey) in GM12878 and K562 cells, as well as in heart, liver, and testis tissues.

**C.** Distributions of the fractions of RAMPAGE read pairs assigned to expressed genes (purple) and host genes of TE-derived TSSs (red) in GM12878 and K562 cells, as well as in heart, liver, and testis tissues. In each boxplot, the center line indicates the median, the box represents the interquartile range, and the whiskers extend to the most extreme values within 1.5× the interquartile range.

**D.** Distance distribution of TE-derived TSSs to known TSSs (from GENCODE and RefSeq) or novel TSSs (from transcripts assembled using matched RNA-seq data). If the distance is less than 1 kb, this TE-derived TSS was defined as high-confidence. The top inset shows the comparison of distance distributions between TE-derived TSSs (bottom) and TE-overlapping TSSs (top).

# Supplementary Figure 2



## **Supplementary Figure 2.** Computational and experimental validation of TE-derived TSSs

**A.** Luciferase reporter assays measuring promoter activity of TE-derived TSS sequences. Data are shown as mean  $\pm$  SEM from four independent experiments. Relative luciferase activity was normalized to Renilla luciferase. Statistical significance was determined using the one-sided Student's t-test.

**B.** DNase signal (left) and ChIP-seq signals for POLR2A (middle) and H3K4me3 (right) are shown in the  $\pm$  1 kb window centered on the summit of RAMPAGE peaks of TE-derived TSSs (black lines) and other TE-overlapping TSSs (grey lines) in GM12878 and K562 cells, as well as in heart, liver, and testis tissues. The signals were averaged over all TSSs in each set.

**C.** Length distributions of the genomic spans of de novo assembled transcripts using RAMPAGE read pairs for high-confidence TE-derived TSSs (red), RAMPAGE-only TE-derived TSSs (blue), and other TE-overlapping TSSs (grey).

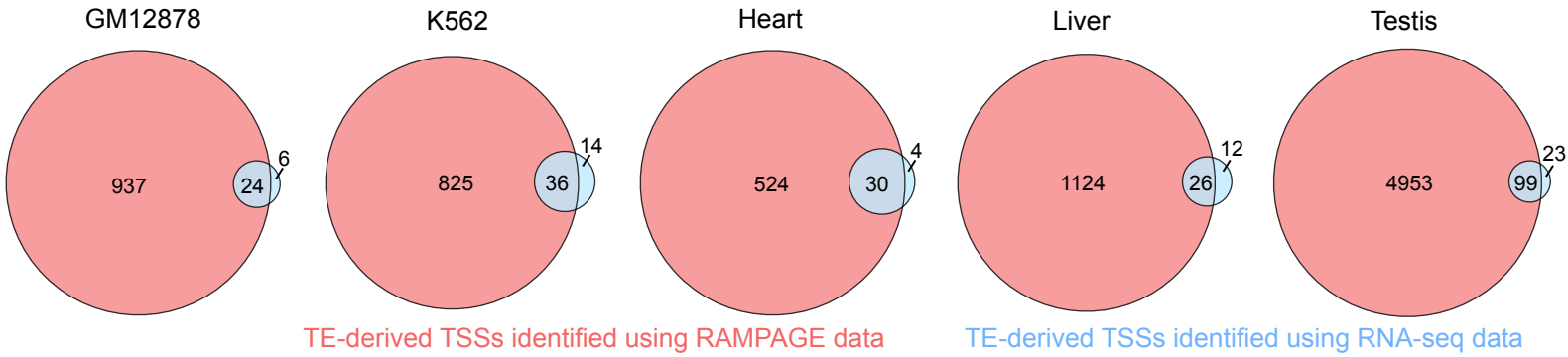
**D.** Exon counts of de novo assembled transcripts using RAMPAGE read pairs for high-confidence TE-derived TSSs (red), RAMPAGE-only TE-derived TSSs (blue), and other TE-overlapping TSSs (grey). In each boxplot, the center line indicates the median, the box represents the interquartile range, and the whiskers extend to the most extreme values within 1.5 $\times$  the interquartile range. One-sided Wilcoxon rank-sum test P-values are shown.

**E.** The 5' RACE sequencing confirmed transcription initiation at three previously unannotated TE-derived TSSs in K562 cells.

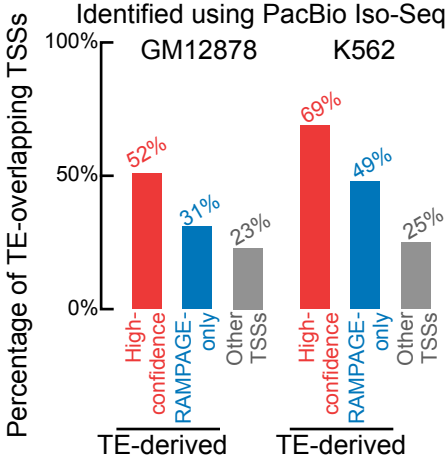
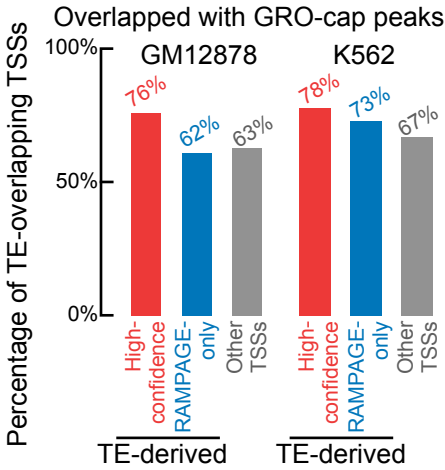
**F.** Reproducibility of TE-derived TSS quantification between RAMPAGE biological replicates. (Left) Scatter plots show pairwise comparisons of normalized RAMPAGE signal intensity across replicates representative human tissues and cell lines (GM12878, K562, heart, and liver), with Pearson correlation coefficients (R) and associated P-values indicated. (Right) The boxplot summarizes correlation coefficients across 50 replicate pairs, demonstrating high reproducibility (median R = 0.90).

# Supplementary Figure 3

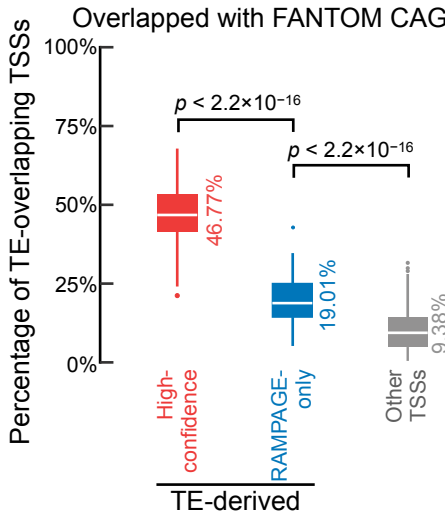
**A**



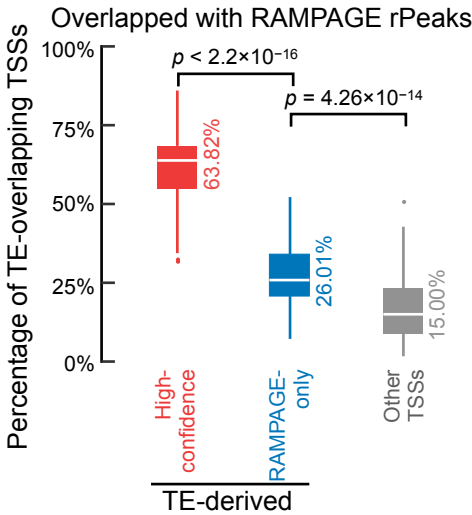
**B**



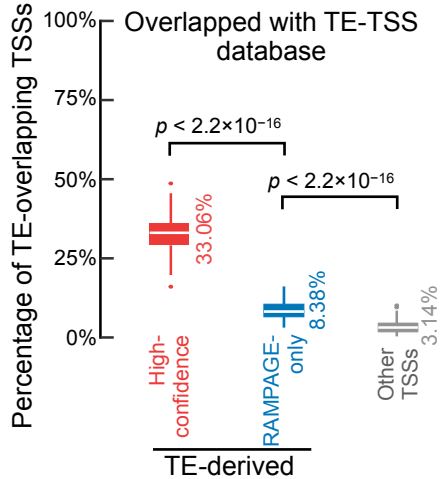
**C**



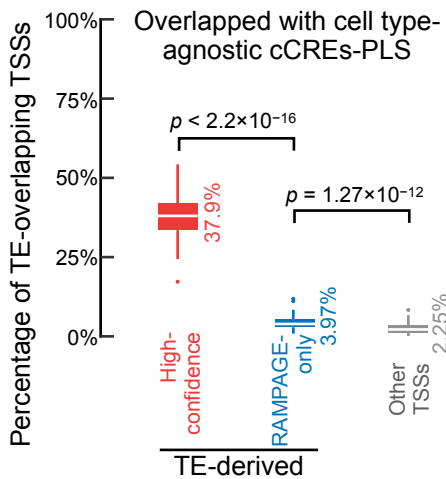
**D**



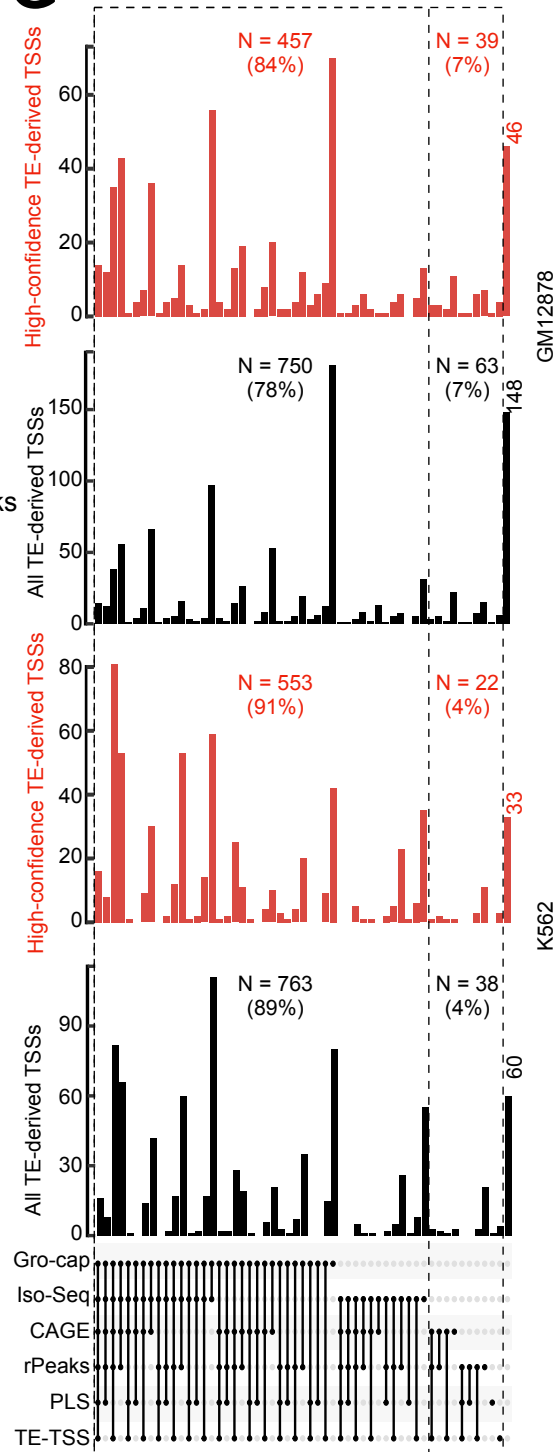
**E**



**F**



**G**



**Supplementary Figure 3.** Cross-validation of TE-derived TSSs with independent TSS annotations

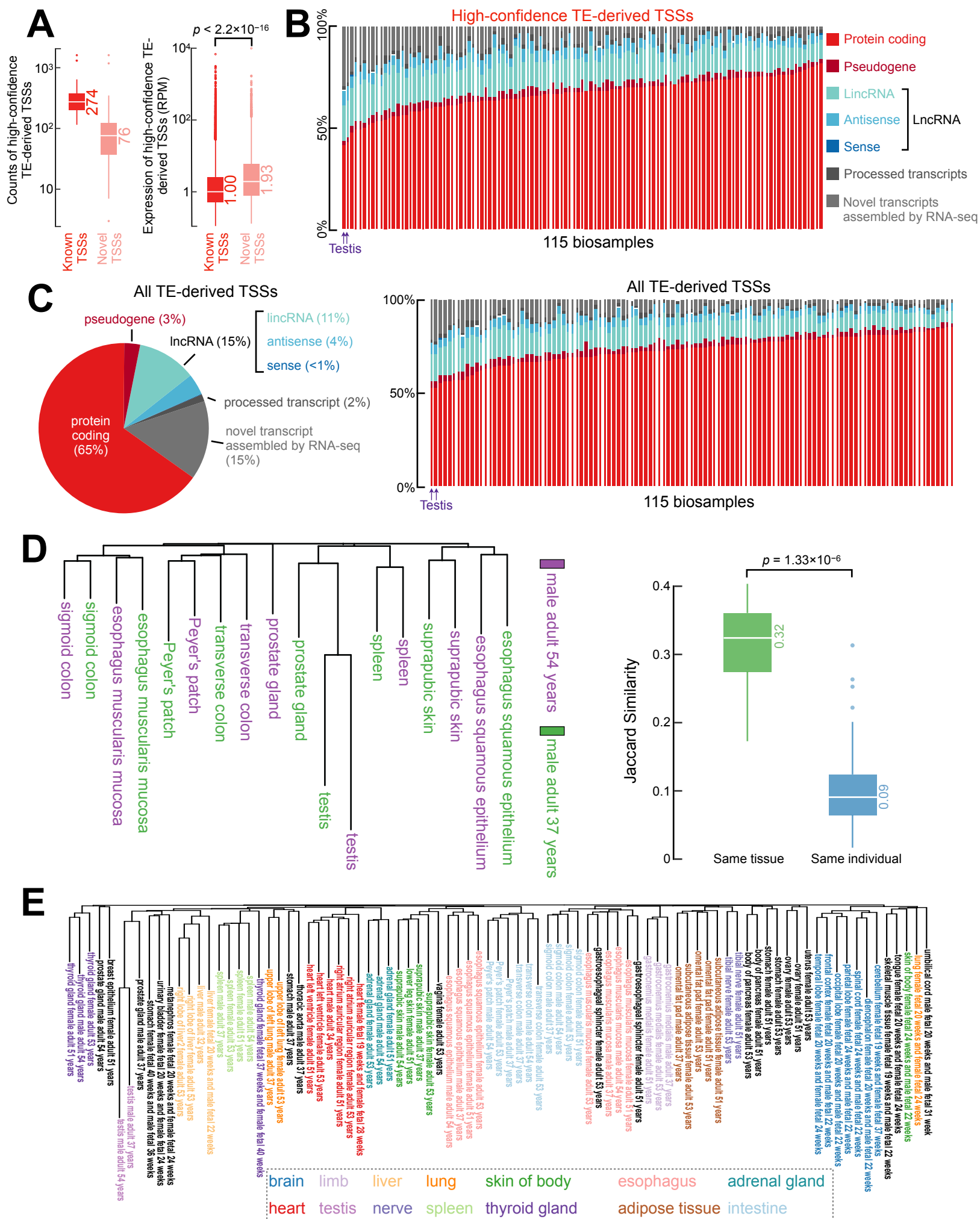
**A.** Comparison of TE-derived TSSs identified using RAMPAGE (red) and RNA-seq (blue) data across representative human tissues and cell lines (GM12878, K562, heart, liver, and testis).

**B.** (Left) Percentages of high-confidence TE-derived TSSs (red), RAMPAGE-only TE-derived TSSs (blue), and other TE-overlapping TSSs (grey) that overlapped with GRO-cap peaks in GM12878 (left bars) and K562 (right bars) cells. (Right) Percentages of high-confidence TE-derived TSSs (red), RAMPAGE-only TE-derived TSSs (blue), and other TE-overlapping TSSs (grey) that were also identified by PacBio Iso-Seq data in GM12878 (left bars) and K562 (right bars) cells.

**C-F.** Percentages of high-confidence TE-derived TSSs (red), RAMPAGE-only TE-derived TSSs (blue), and other TE-overlapping TSSs (grey) that overlapped with FANTOM CAGE peaks (C), RAMPAGE rPeaks (D), the TE-TSS database (E), and cell-type-agnostic cCREs-PLS (F). Each boxplot summarizes data from 115 biosamples, with the center line indicating the median, the box representing the interquartile range, and the whiskers extending to the most extreme values within  $1.5\times$  the interquartile range. One-sided Wilcoxon rank-sum test P-values are shown.

**G.** Upset plots showing the overlap of high-confidence TE-derived TSSs (red) and all TE-derived TSSs (black) with existing curated TSS catalogs in K562 (top) and GM12878 (bottom) cells.

## Supplementary Figure 4



#### **Supplementary Figure 4.** Characteristics of TE-derived TSSs and their host genes

**A.** Counts (left panel) and expression levels (right panel) of high-confidence TE-derived TSSs supported by known GENCODE or RefSeq TSSs (dark red) or novel TSSs from transcripts assembled using RNA-seq data (light red). Each boxplot summarizes data from 115 biosamples, with the center line indicating the median, the box representing the interquartile range, and the whiskers extending to the most extreme values within 1.5× the interquartile range. One-sided Wilcoxon rank-sum test P-values are shown.

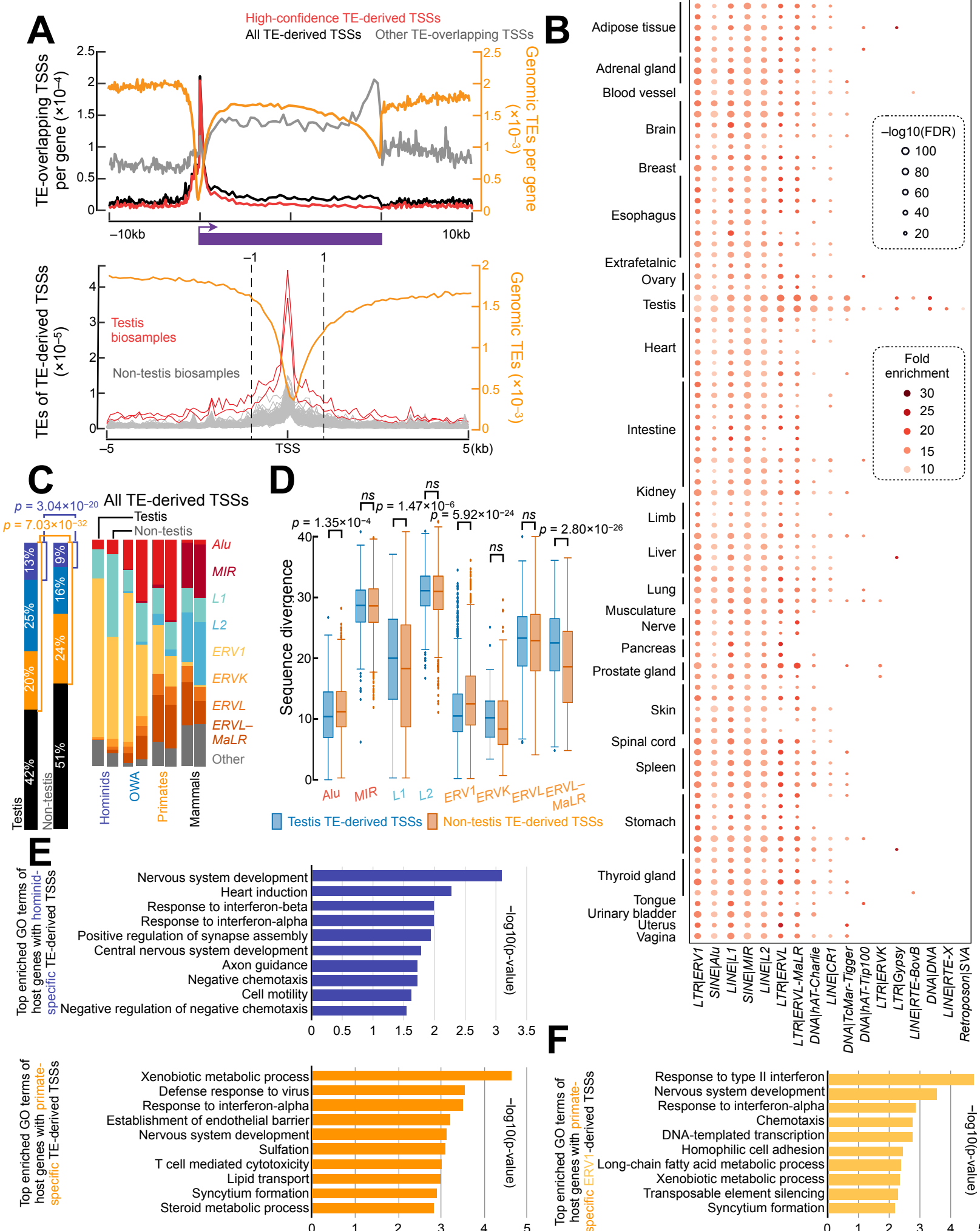
**B.** Bar plots show the types of host genes of high-confidence TE-derived TSSs in each biosample.

**C.** (Left) The pie chart is like Figure 2B but for all TE-derived TSSs. (Right) Bar plots show the types of host genes of all TE-derived TSSs in each biosample.

**D.** (Left) Hierarchical clustering of tissue samples from two donors by their expression profiles of TE-derived TSSs. (Right) Jaccard similarities between samples of the same tissue type from different donors are significantly higher than those across tissues from the same donor, as indicated by the one-sided Wilcoxon rank-sum test p-value. In each boxplot, the center line indicates the median, the box represents the interquartile range, and the whiskers extend to the most extreme values within 1.5× the interquartile range.

**E.** Hierarchical clustering of tissue biosamples by the expression profiles of TE-derived TSSs.

# Supplementary Figure 5



### **Supplementary Figure 5. Specific TE subfamilies contribute to TE-derived TSSs**

**A.** (Top) High-confidence TE-derived TSSs (red line) and all TE-derived TSSs (blue line) are significantly enriched at the 5'-end of genes, while other TE-overlapping TSSs (grey line) show a strong preference at the 3' UTR of genes. In contrast, genomic TEs (orange line) are generally depleted at the 5' and 3'-ends of genes. (Bottom) TE-derived TSSs in testis (red lines) and other tissues (grey lines) are significantly enriched at the 5'-end of genes, while genomic TEs (orange line) are significantly depleted at the 5'-end of genes.

**B.** Enrichment of TE subfamilies in TE-derived TSS formation across tissue types.

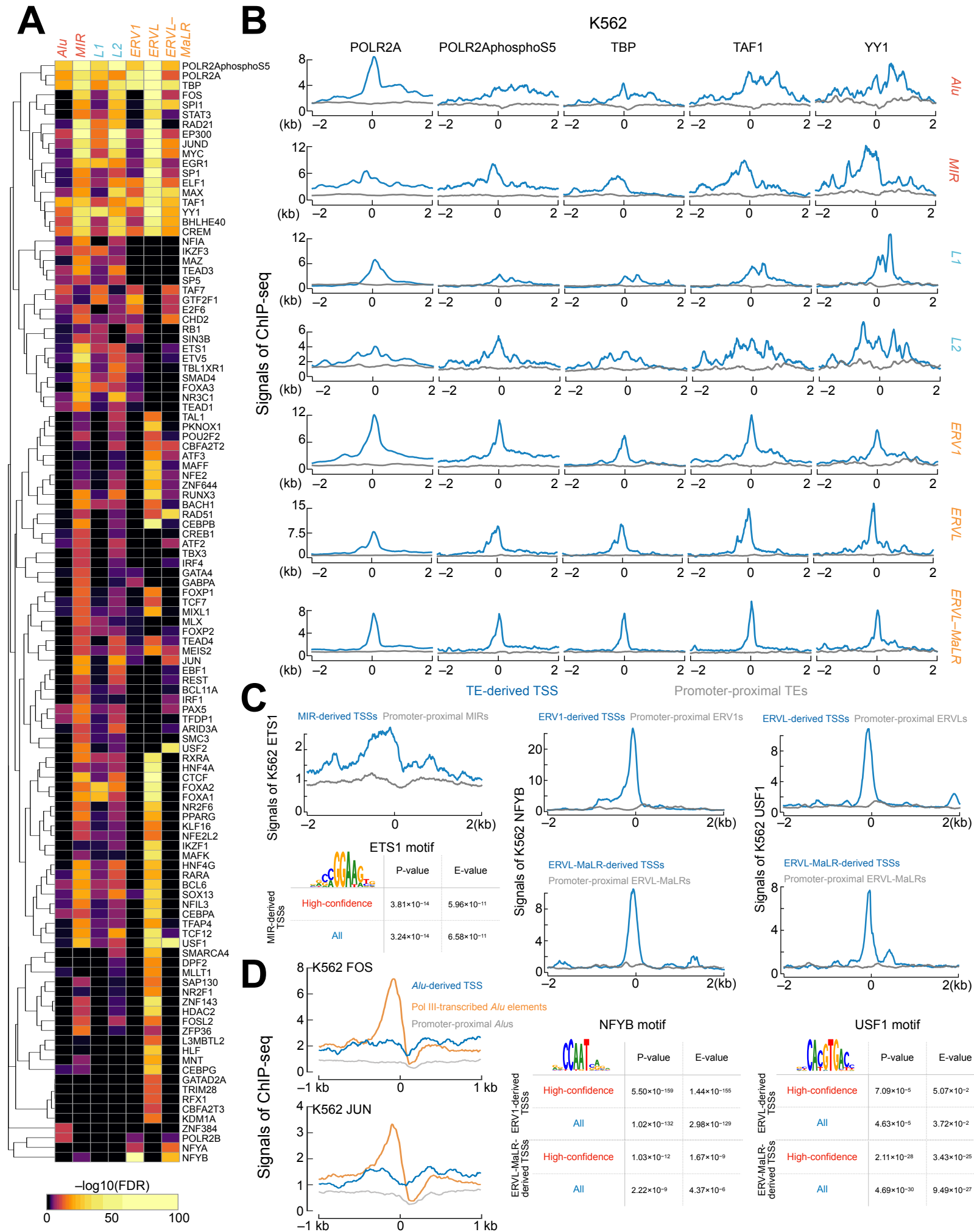
**C.** (Left) According to evolutionary age, TE-derived TSSs in testis and non-testis tissues were clustered into "Hominids", "Old World Anthropoids (OWA)", "Primates", and "Mammals" categories (see Methods for more details). Chi-squared test P-values are shown. (Right) Distribution of TE subfamilies in different evolutionary categories for TE-derived TSSs in testis and non-testis tissues.

**D.** Boxplots comparing the sequence divergence between TE-derived TSSs in testis and non-testis tissues in different TE subfamilies. In each boxplot, the center line indicates the median, the box represents the interquartile range, and the whiskers extend to the most extreme values within 1.5× the interquartile range. One-sided Wilcoxon rank-sum test P-values are shown.

**E.** Top 10 enriched GO terms for host genes with hominid-specific (top) and primate-specific (bottom) TE-derived TSSs, with their respective P-values derived from Fisher's exact test are shown.

**F.** Top 10 enriched GO terms for host genes containing primate-specific ERV1-derived TSSs, with their respective P-values derived from Fisher's exact test are shown.

Supplementary Figure 6



**Supplementary Figure 6.** Enrichment of transcription factor binding at TE-derived TSSs

**A.** Transcription factor enrichment at TE-derived TSSs for each TE subfamily (Fisher's exact test FDR were colored).

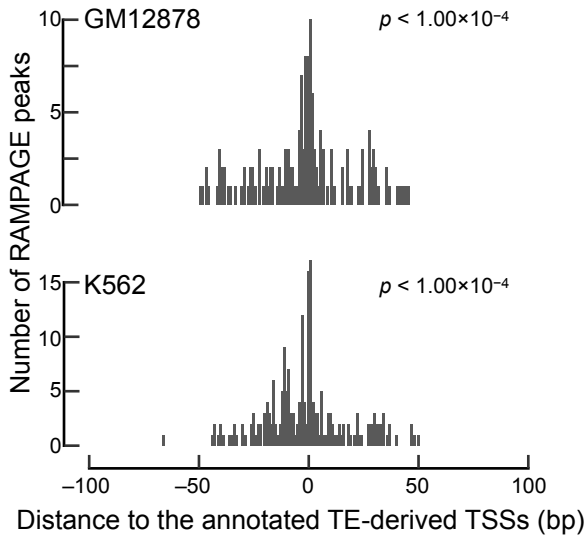
**B.** Average ChIP-seq signals for POLR2A, POLR2AphosphoS5, TBP, TAF1, and YY1 per TE family (rows) in the  $\pm 2$  kb regions around TSSs for TE-derived TSSs (blue line) and promoter-proximal TEs (grey lines) in K562 cell lines.

**C.** (Top) Average ChIP-seq signals for ETS1, NFYB, and USF1 in the  $\pm 2$  kb regions around TSSs for TE-derived TSSs (blue line) and promoter-proximal TEs (grey lines) from selected TE subfamilies in K562 cell lines. (Bottom) Motif enrichment of ETS1, NFYB, and USF1 at TE-derived TSSs from those subfamilies, with motif P-values and E-values calculated using FIMO.

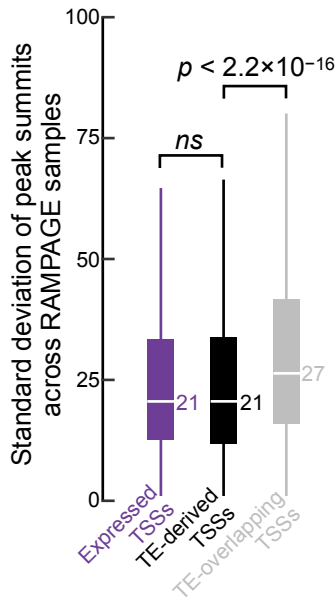
**D.** Average ChIP-seq signals for FOS and JUN in the  $\pm 1$  kb regions around Alu-derived TSSs (blue line), Pol III-transcribed Alu elements (orange line), and promoter-proximal Alu elements (grey lines) in K562 cell lines.

# Supplementary Figure 7

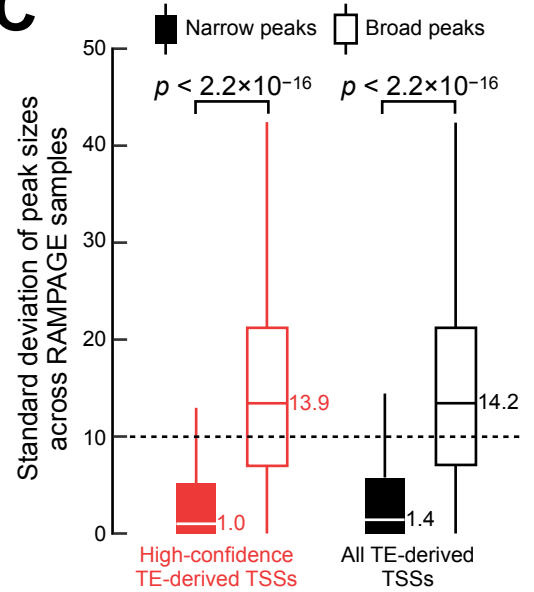
**A**



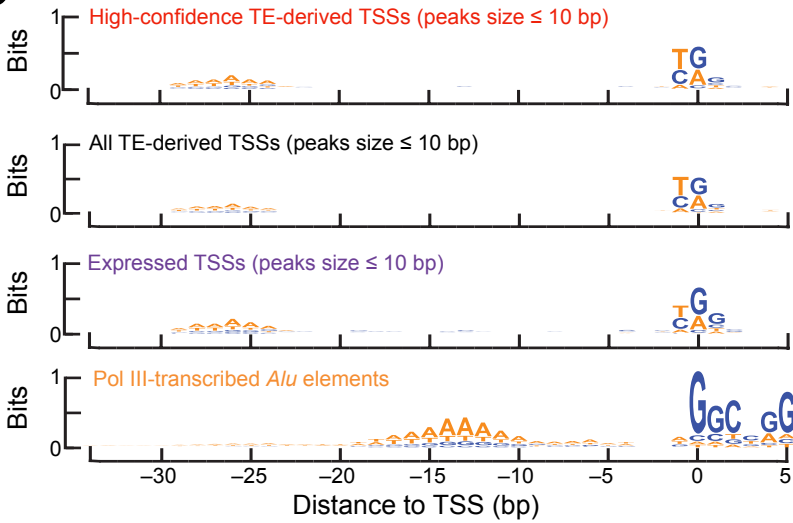
**B**



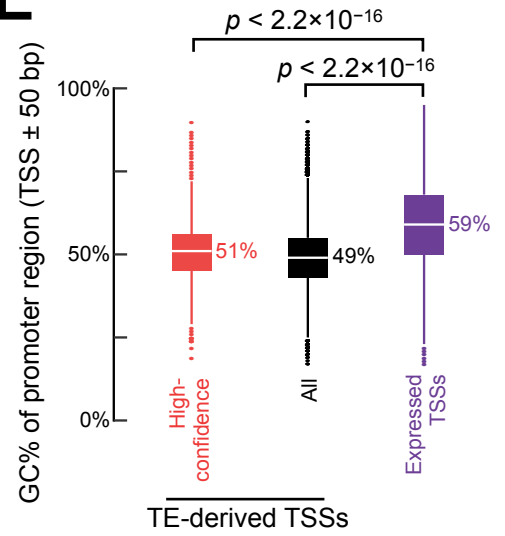
**C**



**D**



**E**



**Supplementary Figure 7.** TE-derived TSSs are enriched for TATA boxes and depleted of CpG islands

**A.** RAMPAGE peaks are enriched near the TE-derived TSSs. The significance of this enrichment was assessed using a P-value, which was obtained through 10,000 simulations of randomly assigning RAMPAGE peaks within a  $\pm 100$  base pair window centered on the TE-derived TSSs.

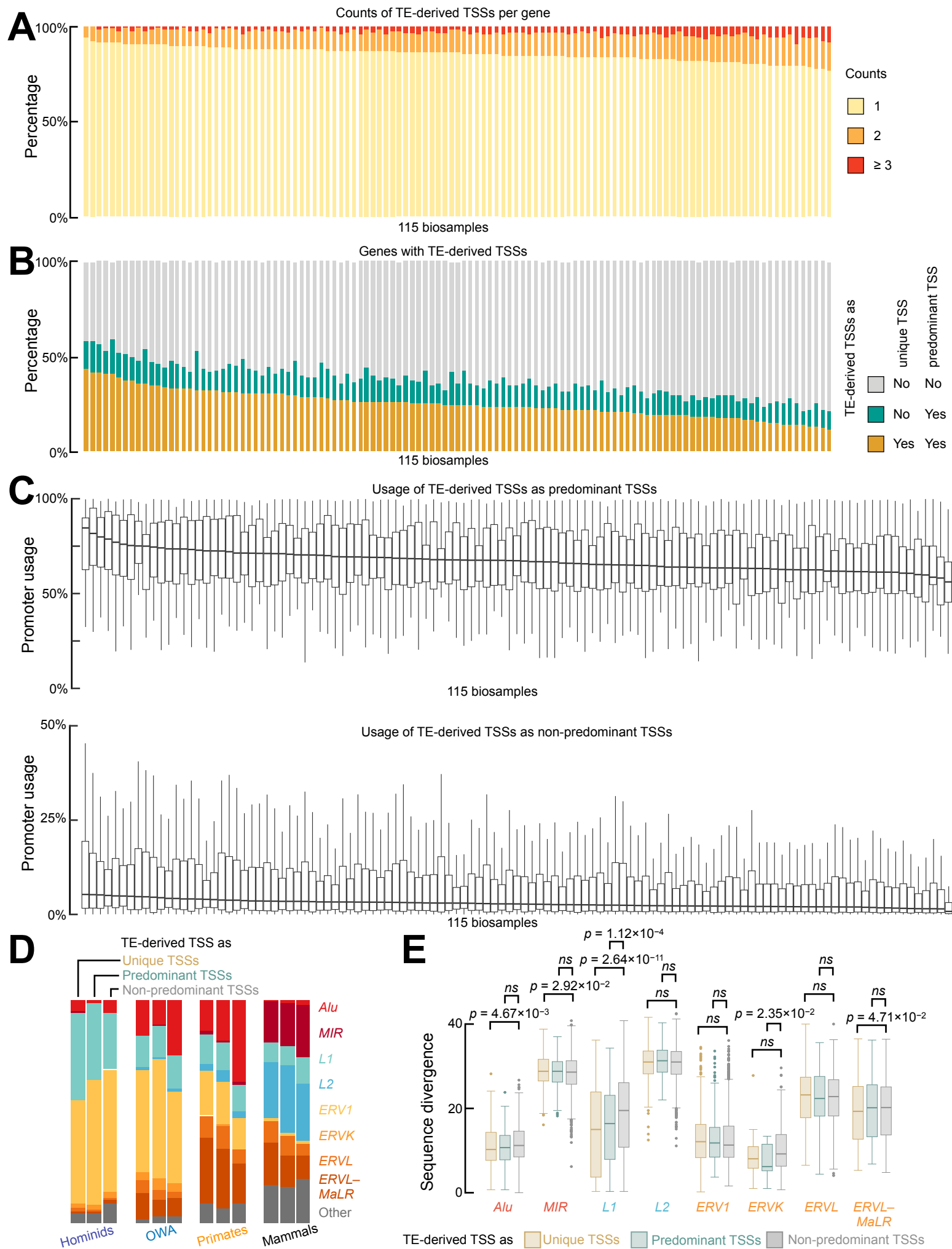
**B.** Standard deviation of peak summits across RAMPAGE samples for TE-derived TSSs (black), mRNA TSSs (purple), and other TE-overlapping TSSs (grey). In each boxplot, the center line indicates the median, the box represents the interquartile range, and the whiskers extend to the most extreme values within  $1.5\times$  the interquartile range. One-sided Wilcoxon rank-sum test P-values are shown.

**C.** Boxplots comparing the standard deviation of narrow and broad peak sizes for TE-derived TSSs across 115 RAMPAGE samples. In each boxplot, the center line indicates the median, the box represents the interquartile range, and the whiskers extend to the most extreme values within  $1.5\times$  the interquartile range. One-sided Wilcoxon rank-sum test P-values are shown.

**D.** Sequence logos of TATA-box and Inr motifs for high-confidence TE-derived TSSs (red), all TE-derived TSSs (black), expressed TSSs (purple), and TSSs of Pol III-transcribed Alu elements (orange).

**E.** GC content of high-confidence TE-derived TSSs (red), all TE-derived TSSs (black), and expressed TSSs (purple). In each boxplot, the center line indicates the median, the box represents the interquartile range, and the whiskers extend to the most extreme values within  $1.5\times$  the interquartile range. One-sided Wilcoxon rank-sum test P-values are shown.

## Supplementary Figure 8



**Supplementary Figure 8.** TE-derived TSSs contribute substantially to gene expression

**A.** Percentage of genes stratified by number of TE-derived TSSs across 115 biosamples.

**B.** Percentage of genes with TE-derived TSSs classified as unique (yellow), predominant (blue), or non-predominant (grey) across 115 biosamples.

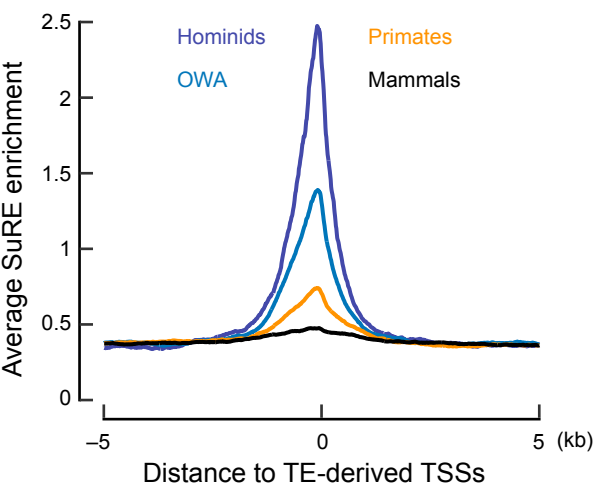
**C.** (Top) Usage of predominant TE-derived TSSs per gene across 115 biosamples. (Bottom) Usage of non-predominant TE-derived TSSs per gene across 115 biosamples. In each boxplot, the center line indicates the median, the box represents the interquartile range, and the whiskers extend to the most extreme values within 1.5× the interquartile range.

**D.** Distribution of TE subfamilies in different evolutionary categories for unique, predominant, and non-predominant TE-derived TSSs.

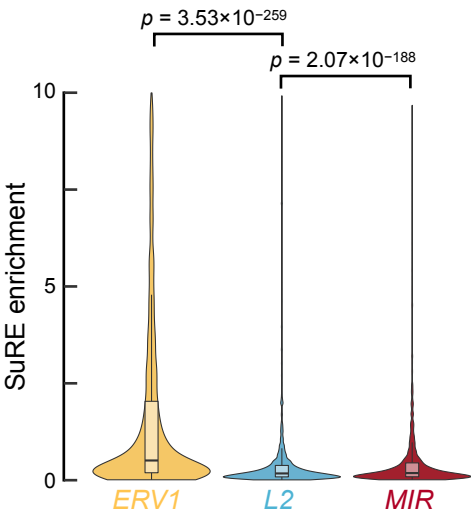
**E.** Boxplots comparing the sequence divergence across unique, predominant, and non-predominant TE-derived TSSs in different TE subfamilies. In each boxplot, the center line indicates the median, the box represents the interquartile range, and the whiskers extend to the most extreme values within 1.5× the interquartile range. One-sided Wilcoxon rank-sum test P-values are shown.

# Supplementary Figure 9

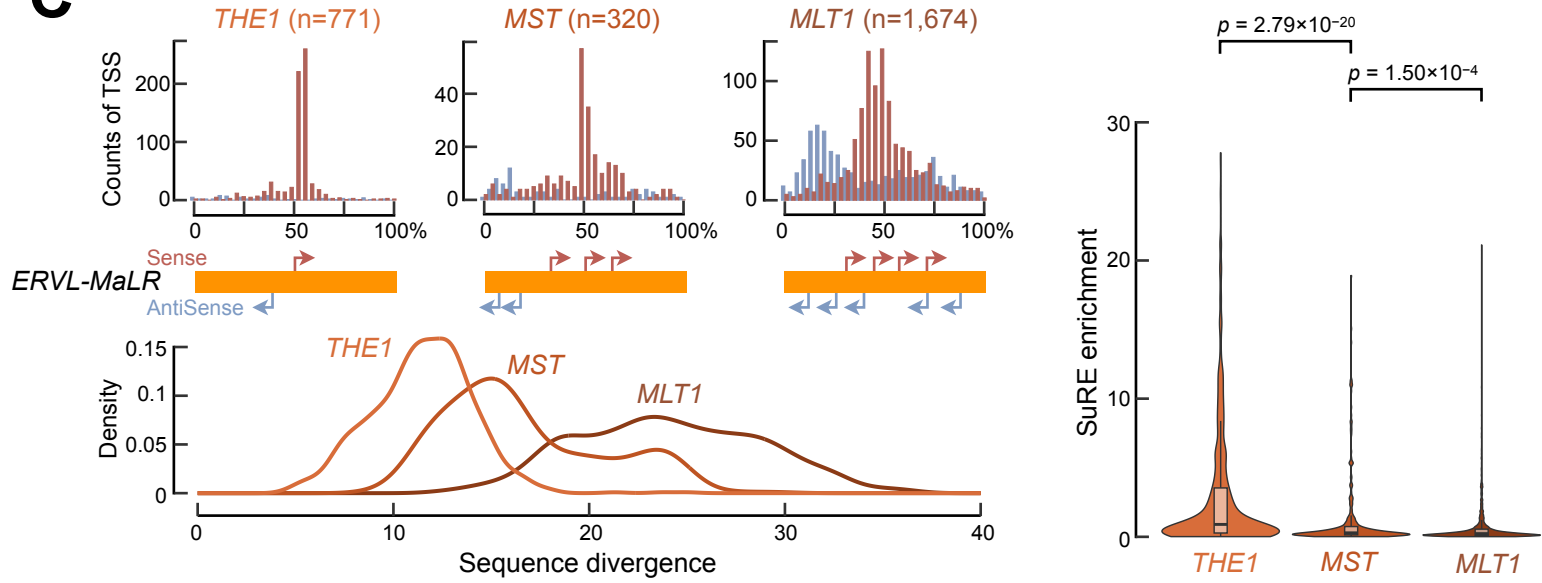
**A**



**B**



**C**



**Supplementary Figure 9.** Young TE-derived TSSs preserve intrinsic promoter architecture to drive precise transcription initiation

**A.** Average SuRE enrichment with respect to TE-derived TSSs belonging to “Hominids” (purple), “OWA” (blue), “Primate” (yellow), and “Mammals” (black) categories.

**B.** SuRE enrichment profiles of TE-derived TSSs arising from ERV1 (yellow), L2 (blue), and MIR (red) families. In each boxplot, the center line indicates the median, the box represents the interquartile range, and the whiskers extend to the most extreme values within 1.5× the interquartile range. One-sided Wilcoxon rank-sum test P-values are shown.

**C.** TSS positions (left top), sequence divergence (left bottom), and SuRE enrichment profiles (right) of TE-derived TSSs arising from different ERVL-MaLR (THE, n = 771; MST, n = 184; MLT, n = 1,674) subfamilies. In each boxplot, the center line indicates the median, the box represents the interquartile range, and the whiskers extend to the most extreme values within 1.5× the interquartile range.