Epigenetic characterization of pseudogenes across human tissues

Yunzhe Jiang, Beatrice Borsari and Mark Gerstein

Genome Res. published online April 15, 2026
Access the most recent version at doi:10.1101/gr.280768.125

Epigenetic characterization of pseudogenes across human tissues

Yunzhe Jiang[1,2], Beatrice Borsari[1,2,3,#], Mark B. Gerstein[1,2,4,5,6,#]

Affiliations

[1] Program in Computational Biology and Biomedical Informatics, Yale University, New Haven, CT 06520, USA

[2] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA

[3] Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona (UB), Barcelona, 08028, Spain

[4] Department of Computer Science, Yale University, New Haven, CT 06520, USA

[5] Department of Statistics & Data Science, Yale University, New Haven, CT 06520, USA

[6] Department of Biomedical Informatics & Data Science, Yale University, New Haven, CT 06520, USA

[#] Corresponding authors: Beatrice Borsari (beatrice.borsari@yale.edu, beatrice.borsari@ub.edu) and Mark B. Gerstein (mark@gersteinlab.org)

ABSTRACT

Pseudogenes have historically been regarded as non-functional remnants of genome evolution. However, relative to other noncoding genomic elements, their promoter architecture and epigenetic regulation remain incompletely understood. Here, we systematically characterize pseudogene promoters and compare them with those of protein-coding genes and long non-coding RNAs. To do this, we integrate matched transcriptomic and epigenomic data across 26 human tissues from the EN-TEx (ENCODE-GTEx) project. We uniformly annotate promoters with chromatin features (histone modifications, chromatin accessibility, and DNA methylation), sequence motifs, and evolutionary conservation, generating an online catalog. Leveraging this catalog, we show that, across multiple tissues, transcribed, unprocessed pseudogenes exhibit chromatin patterns similar to those of active protein-coding genes. In contrast, transcribed, processed pseudogenes show a strikingly different pattern: most lack the canonical hallmarks of transcription (e.g., active histone marks) at their promoters. Instead, their promoters show increased overlap with LINE elements, enrichment for YY1-like binding motifs, and higher Hi-C contact frequency, particularly with distal enhancer-like regulatory regions. Together with their greater conservation (relative to unprocessed pseudogenes), these features suggest that the transcription of processed pseudogenes may require regulatory mechanisms distinct from canonical promoter-associated epigenetic activation.

INTRODUCTION

Since the release of its first draft in 2001, considerable efforts have been devoted to improving the annotation of the human genome (Pruitt et al. 2007; Harrow et al. 2012; Frankish et al. 2019; Yates et al. 2020; Frankish et al. 2023). Early genomic studies primarily focused on the characterization of protein-coding genes, which constitute roughly 2% of the human genome, while less attention was paid to noncoding elements such as long noncoding RNAs (lncRNAs), microRNAs (miRNAs), and pseudogenes. More recently, the advent of high-throughput sequencing technologies, together with the development of multi-omic approaches, has enabled large-scale annotation and quantification of noncoding transcripts, particularly lncRNAs and miRNAs, and has facilitated epigenetic profiling of their genomic loci and regulatory landscapes (Ludwig et al. 2016; Uszczynska-Ratajczak et al. 2018; Kozomara et al. 2019; Zheng et al. 2019).

In contrast, pseudogenes have received less systematic study and remain comparatively undercharacterized as they were traditionally regarded as non-functional genomic remnants, often labeled as “junk DNA.” Pseudogenes are segments of DNA that resemble functional protein-coding genes but have lost coding capacity due to genetic mutations, such as the emergence of premature stop codons or frameshift mutations (Pei et al. 2012). Protein-coding genes that give rise to pseudogenes are referred to as parent genes and are functional paralogs of the pseudogenes. Studying pseudogenes can therefore be technically challenging due to their sequence similarity to parent genes, requiring accurate annotation of the human genome (Cheetham et al. 2020).

Pseudogenes are classified into three biotypes based on their mechanisms of origin (Mighell et al. 2000) (**Fig. 1A**): unprocessed pseudogenes, which derive from duplication of functional protein-coding genes; processed pseudogenes, which derive from retrotransposition of mature

messenger RNAs (mRNAs); and unitary pseudogenes, which accumulate fixed disablements in a given species (e.g., humans) while retaining functional orthologs in others (e.g., mice or non-human primates). Prior work has identified hundreds of both known and previously unannotated processed pseudogenes in the human genome that are widely transcribed (Troskie et al. 2021). Some pseudogenes also regulate the expression of their parent genes (Tam et al. 2008; Poliseno et al. 2010; Han et al. 2011; Singh et al. 2020). Given these regulatory roles, it is essential to understand their transcriptomic patterns and epigenetic regulation across tissues and cell types and how this regulation differs from that of protein-coding genes and lncRNAs.

To address the limited understanding of pseudogene promoter regulation in a tissue-resolved context, we leveraged matched multi-tissue transcriptomic and epigenomic profiles from the ENCODE-GTEx (EN-TEx) project (Rozowsky et al. 2023), spanning 26 adult human tissues. With these data, we first aimed to systematically annotate the epigenetic status of promoters associated with protein-coding genes, lncRNAs, and pseudogenes, generating a publicly available resource to facilitate exploration of the regulatory landscape of pseudogenes. Then, using this framework, we turned to investigate potential differences in epigenetic patterns between protein-coding genes and pseudogenes, as well as among different pseudogene biotypes.

RESULTS

Construction of a multi-tissue catalog of epigenetically and functionally annotated promoters for pseudogenes, lncRNAs, and protein-coding genes

With the goal of characterizing the epigenetic landscape of pseudogenes and comparing it to other gene biotypes in a tissue-specific context, we compiled a comprehensive catalog of promoter regions for pseudogenes, lncRNAs, and protein-coding genes annotated with their epigenetic activity. To annotate these promoter regions with chromatin activity profiles across

tissues, we leveraged EN-TEx epigenomic maps comprising chromatin immunoprecipitation sequencing (ChIP-seq) data for six histone modifications (i.e., H3K27ac, H3K4me3, H3K27me3, H3K36me3, H3K4me1, and H3K9me3), chromatin accessibility assays including DNase I hypersensitive site sequencing (DNase-seq) and assay for transposase-accessible chromatin using sequencing (ATAC-seq), as well as DNA methylation profiling (**Supplemental Table S1**). To this end, we defined promoters as ±1,000 base pair (bp) windows centered on the transcription start sites (TSSs) of all GENCODE v29 transcripts (**Table 1**) and annotated their activity based on the presence of active and repressive epigenetic features. The resulting resource provides a uniform characterization of promoter epigenetic activity across 26 human tissues (**Supplemental Table S2**). In addition to this epigenetic characterization, we included further promoter properties by analyzing guanine-cytosine (GC) content, evolutionary conservation, and mutation spectra based on single nucleotide variant (SNV) data from the Genome Aggregation Database (gnomAD) (Koenig et al. 2024) (**Supplemental Table S3**).

We then used this catalog to investigate the relationship between chromatin state and isoform expression across gene biotypes. We constructed a filtered set of expressed transcripts with well-characterized promoters (**Fig. 1B**). Briefly, we first curated a subset of non-overlapping promoter regions to minimize potential biases arising from overlapping transcripts of different genes. Next, we retained those whose promoters were experimentally supported by the presence of RNA Annotation and Mapping of Promoters for the Analysis of Gene Expression (RAMPAGE) peaks (Moore et al. 2022). Finally, we integrated matched RNA sequencing (RNA-seq) data at the isoform level from the same 26 human tissues. Specifically, for each gene in a given tissue, we selected the most highly expressed transcript, requiring an average expression level of ≥ 1 transcripts per million (TPM) across all available donors and their technical replicates (**Fig. 1C**; **Supplemental Figs. S1** and **S2**). We also used long-read RNA-seq data available from EN-TEx for the heart left ventricle to corroborate expression calls obtained from

short-read RNA-seq. At the gene level, 92% of protein-coding genes, 63% of lncRNA genes, and 59% of pseudogene loci identified as expressed had long-read support. This analysis highlighted that promoters of expressed protein-coding transcripts have stronger experimental support (75%) compared to those of lncRNAs (35%; Fisher's exact test, odds ratio = 5.48, *p*-value < 2.2×10^{-16}) and pseudogenes (37%; Fisher's exact test, odds ratio = 5.06, *p*-value < 2.2×10^{-16}) (**Fig. 1D**). By contrast, we did not observe substantial differences in the proportion of RAMPAGE-supported promoters across different pseudogene biotypes (**Fig. 1E**). In total, we retained 335 non-redundant pseudogenes, 490 lncRNAs, and 14,315 protein-coding transcripts for epigenetic characterization. Given the tissue-specific nature of our analysis, a single transcript could be represented across multiple tissues, yielding 100,657 transcript–tissue pairs for protein-coding genes, 1,022 for lncRNAs, and 1,664 for pseudogenes, which we further analyzed to investigate the interplay between expression and histone modifications at the level of individual tissues (**Table 2**).

Processed pseudogenes lack canonically active histone marks at their promoters despite being expressed

As expected (Hon et al. 2009), most protein-coding promoters exhibit active chromatin marks in the same tissue in which they are expressed, particularly H3K27ac and H3K4me3, with each covering over 75% of the total promoter length (**Fig. 2A**). In contrast, chromatin accessibility is lower, with peaks from ATAC-seq and DNase-seq covering an average of 42% and 24% of the promoter length, respectively. Promoters of lncRNAs exhibit intermediate levels of active mark coverage (~60%), whereas pseudogene promoters show the lowest (~25%). The paucity of active histone marks and open chromatin signatures at pseudogene promoters raised the question of whether this pattern is a general feature of all pseudogenes or specific to particular biotypes. Upon closer examination, we found that promoters of unprocessed and unitary pseudogenes exhibit epigenetic profiles more similar to those of protein-coding and lncRNA

transcripts, whereas processed pseudogenes are markedly depleted of active histone marks and open chromatin signatures (**Fig. 2B**). We also examined other histone modifications available from EN-TEx, including H3K4me1 (typically associated with primed enhancers and promoters) (Heintzman et al. 2007), H3K36me3 (enriched across actively transcribed gene bodies) (Barski et al. 2007), H3K27me3 (a repressive mark found at bivalent promoters) (Bernstein et al. 2006), and H3K9me3 (a hallmark of heterochromatic regions) (Lachner et al. 2001). With the exception of H3K4me1, which shows a distribution across promoter biotypes similar to that of canonically active epigenetic features, the proportions of promoters marked by the other modifications are minimal (**Supplemental Fig. S3**). When analyzed on a per-tissue basis, the profiles of active histone modifications and chromatin accessibility exhibit a consistent pattern across all tissues (**Supplemental Figs. S4–S7**). To ensure that our findings were not driven by transcript selection criteria, we repeated the analyses using alternative promoter definitions and gene expression thresholds. First, in addition to our primary analysis, which focused on the most expressed transcript per gene with an average TPM ≥ 1 across donors, we also included all transcripts meeting this expression level (**Supplemental Fig. S8**). Second, we relaxed the expression threshold to include transcripts with an average TPM ≥ 0.1 across donors (**Supplemental Figs. S9** and **S10**). Third, we redefined promoter regions as the 1,000 bp upstream of each TSS (**Supplemental Fig. S11**). Across all these alternative criteria, the observed chromatin patterns remained highly consistent, underscoring the robustness and reproducibility of our findings.

These findings align with the distribution of DNA methylation signals. We found that processed pseudogenes exhibit higher DNA methylation coverage, regardless of the promoter definition used (e.g., ±1,000 bp around the TSS or 1,000 bp upstream only), and this pattern is conserved across tissues (**Fig. 2C**; **Supplemental Figs. S12–S14**). That is, processed pseudogenes are more frequently heterogeneously methylated or hypermethylated compared to unprocessed and

unitary pseudogenes, lncRNAs, and protein-coding transcripts. Thus, they appear to be not only depleted of active chromatin features but also enriched in this kind of repressive signal. Furthermore, analysis of *in situ* Hi-C data from the transverse colon and skeletal muscle revealed that promoters of processed pseudogenes are more frequently located within inactive genomic compartments (e.g., B compartments) compared to those of protein-coding transcripts (Wilcoxon rank-sum test, *p*-value = 3.6×10^{-7} and 3.5×10^{-16}, respectively, false discovery rate [FDR]-adjusted) (**Fig. 2D**).

To rule out the possibility that these epigenetic differences are driven by underlying differences in expression, we compared the distribution of expression levels across tissues and gene biotypes. We found no significant difference in expression between processed pseudogenes and protein-coding transcripts (Wilcoxon rank-sum test, *p*-value = 0.95, FDR-adjusted) (**Fig. 2E**; **Supplemental Fig. S15**). Notably, despite their lack of active histone marks and open chromatin signatures, processed pseudogenes are more highly expressed than unprocessed pseudogenes (Wilcoxon rank-sum test, *p*-value = 2.0×10^{-31}, FDR-adjusted). These results indicate that epigenetic differences between processed and unprocessed pseudogenes cannot be explained simply by differences in expression magnitude. To further disentangle these regulatory patterns, we next investigated whether pseudogene expression mirrors that of their parent protein-coding genes across tissues. Cross-tissue co-expression analysis revealed that processed pseudogenes display significantly lower correlation with their parent genes than unprocessed pseudogenes (Wilcoxon rank-sum test, *p*-value = 4.1×10^{-12}), suggesting that processed pseudogene expression is less tightly coupled to the regulatory programs of their parent genes (**Fig. 2F**).

Differences in sequence homology and conservation underlie the differential chromatin marking of processed and unprocessed pseudogene promoters

Our findings thus far indicate that, across multiple tissues, processed pseudogenes generally lack canonical promoter-associated chromatin signatures, despite exhibiting expression levels similar to, or even higher than, those of protein-coding genes or unprocessed pseudogenes. We therefore examined whether underlying sequence features could explain these distinct chromatin patterns. Processed pseudogenes originate from the reverse transcription of spliced mRNAs and typically lack intronic regions (**Fig. 1A**). As a result, analyzing promoter sequences downstream of the TSS could introduce systematic biases across gene biotypes: downstream regions overlap exons in processed pseudogenes but introns in protein-coding transcripts, lncRNAs, and unprocessed or unitary pseudogenes. These regions are known to differ in conservation patterns (Park et al. 2014) and mutation rates (Frigola et al. 2017). To mitigate this bias, we restricted all subsequent analyses to the 1,000 bp region upstream of the TSS.

Because processed and unprocessed pseudogenes arise through distinct genomic mechanisms, we first considered the hypothesis that unprocessed pseudogenes tend to exhibit canonical active chromatin marking because they are more likely to inherit promoter regions from their parent genes, which usually display strong promoter-associated activity. In contrast, processed pseudogenes are not expected to inherit these upstream regulatory regions after retrotransposition (D'Errico et al. 2004), potentially explaining their lack of active chromatin features. To test this hypothesis, we quantified the extent to which unprocessed pseudogenes inherit upstream TSS regions from their parent genes (i.e., sequence homology). For each pseudogene–parent gene pair, we aligned the corresponding upstream regions and measured both the proportion of the parent promoter sequence included in the pseudogene locus and the sequence identity across the aligned region (**Fig. 3A**). Unprocessed pseudogenes that inherited upstream promoter sequence exhibit higher coverage of active histone marks and open

chromatin signatures compared to those that did not (**Fig. 3B**; **Supplemental Fig. S16**). In contrast, as expected, processed pseudogenes show near-zero alignment inclusion with promoter regions of parent protein-coding genes, confirming that they do not inherit upstream promoter sequences (Wilcoxon rank-sum test, *p*-value = 1.0×10^{-4}). Together, these results suggest that inheritance of upstream promoter sequence from the parent gene is associated with increased active chromatin marking at unprocessed pseudogene promoters.

In addition to exhibiting low sequence homology to their parent genes, these upstream regions in processed pseudogenes display significantly lower GC content compared to other gene biotypes (Wilcoxon rank-sum test, *p*-value = 1.0×10^{-27} vs. protein-coding transcripts; *p*-value = 4.4×10^{-19} vs. unprocessed pseudogenes, FDR-adjusted) (**Fig. 3C**; **Supplemental Fig. S17**). Given that high GC content is a common hallmark of functional promoters (Sandelin et al. 2007), these findings are in line with the lack of active chromatin features at processed pseudogene promoters and support the notion that these regions do not resemble canonical promoter sequences inherited from the parent gene.

However, these results alone do not explain how processed pseudogenes can be expressed. We therefore explored the genomic sequence context where processed pseudogenes tend to insert. Specifically, we analyzed interspecies conservation of upstream promoter regions by comparing the distribution of average phastCons scores across gene biotypes. Processed pseudogenes exhibit higher conservation than unprocessed pseudogenes, with levels comparable to protein-coding transcripts (Wilcoxon rank-sum test, *p*-value = 0.54 vs. protein-coding transcripts; *p*-value = 2.0×10^{-5} vs. unprocessed pseudogenes, FDR-adjusted) (**Fig. 3D**; **Supplemental Fig. S18**). Nevertheless, this degree of conservation is only weakly correlated with pseudogene expression levels, consistent with previous findings describing a complex relationship between gene expression levels and phylogenetic conservation (Pervouchine et al.

2015; Chan et al. 2009) (**Supplemental Fig. S19**). Altogether, this suggests that, over long evolutionary timescales, there is higher selective pressure on the genomic regions flanking processed pseudogenes, which could potentially support some degree of transcriptional activity without directly modulating expression magnitude.

In contrast, this selection signal is less evident within the human population. The upstream promoter regions of processed and unprocessed pseudogenes harbor a similar fraction of common variants (minor allele frequency [MAF] ≥ 1×10^{-3}; 3.8%, Fisher's exact test, *p*-value = 0.70), and this fraction is higher than that observed for protein-coding transcripts (3.0%, Fisher's exact test, *p*-value < 2.2×10^{-16}) (**Fig. 3E**; **Supplemental Fig. S20A**). Comparisons to parent genes led to similar results, with both processed (Fisher's exact test, odds ratio = 1.39, *p*-value < 2.2×10^{-16}) and unprocessed (Fisher's exact test, odds ratio = 1.16, *p*-value = 3.6×10^{-7}) pseudogenes showing enrichment for common variants. Moreover, variants in upstream promoter regions of processed pseudogenes are less deleterious than those reported by other gene biotypes (Wilcoxon rank-sum test, *p*-value = 0 vs. protein-coding transcripts; *p*-value = 1.5×10^{-137} vs. unprocessed pseudogenes, FDR-adjusted) (**Fig. 3F**; **Supplemental Fig. S20B**).

Processed and unprocessed pseudogene promoters are enriched in distinct classes of transposable elements

Epigenetically, the promoters of processed pseudogenes show some similarities with repetitive elements in the human genome (Liao et al. 2023). In fact, most transposable elements (TEs) are enriched in H3K9me3 (Ninova et al. 2019) (**Supplemental Fig. S21**) and DNA methylation (Ekram et al. 2012), the latter being similarly enriched in the promoters of processed pseudogenes (**Fig. 2C**). However, despite this epigenetic silencing, some TEs, especially long and short interspersed nuclear elements (LINEs and SINEs), are known to be transcriptionally active under certain conditions (Picault et al. 2009; Reichmann et al. 2012; Gnanakkan et al.

2013). For instance, research has shown that H4K16ac activates the transcription of LINE-1 elements in the H9 embryonic stem cell line (Pal et al. 2023).

Thus, we explored the possibility that there could be shared regulatory mechanisms between TEs and processed pseudogenes. We analyzed the overlap between promoter sets and annotated TEs (**Fig. 4A**; **Supplemental Fig. S22**). Pseudogene promoters show a higher frequency of overlap with TEs compared to protein-coding promoters, with distinct patterns between processed and unprocessed pseudogenes. Specifically, among processed pseudogenes, 23% of promoters overlap exclusively with LINEs, 24% with SINEs, and 16% with both LINEs and SINEs. In contrast, unprocessed pseudogenes display a pronounced SINE bias: only 5% of their promoters show LINE-only overlap, whereas 35% overlap exclusively with SINEs and 16% overlap with both LINEs and SINEs. Furthermore, the degree of sequence overlap (as measured by the Jaccard index) with both LINEs and SINEs differs significantly among gene biotypes (**Fig. 4B**). Again, processed and unprocessed pseudogenes show different patterns, reporting the highest overlap with LINEs and SINEs, respectively.

Because pseudogenes originate from a non-random subset of protein-coding genes, we next asked whether enrichment of TEs in pseudogene promoters could be explained by inheritance from corresponding parent gene promoters, particularly for unprocessed pseudogenes, which are likely to retain their original genomic context. To test this, we compared TE overlap frequencies between pseudogene promoters and those of their respective parent genes (**Fig. 4C**). For unprocessed pseudogenes, we observed no significant difference in SINE enrichment relative to their parent gene promoters (two-sample proportion test, 50% vs. 38%, p-value = 0.16). This result is consistent with the possibility that SINE-rich promoter architectures are inherited rather than newly acquired, and therefore cannot be excluded as a contributing factor to the observed SINE bias in unprocessed pseudogene promoters.

For completeness, we extended this analysis to processed pseudogenes, despite their lack of promoter inheritance. In this case, we compared LINE overlap between processed pseudogene promoters and their parent gene promoters and found significantly higher LINE enrichment in processed pseudogene promoters (two-sample proportion test, 39% vs. 13%, *p*-value = 4.8×10^{-5}). This suggests that LINEs are preferentially associated with processed pseudogene promoters independently of the parental promoter context. However, with the current data, we cannot distinguish whether LINE insertions predate or follow integration of the processed pseudogene, leaving the temporal order of these events unresolved.

An additional concern is that epigenetic signals at processed pseudogene promoters could be dominated by the TE sequences themselves. We explored whether the observed epigenetic pattern at these promoters might originate from overlapping TEs rather than from the pseudogenes themselves. To address this concern, we stratified processed pseudogene promoters according to their overlap with LINEs and SINEs. The overlapping and non-overlapping groups showed similar coverage proportions of histone marks and chromatin accessibility signatures, indicating that the observed epigenetic patterns at processed pseudogene promoters are unrelated to the presence of LINEs or SINEs (**Fig. 4D**).

Finally, to evaluate whether TE-associated regulatory features might contribute to pseudogene promoter activity, we examined H4K16ac, a mark previously reported to preferentially activate transcription of LINE-1 elements (Pal et al. 2023). After cross-referencing RNA-seq and CUT&Tag H4K16ac peaks from the H9 embryonic stem cell line with our catalog of pseudogene promoters (Pal et al. 2023; Reese et al. 2023), we did not observe significant enrichment of this mark at processed pseudogene promoters, suggesting that the activating role of H4K16ac may be specific to TEs (**Supplemental Fig. S23**).

YY1-like motifs are enriched at the promoters of processed pseudogenes

The strong co-localization of processed pseudogenes and LINEs, together with their epigenetic similarities, prompted us to investigate whether they share additional regulatory features. Specifically, we hypothesized that processed pseudogenes could, like LINEs, possess an independent mechanism to initiate transcription. LINEs are autonomous TEs whose promoter activity is embedded within the transposed sequence rather than the upstream genomic context. They typically harbor an internal RNA polymerase II promoter, which allows them to be transcribed independently (Beck et al. 2011). SINEs, despite being non-autonomous and relying on LINE-encoded proteins for reverse transcription and integration, are nevertheless transcribed by RNA polymerase III via internal promoter elements derived from their tRNA- or 7SL-related origins, such as *Alu*, B1, and B2 elements (Kramerov and Vassetzky 2011). By analogy, we searched for intrinsic promoter-like activity in processed pseudogenes that might reside near the 5′ end of the retrotransposed sequence, immediately downstream of the annotated TSS. However, among all epigenetic features analyzed, we observed only a modest enrichment of H3K4me3 peaks in the 1,000 bp promoter region downstream of the TSS (McNemar test, *p*-value = 0.03, FDR-adjusted).

Next, we identified transcription factor motifs that could potentially mediate processed pseudogene transcription. We found that 20% of processed pseudogene promoters contain at least one YY1-like transcription factor binding motif (**Fig. 5A**). Specifically, we identified *de novo* human- and mouse-derived YY1-like motifs in 10% and 12% of processed pseudogene promoters, respectively (*p*-value = 1×10^{-38} and *p*-value = 1×10^{-13}) (**Supplemental Fig. S24**). These results indicate significant enrichment of YY1-like motifs in processed pseudogene promoters relative to those of unprocessed pseudogenes, lncRNAs, and protein-coding transcripts. Moreover, we confirmed this enrichment by comparing pseudogene promoters with

their parent genes, suggesting that accumulation of human- and mouse-derived YY1-like motifs in processed pseudogene promoters is independent of parent gene origin (**Fig. 5B**). We also assessed whether the YY1-like motifs exhibit strand bias relative to the promoters of processed pseudogenes. We did not detect any substantial bias, with 58% of occurrences residing on the same strand as the promoter. We obtained similar results for a related transcription factor, YY2, which binds to the same consensus sequence as YY1 but with a low affinity (Kim et al. 2007) and for which we reported a *de novo* motif in ~30% of processed pseudogene promoters (*p*-value = 1×10^{-18}) (**Supplemental Fig. S25**). We also analyzed the distribution of these motifs relative to pseudogene TSSs and observed distinct patterns. Specifically, YY1-like motifs in the promoters of processed pseudogenes display strong peaks centered on the TSS (within 10 bp), whereas those in protein-coding transcripts and unprocessed pseudogenes are distributed more uniformly (**Fig. 5C**; **Supplemental Fig. S26**). This indicates that YY1-like motifs are not only more abundant in processed pseudogene promoters but also preferentially positioned near their TSSs. Of note, YY1 has previously been implicated in the activation of TEs. In humans, the 5′ untranslated region (5′-UTR) of the LINE-1 elements harbors a consensus YY1-binding site within its core promoter, which is essential for accurate transcription initiation (Athanikar et al. 2004). Similarly, in mice, functional YY1-binding sites have been identified in the 5′-UTR monomers of the *Tf_I* subfamily of LINE-1 elements, where YY1 acts as a transcriptional activator (Saha et al. 2024). Given the substantial overlap between processed pseudogene promoters and LINEs reported above (**Fig. 4A**), we next assessed whether any of the 153 motif-associated processed pseudogene promoters overlapped with LINE sequences, and found only one such instance. We also computed the linear distance between YY1-like motifs in promoters and the nearest LINE-1 element. Processed pseudogenes exhibit significantly shorter distances than protein-coding transcripts and lncRNAs, although their distances do not differ significantly from those of unprocessed pseudogenes (**Supplemental Fig. S27**). Overall, these results indicate that the presence of YY1-like motifs at processed pseudogene promoters generally

occurs independently of LINE sequences. Nonetheless, the reduced LINE-1 proximity suggests some degree of shared regulatory architecture between processed pseudogenes and LINE-1 elements.

Besides the YY1- and YY2-like motifs, we also found several other motifs significantly enriched at promoter regions of processed pseudogenes (**Supplemental Fig. S28**). For example, a previously characterized *c-Myc* (bHLH) binding motif was significantly enriched in the promoter regions of processed compared to unprocessed pseudogenes (7% vs. 1%; *p*-value = 1×10^{-3}; *q*-value = 0.0017, FDR-adjusted), consistent with prior evidence showing that MYC can transcriptionally activate pseudogenes such as HMGA1P (Tian et al. 2020).

Processed pseudogene promoters show increased Hi-C contacts and preferential interactions with distal enhancer-like elements

Given that processed pseudogenes lack canonical regulatory signatures at their promoters, we also explored other potential mechanisms that could be implicated in their expression. Specifically, we analyzed whether they may establish contact with other regions in the genome that may be associated with transcriptional regulation. We obtained intact Hi-C loop data from the ENCODE portal that matched our tissue collection and cross-referenced these data against our promoter catalog. We found that pseudogenes, and in particular processed pseudogenes, tend to contact broader genomic intervals (e.g., 10 kb and 25 kb) compared to protein-coding and lncRNA transcripts (**Fig. 5D** and **5E**).

Furthermore, the promoters of processed pseudogenes exhibit a higher contact frequency than those of unprocessed pseudogenes and protein-coding transcripts (Wilcoxon rank-sum test, FDR-adjusted *p*-value = 4.4×10^{-7} and 6.2×10^{-6}, respectively) (**Fig. 5F**). To examine which genomic regions these promoters preferentially contact, we intersected Hi-C interaction intervals

with annotated human candidate *cis*-regulatory elements (cCREs) (Moore et al. 2026). This analysis revealed marked differences in the chromatin interactions of unprocessed and processed pseudogenes. Unprocessed pseudogenes, which display epigenetically active promoters, preferentially contact promoter-like signature (PLS) and proximal enhancer-like signature (pELS) cCREs (**Fig. 5G**; **Supplemental Fig. S29**). In contrast, processed pseudogenes, which are characterized by less active promoter states, more frequently interact with distal enhancer-like signature cCREs (dELS) (**Fig. 5G**). Because YY1 has been implicated in mediating promoter–enhancer looping (Gao et al. 2023; Weintraub et al. 2017), we next examined whether any of the 153 YY1-associated processed pseudogene promoters participated in these Hi-C contacts. However, only a small subset (18 promoters, or 12%) shows such interactions, indicating that YY1 is unlikely to be the main mediator of these long-range contacts. Together, these results suggest that processed pseudogene expression may be influenced by distal regulatory elements, but that such interactions are largely independent of YY1-mediated chromatin looping.

DISCUSSION

In this study, we extensively characterized the promoter activity of pseudogenes across 26 adult human tissues using matched EN-TEx transcriptomic, epigenomic, and three-dimensional chromatin interaction data, complemented by annotations of sequence features, evolutionary conservation, and population genetic variation. We constructed a comprehensive promoter catalog for transcribed pseudogenes, lncRNAs and protein-coding genes, and used this resource to interrogate pseudogene regulation across human tissues. Mining this dataset, we identified distinct regulatory patterns not only between protein-coding genes and pseudogenes, but also among different pseudogene biotypes. Promoter-associated epigenetic profiles are broadly concordant with expression for many unprocessed and unitary pseudogenes, which more frequently exhibit canonical active histone marks and chromatin accessibility, resembling

protein-coding genes. In contrast, even when expressed, processed pseudogenes display a distinct promoter architecture characterized by depletion of canonical active histone marks and chromatin accessibility, coupled with increased DNA methylation and a greater tendency to reside in inactive chromatin compartments. We further demonstrated that these differences reflect both how pseudogenes are formed and where they reside in the genome. Because processed pseudogenes do not inherit upstream promoter sequences from their parent genes (D'Errico et al. 2004), their upstream regions are expected to have a different pattern of conservation than unprocessed pseudogenes. In fact, we observe they tend to reside in more evolutionarily conserved sequences and exhibit association with distinct transposable elements and transcription factors. Together, these results define distinct promoter architectures for expressed processed pseudogenes and motivate testable hypotheses for how they can be transcribed despite limited canonical promoter-associated activation, thereby offering a framework for investigating their regulatory and potential functional roles in the human genome.

Prior to our study, it was unclear whether processed pseudogenes, following retrotransposition events, could acquire epigenetic features resembling normal promoters, and, if not, what alternative regulatory mechanisms might support their transcription. Our motif enrichment and chromatin interaction analyses are consistent with the latter possibility: processed pseudogene promoters are enriched for YY1-like motif occurrences, a feature they did not inherit from their parent genes, and they show increased chromatin interactions whose intervals more often intersect distal enhancer-like cCREs. These findings suggest that, in the absence of a parent-derived promoter, processed pseudogene promoters may adopt regulatory mechanisms that support transcription despite limited canonical epigenetic activation in the immediate vicinity of the annotated TSS.

However, several caveats should be considered when interpreting these results. First, experimental validation is needed to confirm the computational inferences presented here. Although the identified motifs resemble known YY1/YY2 binding sites, they have not yet been experimentally verified. Future work—such as ChIP-seq or CUT&Tag assays targeting YY1 in tissues corresponding to the EN-TEx collection, combined with functional perturbations such as CRISPR interference of pseudogene loci or YY1 knockdown—will be crucial to determine whether YY1 causally mediates processed pseudogene transcription. Second, the mechanism by which YY1-like motifs might mediate transcription of processed pseudogenes remains unclear. The canonical YY1 motif has been implicated in diverse modes of gene activation, from facilitating long-range promoter–enhancer interactions (Gao et al. 2023; Weintraub et al. 2017) to promoting directional transcription initiation (Dudnyk et al. 2024). It also plays key roles in the regulation of TEs, which share epigenetic similarities with processed pseudogenes and often overlap their promoters, particularly in ensuring transcriptional fidelity at human LINE-1 elements (Athanikar et al. 2004). In our data, however, co-localization of the YY1-like motif with either long-range chromatin interactions or LINEs at processed pseudogene promoters is rare. This suggests that, if the motif contributes to processed pseudogene expression, it is unlikely to do so via long-range looping or LINE-associated mechanisms and may instead act through alternative or context-specific regulatory modes.

Overall, our results are consistent when comparing tissues derived from both male and female donors (e.g., stomach and lung) as well as those restricted to a specific sex (e.g., testis and uterus), indicating that sample sex composition does not substantially influence the core results of our analyses. However, this may limit studies exploring sex-specific roles of pseudogenes. Additionally, promoters were defined based on a 2,000-bp window around each reference TSS, which may not fully capture the complexity of promoter architecture and regulatory interactions. Ongoing efforts by the GENCODE project, including integration of CapTrap-Seq data

(Carbonell-Sala et al. 2024) and application of the deep learning model ProCapNet (Cochran et al. 2024), aim to refine promoter annotations and improve accuracy in defining TSSs. Finally, we note that additional histone modifications beyond those profiled in EN-TEx can mark active regulatory regions of the genome and may show enrichment at processed pseudogene promoters (Pradeepa et al. 2016; Wolfe et al. 2021).

Despite these limitations, our work provides a refined understanding of pseudogene promoter architecture and function in human tissues, underscoring the evolutionary and functional diversity of promoter regulation across distinct pseudogene types and establishing a foundation for future investigations into the regulatory roles of pseudogenes and other non-canonical transcripts. Given increasing evidence implicating pseudogenes in human development (Qian et al. 2022) and various diseases (Han et al. 2014; Qi et al. 2021; Liu et al. 2021), further exploration of their epigenetic and transcriptional patterns may reveal novel biological roles and potential therapeutic targets. As our understanding of the non-coding genome continues to evolve, pseudogenes may be viewed not only as vestigial elements but also as epigenetically regulated genomic features with potential functional relevance. We anticipate that the catalog and integrative analyses developed in this study will serve as a resource for the research community, motivating continued exploration into the regulatory potential of the human non-coding genome.

METHODS

Assigning promoter regions of protein-coding transcripts, lncRNAs, and pseudogenes

To streamline the analysis of numerous functional assays and maintain consistency with the ENCODE uniform analysis pipeline (Hitz et al. 2023), we utilized the GENCODE v29 annotation based on the GRCh38 human genome assembly. We included genes and their associated transcripts, focusing specifically on the following gene biotypes as shown in **Table 1**.

For each transcript, we used the GENCODE manually curated annotation and defined the reference TSS as the 5′ end of the annotated transcript, taking transcript strand into account. Unless otherwise specified, promoter regions were defined as a ±1,000 bp window centered on this reference TSS (i.e., 1,000 bp upstream and 1,000 bp downstream relative to the direction of transcription). For sequence feature analyses, to avoid biases introduced by downstream genic context differences across biotypes, we restricted the analysis to the 1,000 bp region upstream of the reference TSS. As a sensitivity analysis, we repeated key epigenetic characterization analyses using an alternative promoter window definition (i.e., the 1,000 bp region upstream of the reference TSS) to confirm the robustness of our conclusions.

Filtering transcripts by annotation, expression, and RAMPAGE peaks

To unambiguously characterize promoters, we removed transcripts whose promoters overlap with those of transcripts from different genes in a head-to-head, tail-to-tail, or unidirectional manner (Wright et al. 2022) in subsequent analyses. In a unidirectional overlap, both promoters are located on the same strand, where one starts before the other has ended. In head-to-head and tail-to-tail overlaps, the promoters are on opposite strands. Head-to-head indicates that the TSSs face each other, whereas tail-to-tail means the termination sites are aligned. In this way, we included a total of 14,253 pseudogenes, 17,624 lncRNAs, and 106,643 protein-coding transcripts (the third row of **Table 2**).

We downloaded isoform-level quantifications from the EN-TEx portal (entex.encodeproject.org) for matched human tissues. Transcripts with an average expression ≥ 1 TPM across donors and technical replicates were retained (the fourth row of **Table 2**). Using this threshold, our estimates of pseudogene abundance and transcription were consistent with previous reports that identified ~10% of all pseudogenes as detectably expressed based on a cutoff of fragments per kilobase of transcript per million mapped reads > 0.5, with approximately 50% of these expressed pseudogenes restricted to a single tissue, recapitulating previously described tissue-specificity patterns (Rozowsky et al. 2023). Finally, for each gene and tissue, the most highly expressed transcript was selected as its representative and used for our primary analyses, yielding 1,267 pseudogenes, 1,126 lncRNAs, and 134,022 protein-coding transcripts across tissues (the fifth row of **Table 2**). For the heart left ventricle, we used available EN-TEx long-read RNA-seq data from a single donor to corroborate expression calls derived from short-read RNA-seq. We mapped each expressed transcript to its corresponding gene and assessed support at the gene level by determining whether that gene had at least one supporting long read assigned.

RAMPAGE is capable of capturing 5'-complete complementary DNAs, allowing for precise identification of TSSs and accurate quantification of promoter activity (Batut and Gingeras 2013). To validate our assignment of promoters, we leveraged a manually integrated collection of representative RAMPAGE peaks (Moore et al. 2022). We employed BEDTools intersect (v2.30.0) (Quinlan and Hall 2010) to find overlaps between the promoters of expressed transcripts and representative RAMPAGE peaks. If a RAMPAGE peak overlapped a promoter by at least 1 bp, we considered the promoter to be experimentally supported (the sixth row of **Table 2**).

Profiling epigenetic patterns with matched tissue-specific data

To characterize epigenetic patterns associated with each transcript, we downloaded tissue-specific histone ChIP-seq peaks, ATAC-seq peaks, and DNase-seq peaks from the EN-TEx portal (entex.encodeproject.org) across multiple human tissues from four donors. Specifically, we downloaded histone ChIP-seq peaks for H3K27ac, H3K4me3, H3K4me1, H3K36me3, H3K9me3, and H3K27me3. Certain tissues, such as the omental fat pad and subcutaneous adipose tissue, were excluded due to insufficient RNA-seq and/or histone ChIP-seq data. Thus, tissue-specific epigenomic data were available for 26 human tissues.

For each assay and tissue, we merged peaks across donors and technical replicates using BEDTools merge (v2.30.0) (Quinlan and Hall 2010) with the -i stdin -c 1 -o count options. We then employed BEDTools intersect (v2.30.0) (Quinlan and Hall 2010) with the -wao option to calculate the base-pair overlap between defined promoters and peaks derived from histone ChIP-seq, ATAC-seq, and DNase-seq data (**Supplemental Table S2**). For these peaks, we calculated the proportion of promoter bases overlapping peaks for each transcript in each tissue. We then aggregated these values, calculating the average proportion of coverage across transcripts and tissues for each biotype.

Additionally, we obtained tissue-specific DNA methylation profiling by array assay across the same tissues. To ensure consistency across donors and technical replicates for each tissue, we used BEDTools merge (v2.30.0) (Quinlan and Hall 2010) with the -c 5 -o mean options to calculate the average beta value for each CpG site. We classified CpG sites based on their beta values into three categories: hypomethylated (beta value < 0.2), hypermethylated (beta value > 0.8), and heterogeneously methylated (0.2 ≤ beta value ≤ 0.8) (Du et al. 2010). For each transcript, we quantified the number of CpG sites based on probe detection. We then categorized the CpG sites according to their methylation status and calculated the proportion of

each category across tissues. We further utilized *in situ* Hi-C data specifically available for two tissues, the transverse colon and skeletal muscle, to support our findings. Genome compartment files for each tissue were downloaded and merged across donors using wiggletools mean (Zerbino et al. 2014). Next, we filtered transcripts specifically expressed in these two tissues and employed bigWigAverageOverBed (Kent et al. 2010) to compute the average signal intensity over the respective promoters of these transcripts.

CUT&Tag peaks for H4K16ac from the H9 embryonic stem cell line were obtained from an external study (GEO: GSE200768) (Pal et al. 2023). Peaks from the three replicates were merged, and the same filtering criteria were applied to retain only expressed transcripts in the H9 embryonic stem cell line. We then calculated the proportion of promoter bases covered by H4K16ac peaks for each protein-coding, lncRNA, and pseudogene transcript.

Alignment of the 1,000 bp upstream regions of the TSSs of pseudogenes and their parent genes

We ran PseudoPipe (Zhang et al. 2006) to infer parent–pseudogene relationships between pseudogenes and protein-coding genes (**Supplemental Table S4**). For each pseudogene and parent gene pair, we extracted the 1,000 bp upstream of the TSS and performed local pairwise alignment (Smith–Waterman) using the following scoring scheme: match = 2, mismatch = −3, gap opening = −5, and gap extension = −2. From the optimal alignment, we obtained the aligned coordinate blocks on the parent sequence. To avoid counting short, potentially spurious matches, we retained only aligned blocks with length ≥ 10 bp and summed their lengths to obtain the total valid aligned length on the parent sequence. Thus, inclusion was defined as the fraction of the upstream region of the parent gene covered by these valid aligned blocks, and identity was computed as the fraction of identical positions in the optimal alignment, defined as the number of identical matches divided by the total alignment length.

Based on inclusion and identity, we partitioned unprocessed pseudogenes into two classes according to whether they inherited the 1,000 bp region upstream of the TSS from their parent genes. Specifically, an unprocessed pseudogene was classified as inherited if at least 50% of its 1,000 bp upstream region aligned to the corresponding 1,000 bp upstream region of the parent gene with sequence identity >75%; all other cases were classified as not inherited.

Characterizing enriched motifs within promoter regions of transcripts

Motif enrichment analysis was performed using HOMER2 (Heinz et al. 2010) with the findMotifsGenome.pl script, employing the options -size 200 -S 10 -bg. Promoter regions of processed pseudogenes were selected across tissues and compared to those of unprocessed pseudogenes and protein-coding transcripts to identify enriched known and *de novo* motifs. We reported motif enrichment *p*-values as reported by HOMER2. For *de novo* motifs, the similarity score between each identified motif and its closest match in the reference database (e.g., HOMER and JASPAR) was also reported. Because multiple transcripts from the same gene can share an identical TSS, we retained a non-redundant set of promoter regions. In the primary analyses, we restricted to loci passing a RAMPAGE peak filter to define high-confidence transcription initiation. For motif enrichment analysis, we included promoter sequences from 761 processed pseudogenes, 471 unprocessed pseudogenes, and 22,156 protein-coding transcripts irrespective of RAMPAGE peak presence. Importantly, the chromatin signature of processed pseudogenes—lack of canonical active histone marks and promoter accessibility—was consistent in both RAMPAGE-positive and RAMPAGE-negative sets.

Quantification and statistical analysis

All analyses and statistical tests in this study were conducted using Python (version 3.9.13) and R (version 4.2.0) (R Core Team 2022), as detailed in the Methods and figure legends. Raw

exact *p*-values for all statistical comparisons are reported in the figure legends and false discovery rate (FDR)-adjusted *p*-values are shown in the figures. Unless otherwise specified, plots were generated using the ggplot2 (Wickham 2016) package in R. All box plots depict the first and third quartiles as the lower and upper bounds of the box, with a central band showing the median value and whiskers representing 1.5x the interquartile range.

Code availability

We have made our catalog of promoters with epigenetic and functional annotation available via Zenodo (https://doi.org/10.5281/zenodo.14934024). Custom scripts and selected intermediate files used in this study are also available on GitHub (https://github.com/gersteinlab/epiPgene) and as Supplemental Code.

ACKNOWLEDGEMENT

We thank Dr. Cristina Sisu (Brunel University London, UK) for valuable feedback on the analyses described in this manuscript. This work was supported by the Albert L. Williams Professorship funds and the National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) under award number U24HG007234. B.B. has been a Serra-Húnter Fellow from Generalitat de Catalunya (Spain) since 2025.

AUTHOR CONTRIBUTIONS

Y.J., B.B., and M.G. conceived the project and designed the study. Y.J. performed the computational analyses. Y.J. and B.B. wrote the original draft of the manuscript, which M.G. subsequently revised.

COMPETING INTEREST STATEMENT

The authors declare no competing interests.

REFERENCES

Athanikar JN, Badge RM, Moran JV. 2004. A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Res* **32**: 3846–3855.

Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**: 823–837.

Batut P, Gingeras TR. 2013. RAMPAGE: promoter activity profiling by paired-end sequencing of 5'-complete cDNAs. *Curr Protoc Mol Biol* **104**: Unit 25B.11.

Beck CR, Garcia-Perez JL, Badge RM, Moran JV. 2011. LINE-1 Elements in Structural Variation and Disease. *Annu Rev Genomics Hum Genet* **12**: 187–215.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**: 315–326.

Carbonell-Sala S, Perteghella T, Lagarde J, Nishiyori H, Palumbo E, Arnan C, Takahashi H, Carninci P, Uszczynska-Ratajczak B, Guigó R. 2024. CapTrap-seq: a platform-agnostic and quantitative approach for high-fidelity full-length RNA sequencing. *Nat Commun* **15**: 5278.

Chan ET, Quon GT, Chua G, Babak T, Trochesset M, Zirngibl RA, Aubin J, Ratcliffe MJH, Wilde A, Brudno M, et al. 2009. Conservation of core gene expression in vertebrate tissues. *J Biol* **8**: 33.

Cheetham SW, Faulkner GJ, Dinger ME. 2020. Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nat Rev Genet* **21**: 191–201.

Cochran K, Yin M, Mantripragada A, Schreiber J, Marinov GK, Kundaje A. 2024. Dissecting the cis-regulatory syntax of transcription initiation with deep learning. 2024.05.28.596138. https://www.biorxiv.org/content/10.1101/2024.05.28.596138v1 (Accessed October 14, 2024).

D'Errico I, Gadaleta G, Saccone C. 2004. Pseudogenes in metazoa: origin and features. *Brief Funct Genomic Proteomic* **3**: 157–167.

Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, Lin SM. 2010. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**: 587.

Dudnyk K, Cai D, Shi C, Xu J, Zhou J. 2024. Sequence basis of transcription initiation in human genome. *Science* **384**: eadj0116.

Ekram MB, Kang K, Kim H, Kim J. 2012. Retrotransposons as a major source of epigenetic variations in the mammalian genome. *Epigenetics* **7**: 370–382.

Frankish A, Carbonell-Sala S, Diekhans M, Jungreis I, Loveland JE, Mudge JM, Sisu C, Wright JC, Arnan C, Barnes I, et al. 2023. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res* **51**: D942–D949.

Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**: D766–D773.

Frigola J, Sabarinathan R, Mularoni L, Muiños F, Gonzalez-Perez A, López-Bigas N. 2017. Reduced mutation rate in exons due to differential mismatch repair. *Nat Genet* **49**: 1684–1692.

Gao Z, Wang M, Smith A, Boyes J. 2023. YY1 Binding to Regulatory Elements That Lack Enhancer Activity Promotes Locus Folding and Gene Activation. *Journal of Molecular Biology* **435**: 168315.

Gnanakkan VP, Jaffe AE, Dai L, Fu J, Wheelan SJ, Levitsky HI, Boeke JD, Burns KH. 2013. TE-array—a high throughput tool to study transposon transcription. *BMC Genomics* **14**: 869.

Han L, Yuan Y, Zheng S, Yang Y, Li J, Edgerton ME, Diao L, Xu Y, Verhaak RGW, Liang H. 2014. The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nat Commun* **5**: 3963.

Han YJ, Ma SF, Yourek G, Park Y-D, Garcia JGN. 2011. A transcribed pseudogene of MYLK promotes cell proliferation. *The FASEB Journal* **25**: 2305.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760–1774.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**: 311–318.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**: 576–589.

Hitz BC, Jin-Wook L, Jolanki O, Kagda MS, Graham K, Sud P, Gabdank I, Strattan JS, Sloan CA, Dreszer T, et al. 2023. The ENCODE Uniform Analysis Pipelines. *bioRxiv* 2023.04.04.535623.

Hon GC, Hawkins RD, Ren B. 2009. Predictive chromatin signatures in the mammalian genome. *Hum Mol Genet* **18**: R195–R201.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**: 2204–2207.

Kim JD, Faulk C, Kim J. 2007. Retroposition and evolution of the DNA-binding motifs of YY1, YY2 and REX1. *Nucleic Acids Res* **35**: 3442–3452.

Koenig Z, Yohannes MT, Nkambule LL, Zhao X, Goodrich JK, Kim HA, Wilson MW, Tiao G, Hao SP, Sahakian N, et al. 2024. A harmonized public resource of deeply sequenced diverse human genomes. *Genome Res* **34**: 796–809.

Kozomara A, Birgaoanu M, Griffiths-Jones S. 2019. miRBase: from microRNA sequences to function. *Nucleic Acids Research* **47**: D155–D162.

Kramerov DA, Vassetzky NS. 2011. Origin and evolution of SINEs in eukaryotic genomes. *Heredity* **107**: 487–495.

Lachner M, O'Carroll D, Rea S, Mechtler K, Jenuwein T. 2001. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**: 116–120.

Liao X, Zhu W, Zhou J, Li H, Xu X, Zhang B, Gao X. 2023. Repetitive DNA sequence detection and its role in the human genome. *Commun Biol* **6**: 954.

Liu Q, Liu X, Zhao D, Ruan X, Su R, Shang X, Wang D, Yang C, Xue Y. 2021. Pseudogene ACTBP2 increases blood-brain barrier permeability by promoting KHDRBS2 transcription through recruitment of KMT2D/WDR5 in Aβ1-42 microenvironment. *Cell Death Discov* **7**: 142.

Ludwig N, Leidinger P, Becker K, Backes C, Fehlmann T, Pallasch C, Rheinheimer S, Meder B, Stähler C, Meese E, et al. 2016. Distribution of miRNA expression across human tissues. *Nucleic Acids Research* **44**: 3865–3877.

Mighell AJ, Smith NR, Robinson PA, Markham AF. 2000. Vertebrate pseudogenes. *FEBS Letters* **468**: 109–114.

Moore JE, Pratt HE, Fan K, Phalke N, Fisher J, Elhajjajy SI, Andrews G, Gao M, Shedd N, Fu Y, et al. 2026. An expanded registry of candidate cis-regulatory elements. *Nature* 1–10.

Moore JE, Zhang X-O, Elhajjajy SI, Fan K, Pratt HE, Reese F, Mortazavi A, Weng Z. 2022. Integration of high-resolution promoter profiling assays reveals novel, cell type–specific transcription start sites across 115 human cell and tissue types. *Genome Res* **32**: 389–402.

Ninova M, Fejes Tóth K, Aravin AA. 2019. The control of gene expression and cell identity by H3K9 trimethylation. *Development* **146**: dev181180.

Pal D, Patel M, Boulet F, Sundarraj J, Grant OA, Branco MR, Basu S, Santos SDM, Zabet NR, Scaffidi P, et al. 2023. H4K16ac activates the transcription of transposable elements and contributes to their cis-regulatory function. *Nat Struct Mol Biol* **30**: 935–947.

Park SG, Hannenhalli S, Choi SS. 2014. Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals. *BMC Genomics* **15**: 526.

Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M, et al. 2012. The GENCODE pseudogene resource. *Genome Biology* **13**: R51.

Pervouchine DD, Djebali S, Breschi A, Davis CA, Barja PP, Dobin A, Tanzer A, Lagarde J, Zaleski C, See L-H, et al. 2015. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. *Nat Commun* **6**: 5903.

Picault N, Chaparro C, Piegu B, Stenger W, Formey D, Llauro C, Descombin J, Sabot F, Lasserre E, Meynard D, et al. 2009. Identification of an active LTR retrotransposon in rice. *Plant J* **58**: 754–765.

Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. 2010. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**: 1033–1038.

Pradeepa MM, Grimes GR, Kumar Y, Olley G, Taylor GCA, Schneider R, Bickmore WA. 2016. Histone H3 globular domain acetylation identifies a new class of enhancers. *Nature Genetics* **48**: 681–686.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **35**: D61–D65.

Qi Y, Wang X, Li W, Chen D, Meng H, An S. 2021. Pseudogenes in Cardiovascular Disease. *Front Mol Biosci* **7**: 622540.

Qian SH, Chen L, Xiong Y-L, Chen Z-X. 2022. Evolution and function of developmentally dynamic pseudogenes in mammals. *Genome Biol* **23**: 235.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

R Core Team. 2022. R: A language and environment for statistical computing. https://www.R-project.org/.

Reese F, Williams B, Balderrama-Gutierrez G, Wyman D, Çelik MH, Rebboah E, Rezaie N, Trout D, Razavi-Mohseni M, Jiang Y, et al. 2023. The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. *bioRxiv* 2023.05.15.540865.

Reichmann J, Crichton JH, Madej MJ, Taggart M, Gautier P, Garcia-Perez JL, Meehan RR, Adams IR. 2012. Microarray analysis of LTR retrotransposon silencing identifies Hdac1 as a regulator of retrotransposon expression in mouse embryonic stem cells. *PLoS Comput Biol* **8**: e1002486.

Rozowsky J, Gao J, Borsari B, Yang YT, Galeev T, Gürsoy G, Epstein CB, Xiong K, Xu J, Li T, et al. 2023. The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models. *Cell* **186**: 1493-1511.e40.

Saha K, Nielsen GI, Nandani R, Zhang Y, Kong L, Ye P, An W. 2024. YY1 is a transcriptional activator of the mouse LINE-1 Tf subfamily. *Nucleic Acids Res* **52**: 12878–12894.

Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA. 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* **8**: 424–436.

Singh RK, Singh D, Yadava A, Srivastava AK. 2020. Molecular fossils “pseudogenes” as functional signature in biological system. *Genes Genom* **42**: 619–630.

Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM, et al. 2008. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**: 534–538.

Tian X, Song J, Zhang X, Yan M, Wang S, Wang Y, Xu L, Zhao L, Wei J-J, Shao C, et al. 2020. MYC-regulated pseudogene HMGA1P6 promotes ovarian cancer malignancy via augmenting the oncogenic HMGA1/2. *Cell Death Dis* **11**: 167.

Troskie R-L, Jafrani Y, Mercer TR, Ewing AD, Faulkner GJ, Cheetham SW. 2021. Long-read cDNA sequencing identifies functional pseudogenes in the human transcriptome. *Genome Biol* **22**: 146.

Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigó R, Johnson R. 2018. Towards a complete map of the human long non-coding RNA transcriptome. *Nat Rev Genet* **19**: 535–548.

Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ, Cohen MA, Nabet B, Buckley DL, et al. 2017. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **171**: 1573-1588.e28.

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wolfe JC, Mikheeva LA, Hagras H, Zabet NR. 2021. An explainable artificial intelligence approach for decoding the enhancer histone modifications code and identification of novel enhancers in Drosophila. *Genome Biol* **22**: 308.

Wright BW, Molloy MP, Jaschke PR. 2022. Overlapping genes in natural and engineered genomes. *Nat Rev Genet* **23**: 154–168.

Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, et al. 2020. Ensembl 2020. *Nucleic Acids Research* **48**: D682.

Zerbino DR, Johnson N, Juettemann T, Wilder SP, Flicek P. 2014. WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. *Bioinformatics* **30**: 1008–1009.

Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, Gerstein M. 2006. PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**: 1437–1439.

Zheng H, Brennan K, Hernaez M, Gevaert O. 2019. Benchmark of long non-coding RNA quantification for RNA sequencing of cancer samples. *GigaScience* **8**: giz145.

FIGURE LEGENDS

Fig. 1. Building a catalog of epigenetically and functionally annotated promoters in the human genome. (A) Schematic representation of different pseudogene biotypes based on their mechanisms of origin, depicted in distinct colors: processed pseudogenes arising from retrotransposition events (left), unitary pseudogenes resulting from mutations in functional protein-coding genes leading to a fixed loss of function in a given species (middle), and unprocessed pseudogenes originating from gene duplication followed by functional decay (right). **(B)** Overview of the study workflow. A catalog of active and repressed promoters was constructed across 26 human tissues using tissue-specific epigenome data. Transcripts were subsequently filtered based on three criteria: (i) annotation, ensuring no overlapping promoters; (ii) expression, requiring an average TPM ≥ 1 across donors and replicates; and (iii) RAMPAGE support for promoters. **(C)** For each gene biotype, the number of expressed transcripts (*y*-axis) detected across an increasing number of tissues (*x*-axis), after applying the cross-tissue expression filtering step in (B). **(D)** Proportion of promoters with RAMPAGE support, categorized by gene biotypes, counting each transcript–tissue pair separately. “Protein-cod.”: protein-coding. **(E)** Analogous to (D), but categorized by pseudogene biotypes. Panels (A) and (B) were created in BioRender. Gerstein, M. (2026) https://BioRender.com/2agdcbr.

Fig. 2. Epigenetic and transcript expression characterization of transcribed promoters. (A) Average proportion of promoter bases (*y*-axis) overlapping H3K27ac, H3K4me3, ATAC-seq, and DNase-seq peaks for protein-coding, lncRNA, and pseudogene transcripts (*x*-axis). Proportions were computed per transcript in each tissue and then aggregated across tissues within each gene biotype. Error bars indicate standard deviations across transcripts and tissues. “Protein-cod.”: protein-coding; “PseudoG.”: pseudogene. **(B)** Same as (A), but shown for the three pseudogene biotypes. Error bars indicate standard deviations across transcripts and tissues. **(C)** Percent stacked bar plot showing, for each gene biotype, the proportion of promoter

CpG sites at different methylation levels across tissues. “Protein-cod.”: protein-coding. **(D)** Ridgeline plot showing, for each gene biotype, the distribution of promoters located within active (> 0; “A”) and inactive (< 0; “B”) genome compartments in the transverse colon (top) and skeletal muscle (bottom). Unitary pseudogenes are excluded due to their small number of expressed transcripts in these tissues. Vertical lines indicate the mean and one standard deviation above and below the mean. Processed pseudogenes are used as the reference group for two-sided Wilcoxon rank-sum tests. ***, FDR-adjusted *p*-value < 0.001; **, FDR-adjusted *p*-value < 0.01; n.s., not significant. Transverse colon: processed vs. protein-coding ($p = 1.2\times10^{-7}$); processed vs. lncRNA ($p = 0.31$); processed vs. unprocessed ($p = 0.15$). Skeletal muscle: processed vs. protein-coding ($p = 1.2\times10^{-16}$); processed vs. lncRNA ($p = 3.6\times10^{-6}$); processed vs. unprocessed ($p = 8.0\times10^{-3}$). **(E)** Violin plot showing transcript expression levels across gene biotypes. For each biotype, we report the number of transcripts included in the analysis at the bottom of the plot (see also Table 2). Processed pseudogenes are used as the reference group for two-sided Wilcoxon rank-sum tests. Transcript counts per biotype across tissues are shown below. ***, FDR-adjusted *p*-value < 0.001; **, FDR-adjusted *p*-value < 0.01; n.s., not significant. Processed vs. protein-coding ($p = 0.95$); processed vs. lncRNA ($p = 5.3\times10^{-31}$); processed vs. unprocessed ($p = 5.9\times10^{-32}$); processed vs. unitary ($p = 6.0\times10^{-3}$). “Protein-cod.”: protein-coding. **(F)** Box plot showing the Spearman’s co-expression correlation coefficient of processed and unprocessed pseudogenes with their parent genes. ***, $p < 0.001$. Processed vs. unprocessed ($p = 4.1\times10^{-12}$).

Fig. 3. Sequence and regulatory features of regions upstream of the TSS in pseudogenes and their parent genes. **(A)** Sequence inclusion and identity for the 1,000 bp region upstream of the TSS in processed and unprocessed pseudogenes, compared with their parent genes. Marginal distributions are shown as box plots alongside. **(B)** Average proportion of promoter bases overlapping ATAC-seq, DNase-seq, H3K27ac, and H3K4me3 peaks for unprocessed

pseudogenes, stratified by whether they inherited the 1,000 bp region upstream of the TSS from their parent genes. Error bars represent standard deviations across transcripts and tissues. **(C)** GC content of the 1,000 bp region upstream of the TSS across gene biotypes. Processed pseudogenes were used as the reference group for two-sided Wilcoxon rank-sum tests. ***, FDR-adjusted *p*-value < 0.001. Processed vs. protein-coding ($p = 2.6\times10^{-28}$); processed vs. lncRNA ($p = 1.6\times10^{-8}$); processed vs. unprocessed ($p = 2.2\times10^{-19}$); processed vs. unitary ($p = 8.0\times10^{-4}$). **(D)** Conservation of the 1,000 bp region upstream of the TSS across gene biotypes, measured by phastCons 20-way scores (shown in logarithmic scale). Unitary pseudogenes were excluded due to the small sample size. Processed pseudogenes were used as the reference group for two-sided Wilcoxon rank-sum tests. ***, FDR-adjusted *p*-value < 0.001; n.s., not significant. Processed vs. protein-coding ($p = 0.54$); processed vs. lncRNA ($p = 0.38$); processed vs. unprocessed ($p = 5.0\times10^{-6}$). **(E)** Proportion of variants across different MAF categories within the 1,000 bp region upstream of the TSS across gene biotypes. “Protein-cod.”: protein-coding. **(F)** Phred-scaled CADD scores for variants within the 1,000 bp region upstream of the TSS across gene biotypes. Due to the large sample size, only 0.5% of variants classified as outliers were randomly selected for visualization. Processed pseudogenes were used as the reference group for two-sided Wilcoxon rank-sum tests. ***, FDR-adjusted *p*-value < 0.001; n.s., not significant. Processed vs. protein-coding ($p = 0$); processed vs. lncRNA ($p = 1.9\times10^{-10}$); processed vs. unprocessed ($p = 7.4\times10^{-138}$); processed vs. unitary ($p = 0.20$).

Fig. 4. Transposable element associations and promoter chromatin features of processed and unprocessed pseudogenes. (A) UpSet plot showing the number of lncRNA and pseudogene promoters overlapping different classes of transposable elements and repetitive elements. Intersection groups with small counts were excluded ($n < 10$). **(B)** Jaccard similarity between promoters of different gene biotypes and LINEs (top) and SINEs (bottom). Processed pseudogenes were used as the reference group for two-sided Wilcoxon rank-sum

tests. ***, FDR-adjusted *p*-value < 0.001; **, FDR-adjusted *p*-value < 0.01; *, FDR-adjusted *p*-value < 0.05; n.s., not significant. LINE: processed vs. protein-coding ($p = 3.0\times10^{-9}$); processed vs. lncRNA ($p = 8.6\times10^{-5}$); processed vs. unprocessed ($p = 0.05$). SINE: processed vs. protein-coding ($p = 0.004$); processed vs. lncRNA ($p = 0.11$); processed vs. unprocessed ($p = 0.60$). **(C)** Proportions of promoters overlapping different classes of transposable elements (LINE, SINE, LTR, DNA, and simple repeat) for all parent genes and all protein-coding genes (left), processed pseudogenes and their parent genes (middle), and unprocessed pseudogenes and their parent genes (right). **(D)** Epigenetic characterization of processed and unprocessed pseudogene promoters, comparing promoters that overlap annotated LINEs or SINEs (left) with those without any overlap (right). Proportions were calculated for each transcript in each tissue and then aggregated across tissues for processed and unprocessed pseudogenes. Error bars represent standard deviations across transcripts and tissues.

Fig. 5. Potential factors regulating the expression of processed pseudogenes. (A) Bar plots showing the proportion of promoters containing human- and mouse-derived YY1-like motifs enriched in processed pseudogenes, compared with promoters of protein-coding transcripts, lncRNAs, and unprocessed pseudogenes. **(B)** Bar plots showing the proportion of parent genes of processed pseudogenes whose promoters harbor human- and mouse-derived YY1-like motifs. **(C)** Density plot showing the positional distribution of YY1-like motifs relative to TSSs of processed pseudogenes and protein-coding transcripts. The *p*-values were calculated using Fisher's exact test by comparing the proportion of motif occurrences falling within ±10 bp of the TSS between processed pseudogenes and protein-coding transcripts. **(D)** Pie chart showing the proportion of expressed pseudogenes from different biotypes that contact genomic regions of varying lengths across tissues. **(E)** Similar to (D), percent stacked bar plots showing the proportion of expressed protein-coding and lncRNA transcripts that contact genomic regions of varying lengths across tissues. **(F)** Violin plot showing the distribution of observed contact

frequencies between transcripts from different gene biotypes and other genomic regions. Within each violin, the white point represents the mean, and the vertical line denotes the range of one standard deviation above and below the mean. Processed pseudogenes were used as the reference group for two-sided Wilcoxon rank-sum tests. ***, FDR-adjusted *p*-value < 0.001; n.s., not significant. Processed vs. protein-coding ($p = 3.1\times10^{-6}$); processed vs. lncRNA ($p = 3.3\times10^{-4}$); processed vs. unprocessed ($p = 1.1\times10^{-7}$); processed vs. unitary ($p = 0.11$). **(G)** Bar plot showing the average proportion of bases in contacted genomic regions that overlap annotated cCREs, stratified by gene biotype. Error bars indicate standard errors. dELS, distal enhancer-like signature; pELS, proximal enhancer-like signature; PLS, promoter-like signature; TF, transcription factor binding site.

Figure 1

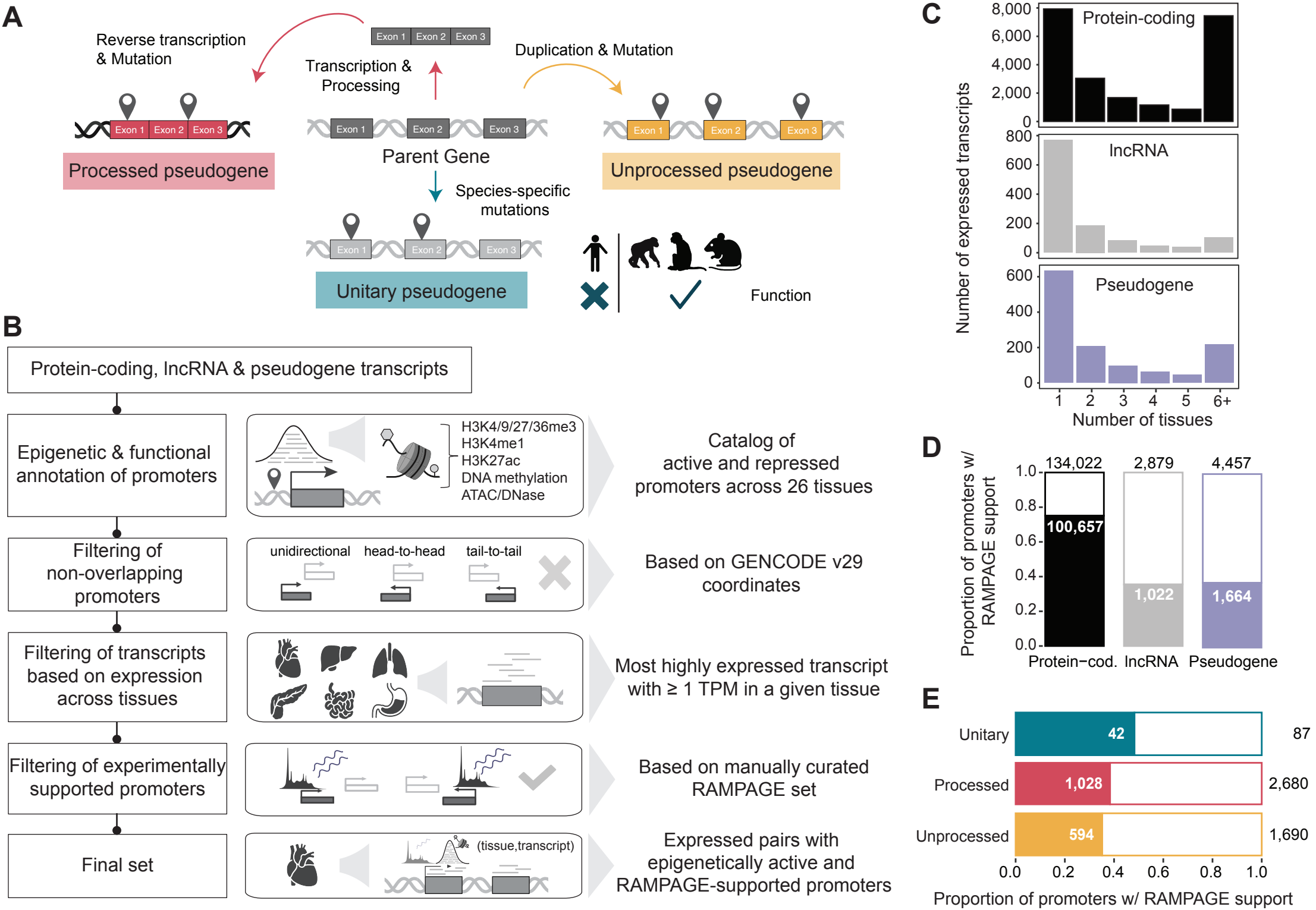
A
Reverse transcription & Mutation
Transcription & Processing
Duplication & Mutation
Exon 1
Exon 2
Exon 3
Processed pseudogene
Parent Gene
Unprocessed pseudogene
Species-specific mutations
Unitary pseudogene
Function
B
Protein-coding, lncRNA & pseudogene transcripts
Epigenetic & functional annotation of promoters
H3K4/9/27/36me3
H3K4me1
H3K27ac
DNA methylation
ATAC/DNase
Catalog of active and repressed promoters across 26 tissues
Filtering of non-overlapping promoters
unidirectional
head-to-head
tail-to-tail
Based on GENCODE v29 coordinates
Filtering of transcripts based on expression across tissues
Most highly expressed transcript with ≥ 1 TPM in a given tissue
Filtering of experimentally supported promoters
Based on manually curated RAMPAGE set
Final set
(tissue,transcript)
Expressed pairs with epigenetically active and RAMPAGE-supported promoters
C
Protein-coding
lncRNA
Pseudogene
Number of expressed transcripts
8,000
6,000
4,000
2,000
0
800
600
400
200
600
400
200
1
2
3
4
5
6+
Number of tissues
D
Proportion of promoters w/ RAMPAGE support
134,022
2,879
4,457
100,657
1,022
1,664
1.0
0.8
0.6
0.4
0.2
0.0
Protein−cod.
lncRNA
Pseudogene
E
Unitary
42
87
Processed
1,028
2,680
Unprocessed
594
1,690
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of promoters w/ RAMPAGE support

Figure 2

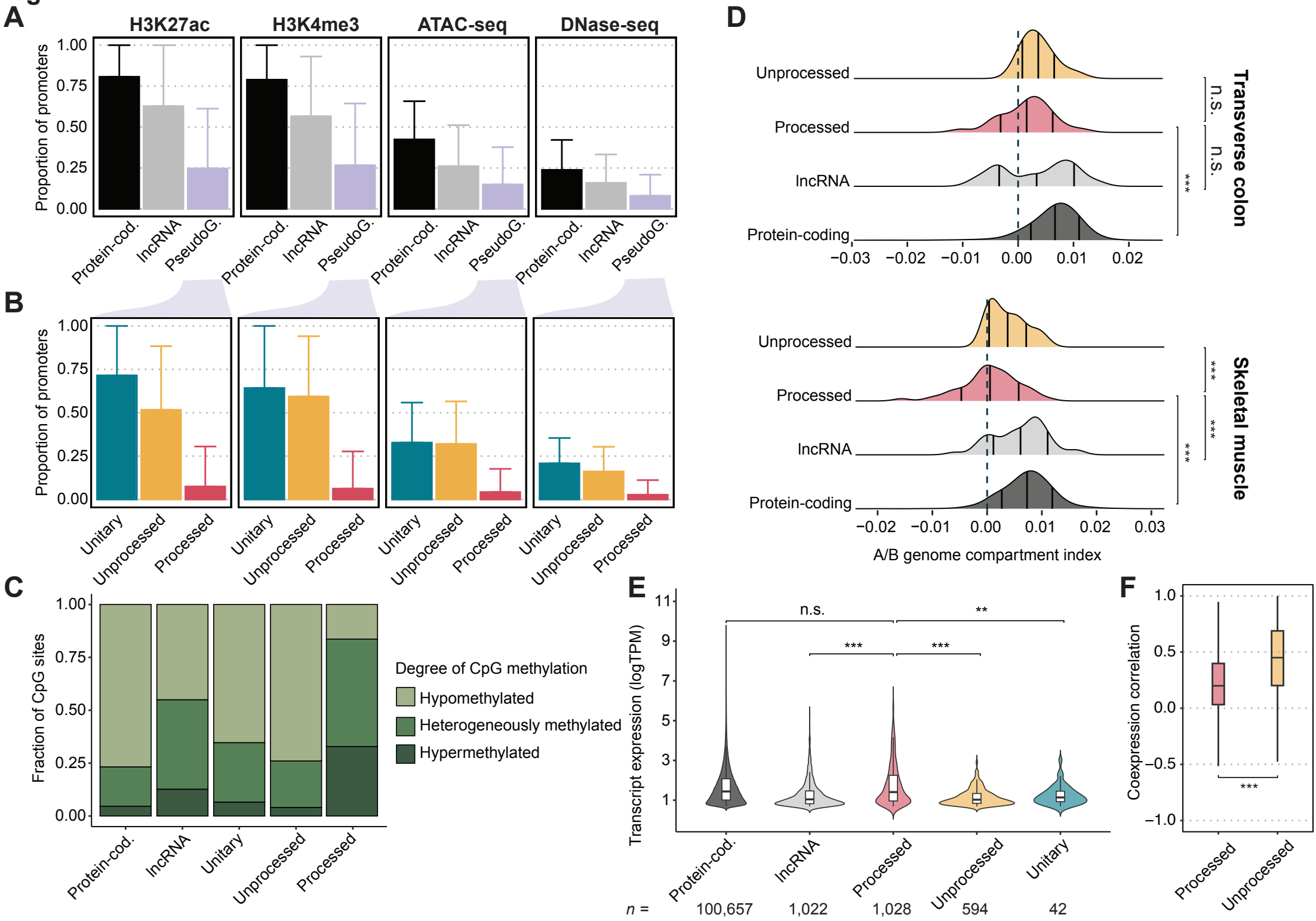
A
H3K27ac
H3K4me3
ATAC-seq
DNase-seq
Proportion of promoters
1.00
0.75
0.50
0.25
0.00
Protein-cod.
lncRNA
PseudoG.
B
Proportion of promoters
Unitary
Unprocessed
Processed
C
Fraction of CpG sites
Degree of CpG methylation
Hypomethylated
Heterogeneously methylated
Hypermethylated
Protein-cod.
lncRNA
Unitary
Unprocessed
Processed
D
Unprocessed
Processed
lncRNA
Protein-coding
n.s.
***
Transverse colon
Skeletal muscle
−0.03
−0.02
−0.01
0.00
0.01
0.02
0.03
A/B genome compartment index
E
Transcript expression (logTPM)
n.s.
**
***
11
9
7
5
3
1
n = 100,657 1,022 1,028 594 42
F
Coexpression correlation
1.0
0.5
0.0
−0.5
−1.0
Processed
Unprocessed

Figure 3

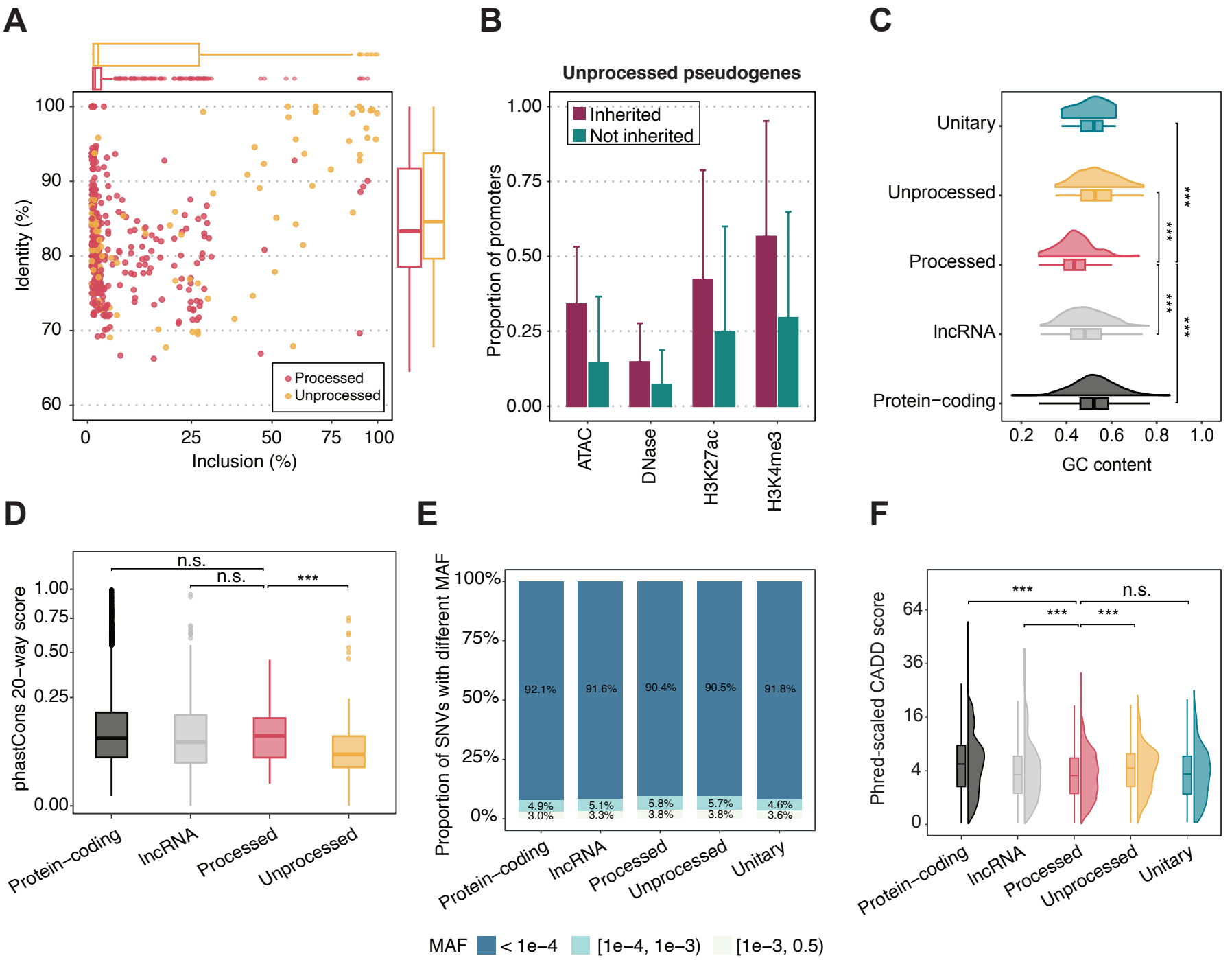
A
Identity (%)
Inclusion (%)
Processed
Unprocessed
B
Unprocessed pseudogenes
Inherited
Not inherited
Proportion of promoters
ATAC
DNase
H3K27ac
H3K4me3
C
Unitary
Unprocessed
Processed
lncRNA
Protein−coding
GC content
D
phastCons 20−way score
n.s.
n.s.
***
Protein−coding
lncRNA
Processed
Unprocessed
E
Proportion of SNVs with different MAF
92.1%
91.6%
90.4%
90.5%
91.8%
4.9%
3.0%
5.1%
3.3%
5.8%
3.8%
5.7%
3.8%
4.6%
3.6%
Protein−coding
lncRNA
Processed
Unprocessed
Unitary
MAF
< 1e−4
[1e−4, 1e−3)
[1e−3, 0.5)
F
Phred-scaled CADD score
***
n.s.
Protein−coding
lncRNA
Processed
Unprocessed
Unitary

Figure 4

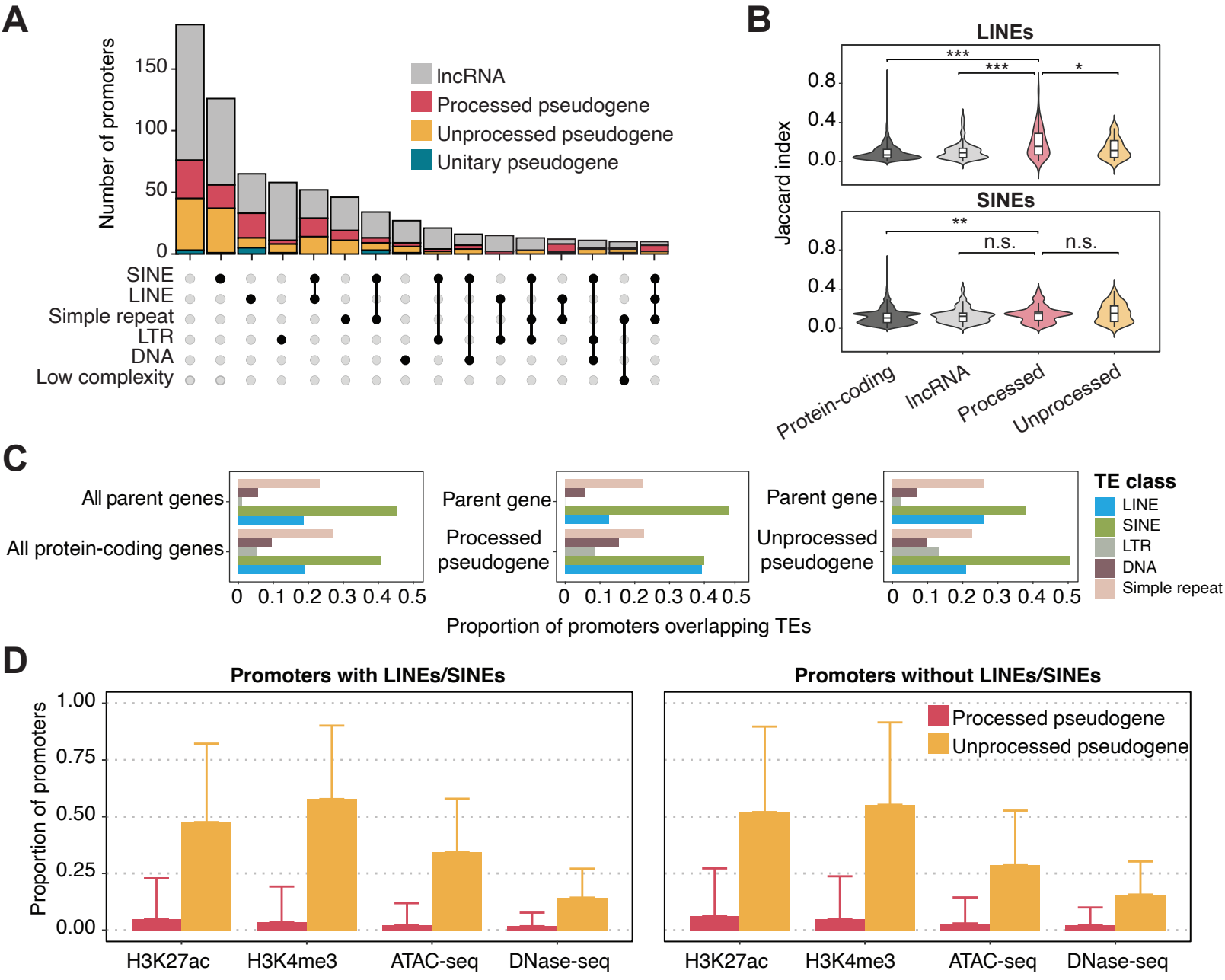
A
Number of promoters
150
100
50
0
lncRNA
Processed pseudogene
Unprocessed pseudogene
Unitary pseudogene
SINE
LINE
Simple repeat
LTR
DNA
Low complexity
B
LINEs
SINEs
Jaccard index
0.8
0.4
0.0
***
***
*
**
n.s.
n.s.
Protein-coding
lncRNA
Processed
Unprocessed
C
All parent genes
All protein-coding genes
Parent gene
Processed pseudogene
Unprocessed pseudogene
0 0.1 0.2 0.3 0.4 0.5
Proportion of promoters overlapping TEs
TE class
LINE
SINE
LTR
DNA
Simple repeat
D
Promoters with LINEs/SINEs
Promoters without LINEs/SINEs
Proportion of promoters
1.00
0.75
0.50
0.25
0.00
Processed pseudogene
Unprocessed pseudogene
H3K27ac
H3K4me3
ATAC-seq
DNase-seq

Figure 5

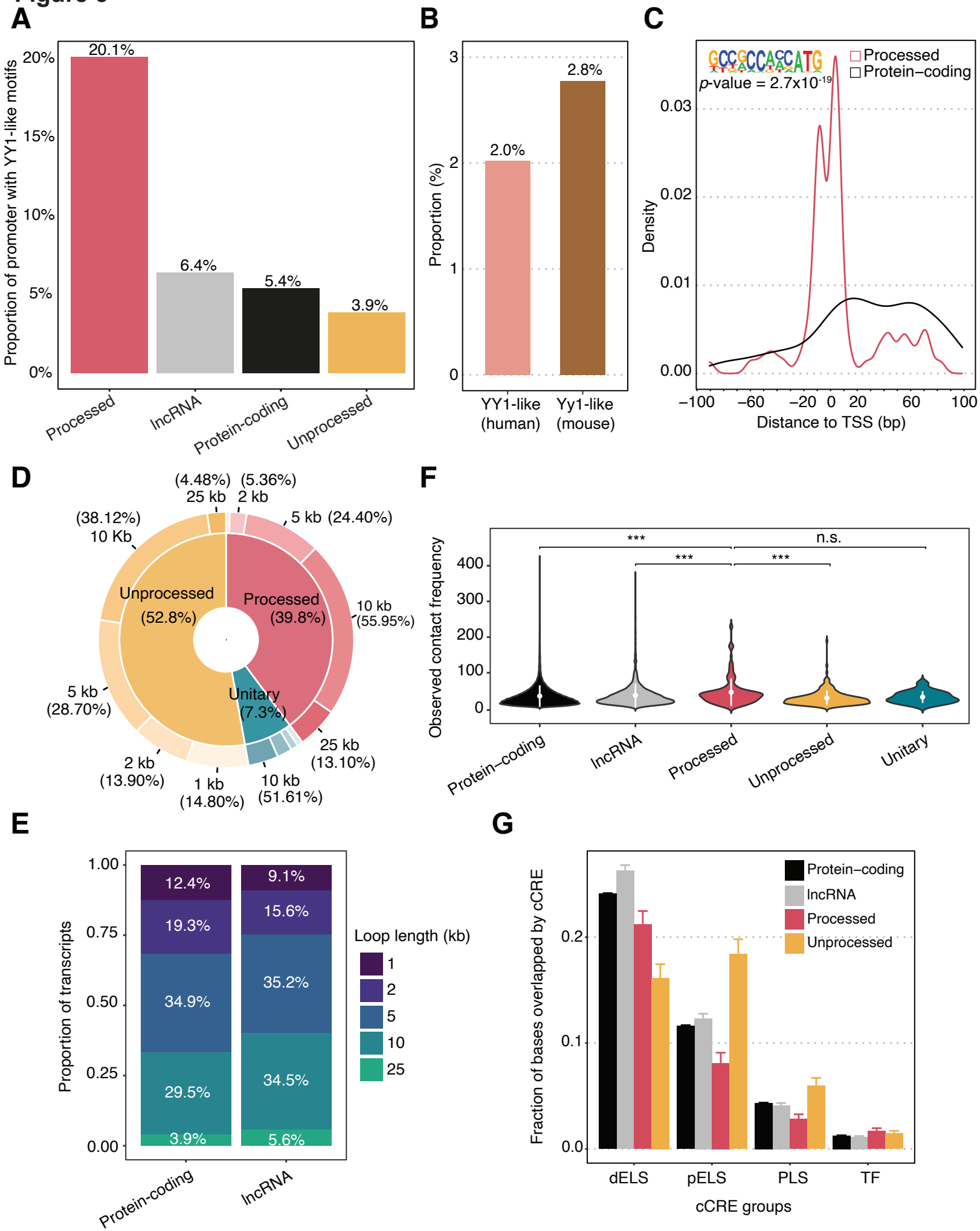

Table 1. Gene biotypes included in this study based on GENCODE v29 annotation

Category	GENCODE v29 gene biotype terms
Protein-coding	protein_coding
lncRNA	3prime_overlapping_ncrna, antisense, bidirectional_promoter_lncRNA, macro_lncRNA, non_coding, processed_transcript, sense_intronic, sense_overlapping, lincRNA
Pseudogene	processed_pseudogene, unprocessed_pseudogene, unitary_pseudogene, transcribed_processed_pseudogene, transcribed_unprocessed_pseudogene, transcribed_unitary_pseudogene, translated_processed_pseudogene, translated_unprocessed_pseudogene

Table 2. Gene and transcript counts across filtering steps and biotypes. GENCODE annotations and RAMPAGE data are tissue-agnostic (rows 1–3 and 6), whereas expression profiles (rows 4–5) are tissue-specific. For the latter, we report both the total number of transcripts identified across 26 tissues (left, underlined) and the number of nonredundant (unique) transcripts identified across tissues (right, bold).

	Pseudogene								Protein-coding		lncRNA	
	Processed		Unprocessed		Unitary		Total					
GENCODE v29 annotated genes	10,672		3,531		219		14,422		19,922		14,889	
GENCODE v29 annotated transcripts	11,255		5,503		729		17,487		151,113		28,149	
Transcripts with non-overlapping promoters	9,501		4,232		520		14,253		106,643		17,624	
Expressed transcripts across tissues	2,780	**778**	2,363	**617**	157	**62**	5,300	**1,511**	259,084	**43,386**	3,521	**1,449**
Most highly expressed transcript per gene	2,680	**761**	1,690	**471**	87	**35**	4,457	**1,267**	134,022	**22,156**	2,879	**1,231**
RAMPAGE-supported transcripts	1,028	**151**	594	**167**	42	**17**	1,664	**335**	100,657	**14,315**	1,022	**490**