RESEARCH ARTICLE SUMMARY

MOLECULAR BIOLOGY

mRNA initiation and termination are spatially coordinated

Ezequiel Calvo-Roitberg†, Christine L. Carroll†, GyeungYun Kim, Valeria Sanabria, Sergey V. Venev, Steven T. Mick, Joseph D. Paquette, Maritere Uriostegui-Arcos, Job Dekker, Ana Fiszbein*‡, Athma A. Pai*‡

Full article and list of author affiliations:
https://doi.org/10.1126/science.ado8279

INTRODUCTION: Initiation and termination are key steps in the synthesis of mature mRNA molecules. Recent high-throughput analyses have suggested that most mRNA isoform diversity comes from the use of alternative initiation and termination sites. These choices can lead to variable protein conformations or different 5′ and 3′ untranslated regions that influence mRNA localization, translation efficiency, and stability. Thus far, these processing events have primarily been studied in isolation, with little insight into how decisions are coordinated across the transcript to govern the ultimate fate and function of mRNA molecules.

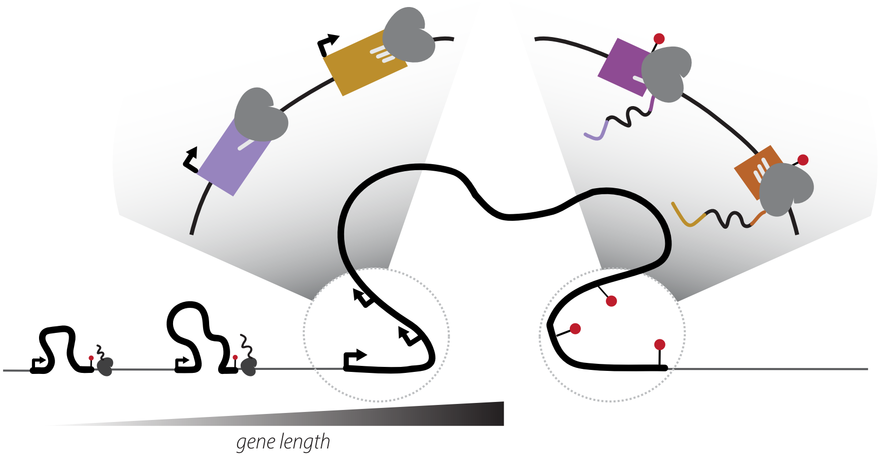
RATIONALE: Analyzing short-read RNA-sequencing (RNA-seq) data across human tissues, we observed that the numbers of alternative transcription start sites (TSSs) and alternative 3′ polyadenylation sites (PASs) are correlated across genes. This led us to hypothesize that genes may be structured to use these sites in a coordinated manner. We studied how alternative RNA processing decisions at the terminal ends of genes are coordinated, aiming to better understand the regulation of full-length RNA isoforms.

RESULTS: Through a systematic analysis of hundreds of long-read RNA-seq datasets across mammalian tissues, we find that mRNA initiation and termination site choice are directly coupled. Notably, this coupling is based on the order in which sites appear in the genome, with transcripts that start at an upstream TSS preferentially ending at an upstream PAS and those starting at a downstream TSS using a downstream PAS. This revealed a positional initiation termination axis (PITA), which governs coupled mRNA terminal end choices, independent of tissue- or context-specific regulation. Consistently, dCas9-CRISPR perturbations show that mRNA 5′ end choice directly influences mRNA 3′ end choice. PITA is strongly associated with the length of genomic features; PITA genes are longer and transcribed faster, have greater distances between alternative TSSs or PASs, and exhibit distinct chromatin features at TSSs. We see that the rate of RNA Polymerase II (RNAPII) elongation is dependent on where transcription begins, with downstream TSSs associated with faster RNAPII elongation. Overall, we found that PITA coupling depends on a combination of sustained RNAPII trafficking and progressively stronger alternative PAS sequences across longer genes. Specifically, slower RNAPII molecules originating at upstream TSSs are more likely to use weaker upstream PASs, whereas faster RNAPII molecules from downstream TSSs can reach stronger downstream PASs. Together, our data support a model integrating sequence and kinetic features to propose that full-length isoform expression is governed by RNAPII elongation rates within and across human genes.

CONCLUSION: This study reveals widespread ordinal coupling between alternative mRNA starts and ends across tissues and species. Our results suggest an interplay between mRNA processing and the evolution of gene architecture. By showing that PITA coupling is governed by transcription elongation dynamics, we extend the classical “window of opportunity” paradigm from splicing to full-length mRNA isoforms. This work lays the foundations for new areas of investigation into the spatial control of mRNA expression and co-transcriptional RNA processing. □

*Corresponding author. Email: anafisz@bu.edu (A.F.); athma.pai@umassmed.edu (A.A.P.)
†These authors contributed equally to this work. ‡These authors contributed equally to this work.

Cite this article as E. Calvo-Roitberg *et al.*, *Science* 390, eado8279 (2025). DOI: 10.1126/science.ado8279



Transcription elongation rates drive the positional coordination of mRNA initiation and termination. RNAPII molecules initiating at downstream TSSs elongate faster throughout the gene, coupling the usage of these sites with downstream, stronger PASs in longer genes.

MOLECULAR BIOLOGY

mRNA initiation and termination are spatially coordinated

Ezequiel Calvo-Roitberg[1]†, Christine L. Carroll[2]†, GyeungYun Kim[2], Valeria Sanabria[1], Sergey V. Venev[3], Steven T. Mick[2], Joseph D. Paquette[1], Maritere Uriostegui-Arcos[2], Job Dekker[3,4], Ana Fiszbein[2,5]*‡, Athma A. Pai[1]*‡

Transcriptional initiation and termination decisions drive messenger RNA (mRNA) isoform diversity but the relationship between them remains poorly understood. By systematically profiling joint usage of transcription start and end sites, we observed that mRNA using upstream starts preferentially use upstream end sites and that the usage of downstream sites is similarly coupled. Our results suggest a positional initiation termination axis (PITA), in which usage of alternative terminal sites are coupled based on their genomic order. PITA is enriched in longer genes with distinct chromatin features. We find that mRNA 5′ start choice directly influences 3′ ends depending on RNA polymerase II trafficking speed. Our results indicate that spatial organization and transcriptional dynamics couple transcription initiation and mRNA 3′ end decisions to define mRNA isoform expression.

Mechanisms involved in mRNA processing are regulated by disparate molecular machineries and subjected to different global regulatory constraints to drive tissue- or context-specific transcriptomes (1). Thus, exon usage within mRNAs has historically been studied as independent events that contribute individually to the composition of full-length isoforms (2, 3). However, increasing evidence suggests cooperative regulation by RNA processing mechanisms across a gene (4–7). Recent studies have found that factors involved in transcription, mRNA splicing, and cleavage and polyadenylation (CPA) may have distinct secondary roles in another step of RNA processing (8–12). These investigations have mostly focused on connections between splicing and either transcription initiation or CPA, with less known about direct connections between transcription initiation and CPA.

More than 70% of mammalian genes express alternative transcription start sites (TSSs) or polyadenylation sites (PASs) across cellular contexts (4, 13, 14), with four to five annotated TSSs and/or PASs per gene on average (15, 16). Alternative mRNA terminal end usage underlies most variation in isoform usage across cells, tissue types, and cellular contexts (4, 17–19). Alternative TSSs and PASs influence the composition and length of coding sequences and 5′ and 3′ untranslated regions (UTRs), respectively, with impacts on RNA stability, localization, and translation efficiency (20–22). Understanding whether and how these decisions are coordinated across the gene will inform the regulation of full-length isoforms, the proteins encoded by them, and effects on protein expression.

There are many shared regulatory features between the 5′ and 3′ ends of genes. For instance, RNA polymerase II (RNAPII) phosphorylation events required for promoter-proximal transcription elongation are also required for productive elongation at the end of the gene (9, 23, 24). RNAPII elongation rates can influence both alternative TSS and PAS usage based on the position of sites within genes (25–31). Decisions at the terminal ends of genes might also be spatially associated (32). Recent high-resolution measurement of chromatin conformation found dynamic loops anchored at TSSs and extruding through the body of the gene (33, 34), suggesting that both two- and three-dimensional (2D and 3D, respectively) architecture may be important for mRNA synthesis. These findings all hint at spatiotemporal connections between transcription initiation and CPA.

We set out to answer two fundamental questions in RNA biology: Are transcription initiation and termination coregulated, and which mechanisms drive coordinated RNA processing? Through an analysis of genome-wide datasets across tissues, we find that genes have similar numbers of alternative sites at both their 5′ and 3′ ends and that there is a global bias toward coordinated usage of alternative TSSs and PASs based on their ordinal position within genes. Our analyses reveal a positional initiation termination axis (PITA) that intrinsically governs mRNA terminal end choices, independent of tissue- or context-specific regulation. We show that this coupling between transcription start and end sites is mostly driven by the dynamics of transcription.

Results

The selection of mRNA terminal (5′ and 3′) ends is a highly regulated process that often involves a choice between multiple alternative sites and exons (Fig. 1A). To understand the extent to which mRNA isoform diversity in genes is driven by the usage of alternative terminal exons, we systematically characterized the landscape of terminal exon usage across human cell types. We previously developed the hybrid-internal-terminal (HIT) index, a metric that uses RNA-sequencing (RNA-seq) data to identify and quantify the relative usage of alternative terminal exons across the genome (35). We applied the HIT index framework to 17,350 RNA-seq samples across 54 human tissues from the genotype tissue expression (GTEx) project (36) and identified 66,673 and 63,399 annotated alternative first and last exons, respectively, across all samples. The distributions of the number of first and last exons per gene were similar: The number of genes with one, two, three, etc. first exons closely matched the number of genes with only one, two, three, etc. last exons, respectively (Fig. 1B). Genes with alternative first exons (AFEs) are significantly more likely than expected to also use alternative last exons (ALEs) (hypergeometric test P-value $< 2.2 \times 10^{-16}$; fig. S1A).

Coordinated terminal exon usage based on ordinal position

This led us to speculate that genes might utilize similar numbers of first and last exons to jointly regulate the usage of terminal ends. Among genes with multiple AFEs and ALEs, there is a strong correlation between the expression of first and last exons within a gene, such that genes expressing two AFEs are more likely to use two ALEs, genes using three AFEs often use three ALEs, and so on (Fig. 1C, inset). Although the strength of this effect varies across tissues, 97.8% of samples showed positive correlations (Fig. 1C and fig. S1B) between quantities of first and last exons identified despite tissue-specific regulation of alternative terminal end usage. The presence of similar numbers of alternative sites at both gene ends suggests the potential for regulatory coupling of transcription and CPA decisions based on gene organization, so we next asked whether the usage of first and last exons was associated with their position within the gene. We measured exon usage with a percent spliced-in (PSI) value (37), which estimates the proportion of mRNA transcripts that include a specific terminal exon. As a baseline, genes that predominantly use a single first exon (PSI > 0.95) also tend to use a single last exon (fig. S1C). More broadly, we found that the usage of terminal exons with similar ordinal positions was correlated across genes (Fig. 1D). Specifically, increased usage of the most upstream AFE (AFE 1) was associated with increased usage of the most upstream ALE

[1]RNA Therapeutics Institute, University of Massachusetts Chan Medical School, Worcester, MA, USA. [2]Department of Biology, Boston University, Boston, MA, USA. [3]Department of Systems Biology, University of Massachusetts Chan Medical School, Worcester, MA, USA. [4]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [5]Faculty of Computing & Data Sciences, Boston University, Boston, MA, USA. ***Corresponding authors. Email: anafisz@bu.edu (A.F.); athma.pai@umassmed.edu (A.A.P.)** †These authors contributed equally to this work. ‡These authors contributed equally to this work.

(ALE 1), AFE 2 usage positively correlated with ALE 2 usage, etc. Concordantly, we observed that usage of AFE 1 negatively correlated with usage of ALE 2, indicating a complex coupling relationship based on the ordinal positions of the terminal exons within genes. Restricting the analysis to genes that use exactly three AFEs and three ALEs confirmed this ordinal coupling pattern (Fig. 1D, inset), suggesting that not only are the number of terminal exons interconnected, but their usage is coupled as well. These findings indicate that gene architecture plays a pivotal role in determining coupling between terminal exons and thus regulating isoform diversity in genes.

PITA defines terminal exon coupling

The observed correlation of terminal exon usage could arise in one of two ways. Terminal sites may be directly coupled, whereby exons with similar ordinal positions co-occur on the same mRNA molecules (fig. S1D). Alternatively, coupling may be indirect, such that exons with similar ordinal positions are independently favored across transcripts, leading to apparent correlations in the short-read RNA-seq data. To distinguish between these possibilities and identify specific genes with coupled terminal end usage, we analyzed long-read isoform sequencing (Iso-Seq) data across human tissues and cells from the ENCODE Project ([38], [39]) and evaluated whether terminal exons were coupled within the same molecules. Specifically, we wanted to quantify how often an mRNA that started at a particular TSS was likely to terminate at a PAS with a similar ordinal position. For example, after conditioning on full-length reads [methods, fig. S2; ([40], [41])], *MYO10* has 184 long mRNA reads in H9 cells, distributed across three TSSs and two primary PASs. Of the 128 reads that begin at the first TSS, 94% terminate at the first primary PAS and 59% of the remaining 56 reads, which start downstream of the first TSS, terminate at the second primary PAS (Fig. 2A). Thus, long reads from *MYO10* support direct coupling of terminal exons based on their ordinal position, consistent with evidence from short-read RNA-seq data. These observations support the existence of an

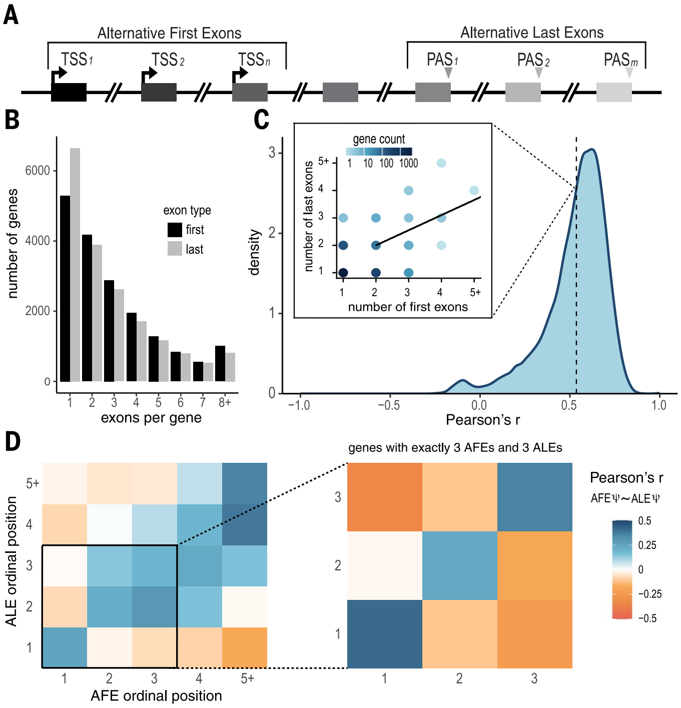


Fig. 1. A positional relationship between sites of mRNA initiation and termination in human genes. (**A**) Schematic of a gene with alternative start and end sites. AFEs (black/dark gray) and their respective TSSs are shown on the left and the ALEs (light gray) and their respective PASs are shown on the right. Exons are numbered by the order in which they appear in the direction of transcription (ordinal position). (**B**) The total number of protein coding genes (y-axis) that use a given number of annotated AFEs (black) or ALEs (gray), aggregated across all GTEx tissues. (**C**) Distribution of Pearson's r between the number of expressed AFEs and ALEs per gene in all GTEx samples ($n = 17{,}350$; mean $r = 0.53$). Inset shows a representative ovary sample, marked by the dashed line of the distribution plot, in which genes expressing n AFEs (x-axis) and n ALEs (y-axis) are depicted. Color intensity represents the number of genes exhibiting each unique AFE–ALE count combination. The trend line (black line) reflects the correlation between the number of expressed AFEs and ALEs in genes using multiple AFE and ALEs (Pearson's $r = 0.55$, P-value $= 1.85 \times 10^{-8}$). (**D**) Heatmap of Pearson's r for pairwise correlations between the relative usage (Ψ) of AFEs and ALEs based on their ordinal position for genes with multiple first and last exons (left panel, $n = 1{,}560{,}899$ exons) and genes with exactly three AFEs and three ALEs (right panel, $n = 63{,}811$ exons).

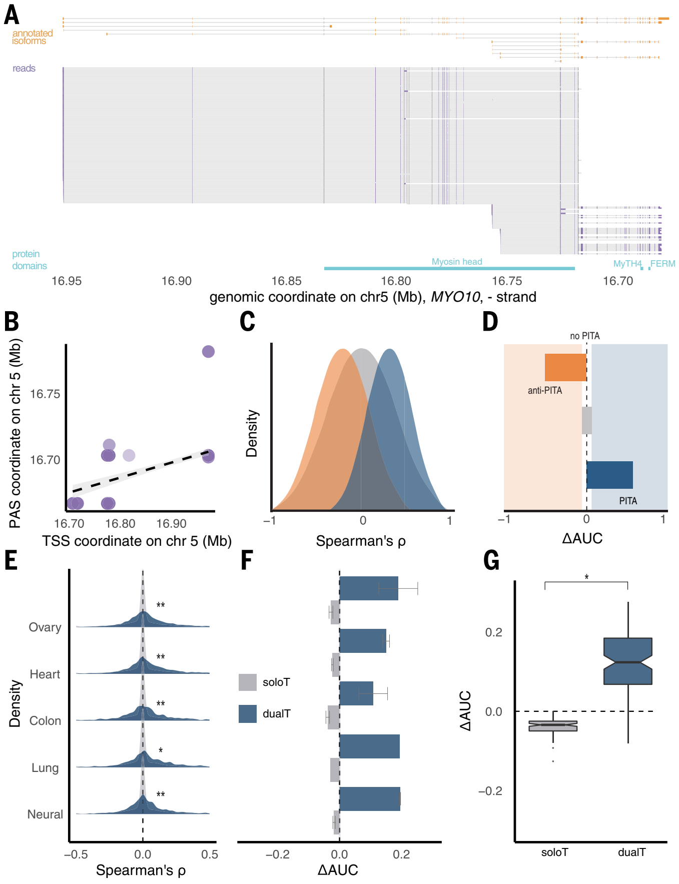


Fig. 2. Positional coupling of TSS-PAS usage occurs within individual mRNA molecules. (**A**) Annotated isoforms (top, orange), a subset of LRS reads (middle, purple with introns in thin black lines), and annotated protein domains (bottom, light blue) in H9 cells for MYO10. (**B**) Correlation between LRS read start (x-axis) and end coordinates (y-axis) for MYO10 (Spearman's $\rho = 0.66$). (**C**) Schematic of expected genome-wide distributions for Spearman's ρ showing possible shifts toward negative correlations (orange, anti-PITA), positive correlations (blue, PITA), or an unbiased distribution (gray, no PITA) using fictional data. (**D**) Schematic of expected ΔAUCs for the categories outlined in (C). ΔAUC is defined as $\mathrm{AUC}_{\rho>0}$ to $\mathrm{AUC}_{\rho<0}$. Distribution of Spearman's ρ (**E**) and the mean ΔAUC across samples (**F**) for solo termini genes (soloT, gray) and dual alternative termini genes (dualT, blue) in five human tissue types. K-S test, *P-value $< 10^{-8}$; ** P-value $< 10^{-16}$ for (E). (**G**) Distribution of ΔAUC values across 109 LRS samples for soloT (gray) and dualT genes (blue). t-test *P-value $< 10^{-16}$.

intramolecular regulatory paradigm that we call PITA, which underlies the coupling of transcription initiation and CPA decisions.

To evaluate the extent to which intramolecular coupling of ordinal terminal exons occurs globally, we estimated a Spearman's ρ to look at the rank correlation between TSS and PAS positions across reads within a gene. *MYO10* has a Spearman's ρ of 0.66 (Fig. 2B). Since rank corresponds to the ordinal position across terminal sites, positive correlations indicate ordinal coupling (PITA) whereas negative correlations indicate coupling of exons according to inverse ordinal position (anti-PITA) (schematic of possible distributions shown in Fig. 2C). We then used area under the curve (ΔAUC) to quantify whether the distribution of Spearman's ρ per sample shifted away being centered around zero (methods), in which positive or negative ΔAUCs would again indicate genome-wide enrichments of PITA or anti-PITA genes, respectively (schematic of possible ΔAUCs shown in Fig. 2D).

When looking at genes that use multiple first exons and multiple PASs within a cell type [dual alternative termini (dualT)], we observed substantial enrichments of PITA genes across multiple human cells and tissues (shown for five tissues in Fig. 2, E and F). We considered genes that use only one first exon or one PAS within each sample [solo termini (soloT)] as a control that accounts for variability in long read start and end coordinates (fig. S3A) and observe that these genes consistently show weakly negative ΔAUCs, indicating that our observations are not due to biases in the metric or identification of terminal ends from long reads. These trends are also not biased by read depth (fig. S3B), gene expression (fig. S3C), read thresholds (fig. S3D), read length (fig. S3F), or other spurious correlations in the data (assessed with permutations, fig. S3E). Notably, the ΔAUCs are even more biased toward positive values when we calculated Pearson correlations using ordinal positions of sites rather than genomic coordinates of reads (fig. S3, G and H) and the enrichment remains across different thresholds for calculating the ΔAUC (fig. S3I). On average, the distribution of PITA coupling across genes is more similar between replicates than nonreplicates (fig. S3J). Finally, we observe a similar enrichment of PITA genes when we use data from orthogonal experimental approaches [transcript isoform sequencing (TIF-seq) and complementary DNA–polymerase chain reaction (cDNA-PCR) Nanopore sequencing datasets; fig. S4]. Overall, we observe that 80% of samples show evidence for greater than expected PITA coupling, with a 3 to 14% enrichment of PITA genes coupling across 109 Iso-Seq samples from 47 tissues and cell types (Fig. 2G) corresponding to 58 to 460 genes using conservative thresholds (methods). Together, our results suggest that there is direct coupling of terminal exon usage and we can identify genes that are enriched for coupled terminal exon usage. This relationship is biased toward coupled usage of terminal exons with similar ordinal positions in a gene (Fig. 3A).

PITA is independent of tissue-specific mRNA regulation

mRNA terminal ends are often regulated across tissues, with abundant tissue-specific usage of both TSSs and PASs. Indeed, 83% of genes expressed in the long-read dataset are regulated in a tissue-specific manner, such that they use distinct isoforms across tissues. It is also known that tissue- or context-specific regulation of alternative TSSs or PASs results in global shifts toward preferential usage of upstream or downstream sites ([20], [42–46]). Even among dualT genes, there is usually one predominant isoform expressed, likely due to tissue-specific regulation of isoform usage and gene expression. When we control for this differential isoform usage and gene expression, the enrichment of PITA genes increases (9 to 23% enrichment; fig. S3B). Thus, PITA coupling within dualT genes is likely not a byproduct of global tissue-specific signatures, but rather an independent driver of full-length isoform usage. To investigate this regulatory paradigm further, we asked whether genes that show strong tissue-specific regulation of terminal sites also show PITA coupling across tissues. Specifically, we looked at genes that use unique termini within tissues but alternative TSSs and alternative PASs across tissues (tissue-dualT). For example, *DNJC11* expresses distinct TSSs and PASs across lung, induced pluripotent stem cells (iPSCs), and astrocyte cells (Fig. 3B). However, across cell types there is a significant correlation of *DNJC11* TSS and PAS usage, with the expressed isoforms showing a preference for PITA coupling (Spearman's ρ = 0.92). Using a tissue-ΔAUC metric in which Spearman's ρ is calculated using reads across tissues for tissue-dualT genes (methods), we observe an enrichment of PITA coupling that is stronger than that observed in most individual tissues (17 to 19% enrichment; Fig. 3C) corresponding to 982 genes showing cross-tissue PITA regulation. This observation suggests that PITA coupling is a pervasive phenomenon in which the coordination of transcript starts and ends might be influenced by gene architecture even when only one isoform is expressed per tissue or constitutively expressed within a cell type.

The preferential usage of PITA isoforms may enable the regulation of RNA or protein features related to the positioning of elements within a gene. For instance, if PITA coupling allows for the regulation of isoforms with specific functional significance, then PITA genes may show conserved features specifically at the terminal ends of transcripts. Consistent with this, the upstream and downstream TSS and PAS regions for dualT genes with PITA coupling are more conserved than similar regions for non-PITA or anti-PITA dualT genes (Fig. 3D). Furthermore, genes expressing PITA isoforms more often encode for a greater variety of protein domains than anti-PITA or non-PITA genes (methods, Fig. 3E), despite all of these genes containing similar numbers of domains on average (fig. S5, A and B). For instance, *MYO10* PITA isoforms starting at the upstream TSS are able to encode a myosin head domain whereas PITA isoforms starting at the downstream PAS are able to encode MyTH4 and FERM domains (Fig. 2A, bottom track). Together, these observations suggest that PITA coupling may help to shape mRNA function or proteome diversity through coordinated exon usage across the terminal ends of an isoform. Notably, we do not observe either increased conservation or protein domain diversity for genes with intertissue PITA coupling (fig. S5, C and D), suggesting that all tissue-dualT genes may already be under stronger functional constraints due to cell type–specific isoform regulation ([18], [21], [47]).

Alternative first exon usage directly influences last exon usage

These correlative analyses suggest a potential causal relationship between the ordinal, coupled usage of TSSs and alternative PASs. To directly test this, we used dCas9-CRISPR activation or interference approaches to perturb 27 specific AFEs across 16 genes. We achieved specific activation or repression for eight AFEs, and six of these perturbations changed downstream ALE expression in a manner consistent with PITA coupling. For example, activating AFE1 of *ZNF638* led to an increase in the expression of ALE1 and a concurrent decrease in the expression of both AFE2 and ALE2 (Fig. 3F, left). Similarly, activating AFE1 of *TP53* caused a comparable increase in the expression of ALE1, with no substantial changes in the expression of AFE2 or ALE2 (fig. S6A). Activation of AFE2 had similar effects for three genes: Activating AFE2 of the *MAST1* gene resulted in increased expression of ALE2 whereas both AFE1 and ALE1 decreased (Fig. 3F, middle) and activation of AFE2 in both *SYT9* and *LEPR* predominantly increased expression of ALE2 (fig. S6, B and C). Finally, repressing AFE2 of *SWI5* reduced the expression of ALE2 without considerably altering the expression of AFE1 or ALE1 (Fig. 3F, right). These results provide direct evidence that AFE choice can regulate ALE expression, supporting a model in which terminal exon usage is coordinated in an ordinal and functionally coupled manner.

Notably, perturbing the usage of PAS elements did not induce changes in AFE usage that are consistent with PITA coupling. Upon deletion of the downstream ALE in *ZNF638* [392 nucleotides (nt), fig. S6D], we observe substantial reductions in the expression of both ALE1 and ALE2 but a major increase in the expression of AFE1 (fig. S6E). This result

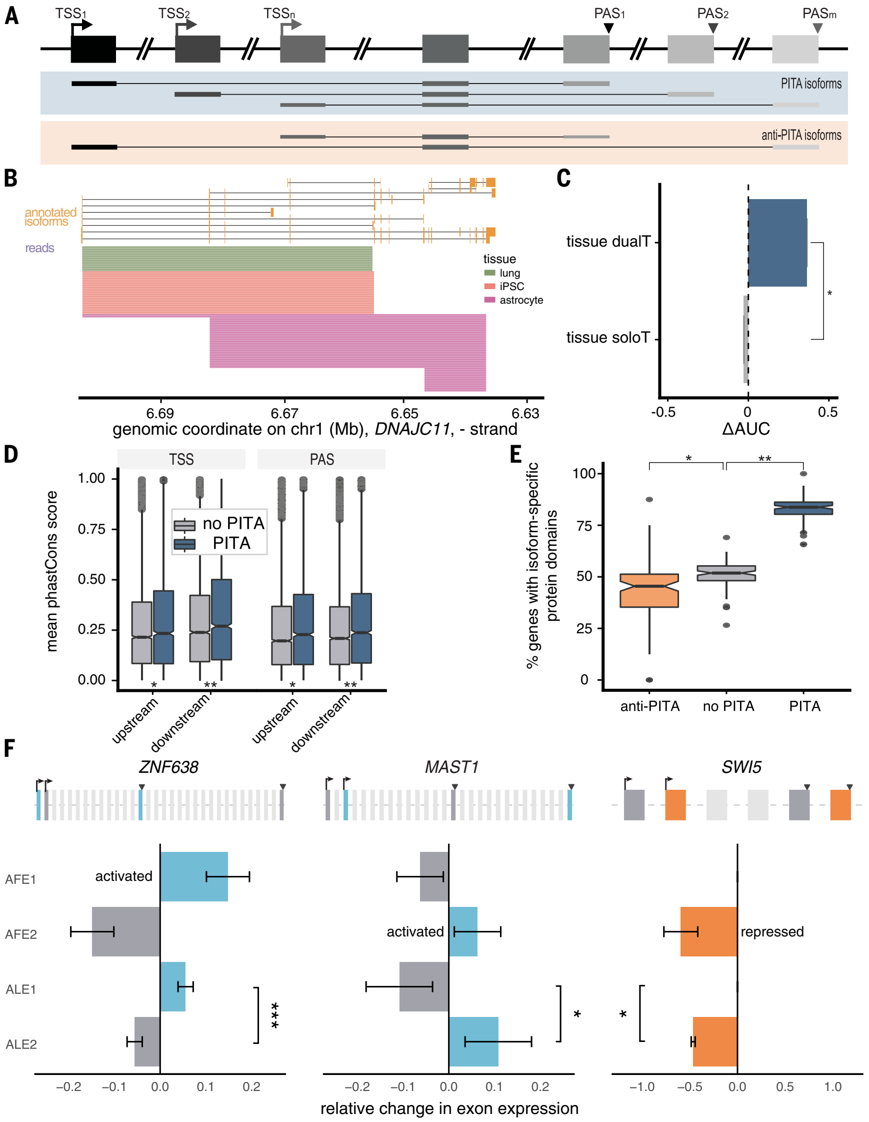


Fig. 3. Transcription start sites directly regulate transcript end site usage to govern the expression of functionally distinct isoforms. (**A**) Schematic of mRNA isoforms based on PITA classification. PITA isoforms preferentially use TSSs and PASs that are ordinally similar (light blue) whereas anti-PITA isoforms preferentially use ordinally different TSSs and PASs (light orange). (**B**) Annotated mRNA isoforms (orange) and a randomly subsampled proportion of reads for DNAJC11 in lungs (green), iPSCs (pink), and astrocytes (purple). (**C**) ΔAUC values for genes with dual alternative termini across tissues. Error bars represent standard error across 100 samples of reads across tissues. t-test *P-value $< 10^{-16}$. (**D**) Conservation scores (mean phastCons score, y-axis) in a 400-nt region around each terminal site of the two most highly expressed isoforms for genes with dual alternative termini. K-S test *P-value $< 10^{-3}$; **P-value $< 10^{-7}$. (**E**) Percentage of dual alternative termini genes (y-axis) whose isoforms overlap different annotated protein domains. t-test *P-value $< 10^{-7}$; **P-value $< 10^{-16}$. (**F**) CRISPR modulation of a given first exon drives concordant changes in the corresponding last exon of the same ordinal position in three protein-coding genes expressed in HEK293T-A2 cells. Changes in exon expression were quantified relative to control samples (methods). Error bars depict the standard error of means. The t-test measured whether the corresponding last exon of the same ordinal position exhibited a larger directional change than the noncorresponding last exon. CRISPR activation of *ZNF638* AFE1 resulted in an increase in ALE1 (left); CRISPR activation of *MAST1* AFE2 resulted in an increase in ALE2 (middle); and CRISPR interference of *SWI5* AFE2 resulted in a decrease in ALE2 (right). *P-value < 0.05, **P-value < 0.01, ***P-value < 0.001.

indicates that PAS perturbations can influence AFE usage but that these regulatory effects are not involved in PITA coupling. These findings suggest that PITA represents a unidirectional regulatory mechanism by which alternative TSSs influence ordinally paired alternative PAS usage, but PAS usage does not drive PITA coupling.

PITA coupling depends on gene length

PITA coupling suggests that the architectures of genes are configured to enable coregulation of transcript start and end sites. This coupling involves long-distance coordination across transcription initiation and polyadenylation sites that are often separated by thousands to hundreds of thousands of base pairs in DNA space and hundreds to thousands of nucleotides in pre-mRNA space. Thus, we first evaluated whether the DNA or RNA distance between these sites was correlated with the strength of PITA coupling across genes. PITA coupling is more likely to occur in longer genes, defined as the total genomic length between the upstream most-expressed TSS and downstream most-expressed PAS (Fig. 4A). However, PITA coupling is not associated with the pre-mRNA length of the resulting isoforms (Fig. 4A) or the mRNA lengths of the isoforms expressed (estimated by average read length, methods and fig. S7A). This observation suggests that PITA coupling may be driven by transcriptional or co-transcriptional mechanisms and regulatory decisions.

A gene can be broken down into many regions, all of which contribute independently to the total length of a gene. To understand how the length of a gene may influence PITA coupling, we calculated the lengths of the regions involved in transcriptional regulation (interval between TSSs), in splicing regulation (interval between downstream-most TSS and upstream-most PAS), and in cleavage and polyadenylation regulation (interval between PASs). We find that both the TSS and PAS intervals are strongly associated with PITA, with longer intervals between alternative TSSs and alternative PASs leading to stronger PITA coupling (Fig. 4B). However, the length of the internal pre-mRNA interval is mostly not associated with the strength of PITA coupling, except for long genes in which the increased number of alternative TSSs and PASs limits the lengths of internal pre-mRNA intervals. To understand whether the TSS and PAS intervals were directly associated with PITA regulation, we used a comparative genomics approach and treated evolutionary changes as natural perturbations that vary gene length. Using Iso-Seq data from mouse tissues matched to the human tissues analyzed [ENCODE (38, 39); methods], we identified species-specific PITA genes as described above (fig. S7, B and C). We find that genes only showing PITA coupling in human cells are, on average, significantly longer in the human genome than in the mouse genome and correspondingly, mouse-specific PITA genes are significantly longer in the mouse genome [Kolmogorov-Smirnov (K-S) test P-value < 0.001; Fig. 4C and fig. S7D]. Genes that show no PITA coupling in either species do not vary in gene length. This pattern of species-specific PITA genes having longer genomic intervals in the PITA species also holds for TSS and PAS intervals. The opposite is also true as genes with the largest difference in gene length or TSS and PAS interval lengths between species are enriched for species-specific PITA genes in species with longer lengths (fig. S7E). Together, these observations suggest that the strength of PITA coupling is directly associated with longer genomic distances.

Since PITA coupling shows a relationship with both TSS and PAS intervals, we wanted to understand whether certain genes were more

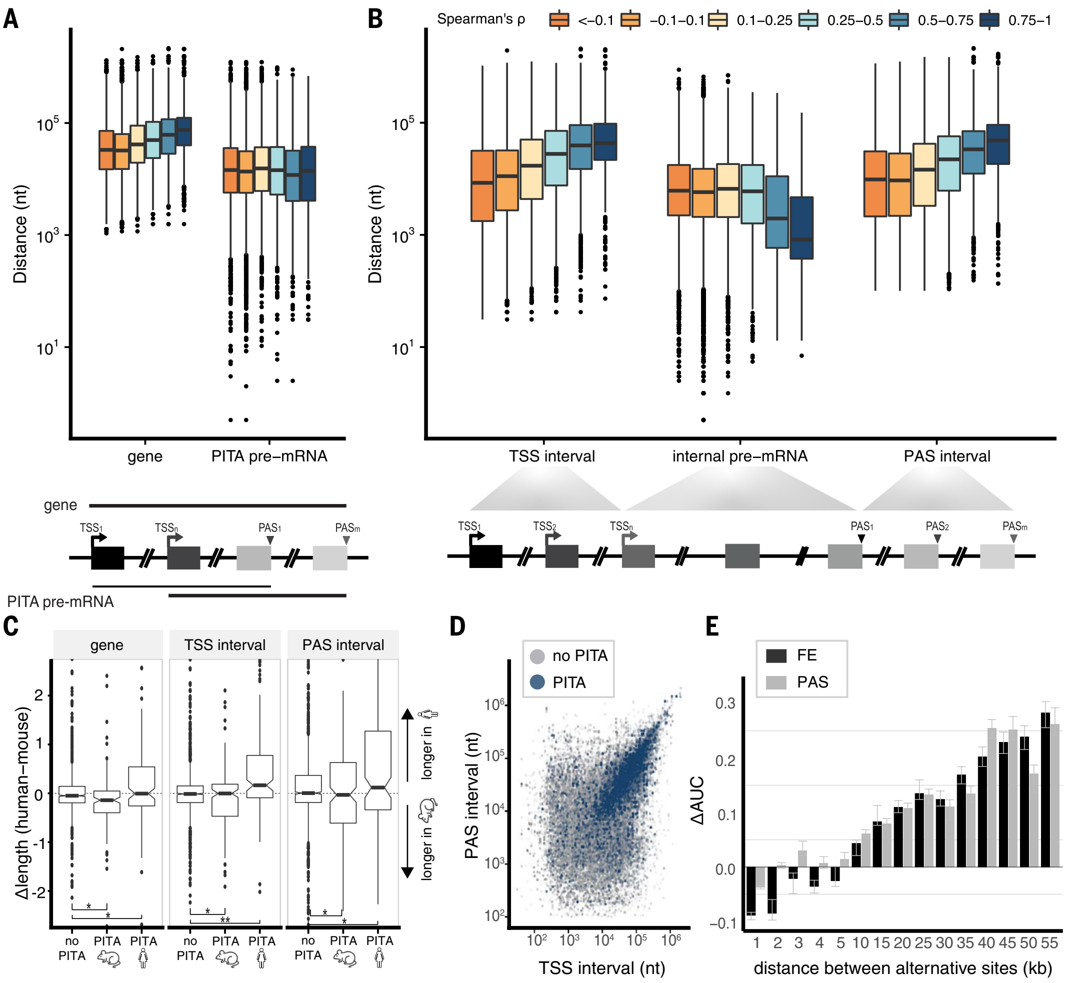


Fig. 4. PITA occurs more often in longer genes. (**A**) Distribution of the lengths (*y*-axis) of dual alternative termini genes (left) and PITA pre-mRNAs (right) for genes within bins of Spearman's ρ values (colors). (**B**) Distribution of the maximum distances (*y*-axis) between TSSs (left), the downstream-most TSS and upstream-most PAS (internal pre-mRNA, middle), and PASs (right) for dual alternative termini genes within varying bins of Spearman's ρ values (colors). (**C**) Distribution of the change in feature lengths between human and mouse orthologs (*y*-axis) for genes that are not PITA in either species, PITA only in mice, or PITA only in humans. To account for global differences in gene lengths between species, distances were first normalized by the mean distance within each species for each feature. K-S Test **P*-value $< 10^{-3}$; ***P*-value $< 10^{-7}$. (**D**) Correlation between the TSS intervals (*x*-axis) and PAS intervals (*y*-axis) for dual alternative termini genes non-PITA genes (gray, Spearman's $\rho < 0.3$, Pearson's $r = 0.59$; *P*-value $< 2.2 \times 10^{-16}$) and PITA genes (blue, Spearman's $\rho \geq 0.3$, Pearson's $r = 0.82$; *P*-value $< 2.2 \times 10^{-16}$). (**E**) ΔAUC values binned by distances between FEs or PASs (black and gray, respectively) across pairwise comparisons between mRNA isoforms. Error bars represent bootstrapped 95% confidence intervals (CI).

likely to show an association with TSS intervals and others with PAS intervals. Given our initial observation that genes are more likely to have similar numbers of first and last exons and a correlation between gene length and terminal interval length (fig. S8, A and B), we first asked whether genes are also more likely to have similar transcription and CPA region intervals. Although TSS and PAS intervals are weakly correlated globally (Pearson's r = 0.59), this correlation is even stronger for PITA genes (Pearson's r = 0.82; Fig. 4D). The change in TSS interval length between human and mouse is also significantly correlated to the change in interspecies PAS interval length for human- and mouse-specific PITA genes (fig. S8C). Finally, we asked if there was a minimum distance between alternative sites at which PITA coupling is more likely to occur. PITA coupling is enriched when any two alternative TSSs or PASs are separated by at least 5 kb, regardless of whether the alternative PASs are present on the same exon or different exons (Fig. 4E and fig. S8D). Together, these observations suggest that there are selective pressures to coordinate the lengths of genomic regions that are involved in the regulation of alternative transcription and CPA.

Distinct promoter regions and 3D architecture around PITA-coupled TSSs and PASs

The association between PITA coupling and the lengths of genomic intervals suggests that PITA may be driven by mechanisms related to transcriptional regulation of PITA genes, including the relative positioning of regulatory elements at the beginning of genes (*34*). We hypothesized that alternative TSSs may have distinct regulatory patterns in PITA genes. To test this, we used ultradeep data from human embryonic stem (H1-ESC) and foreskin fibroblast (HFF) cells [4D Nucleome (*48*)] to estimate the positioning and strength of several promoter-defining marks such as chromatin accessibility, H3K4me3, CTCF binding, and 3D chromatin characteristics (measured by insulation strength from micro-capture (micro-C) data indicative of boundaries between structural chromatin domains). Upstream TSSs in both PITA and non-PITA genes are characterized by high chromatin accessibility, strong insulation, and well-positioned H3K4me3 and CTCF patterns (figs. S9A and S10A), regardless of the absolute or relative expression of the TSS. However, the presence of insulation and promoter-defining marks at downstream TSSs is correlated with the relative usage of the TSS. This is especially true for signatures of chromatin domains, in which major TSSs (PSI > 0.5), including downstream TSSs, have equally strong insulation in both non-PITA and PITA genes whereas minor TSSs have insulation in PITA genes but not in non-PITA genes (figs. S9B and S10B). PITA genes also appear to have an additional insulation subdomain across the TSS interval, bounded by the upstream-most and downstream-most TSS (figs. S9C and S10C). By contrast, insulation at PASs appears to be less precisely positioned—likely because PASs are less well defined and more variable than TSSs between cells ([29], [49])—and therefore more difficult to capture at the same resolution as TSSs, as shown before ([50]). Together, these observations suggest that alternative TSSs in PITA genes may have distinct chromatin conformations that could facilitate downstream transcriptional dynamics.

Finally, we used ultradeep micro-C data ([48], [51]) to investigate whether there is any evidence for specific localized physical interactions between the terminal ends of PITA or non-PITA genes and/or whether PITA genes display more global differences in chromatin conformation at the level of chromatin domains and domain boundaries, as compared with non-PITA genes (methods). To confidently distinguish signals across individual sites, these analyses were limited to ~150 and ~60 extremely long non-PITA and PITA genes, respectively, with at least 2.5 kb between alternative TSSs and PASs and a minimum gene length of 8 kb (methods). Previously, it has been shown by Hi-C that active TSSs and PASs display insulation ([32]): These loci prevent long-range chromatin interactions across them, leading to the formation of structural chromatin domains containing active genes from start to end, and demarcated by insulating boundaries ([32], [34], [50], [52]). Consistently, we found that both non-PITA and PITA genes form structural domains defined across the entire gene, with insulation boundaries at upstream TSS and downstream PAS (figs. S9D and S10D). In addition, aggregate analyses suggest that PITA genes might form multiple overlapping structural domains whose boundaries are defined by relationships between the upstream TSS–upstream PAS and downstream TSS–downstream PAS pairs (figs. S9D and S10D), which could reflect or reinforce relationships between mRNA initiation, termination, and isoform expression.

RNAPII elongation rates mediate the relationship between mRNA ends

The length of a gene influences the total distance that RNAPII must travel to transcribe full-length mRNAs. The increased time necessary to transcribe these longer genes, with greater spacing between TSSs or PASs, may also create a coupling between transcriptional dynamics and the choice of alternative RNA processing sites. To test whether altered RNAPII elongation rates could directly influence PITA coupling, we analyzed publicly available RNA-seq data from human cells overexpressing RNAPII mutants that elongate either slower or faster than wild-type (WT) RNAPII, paired with a control line (Fig. 5A) ([53]). Although the slow RNAPII markedly disrupted the

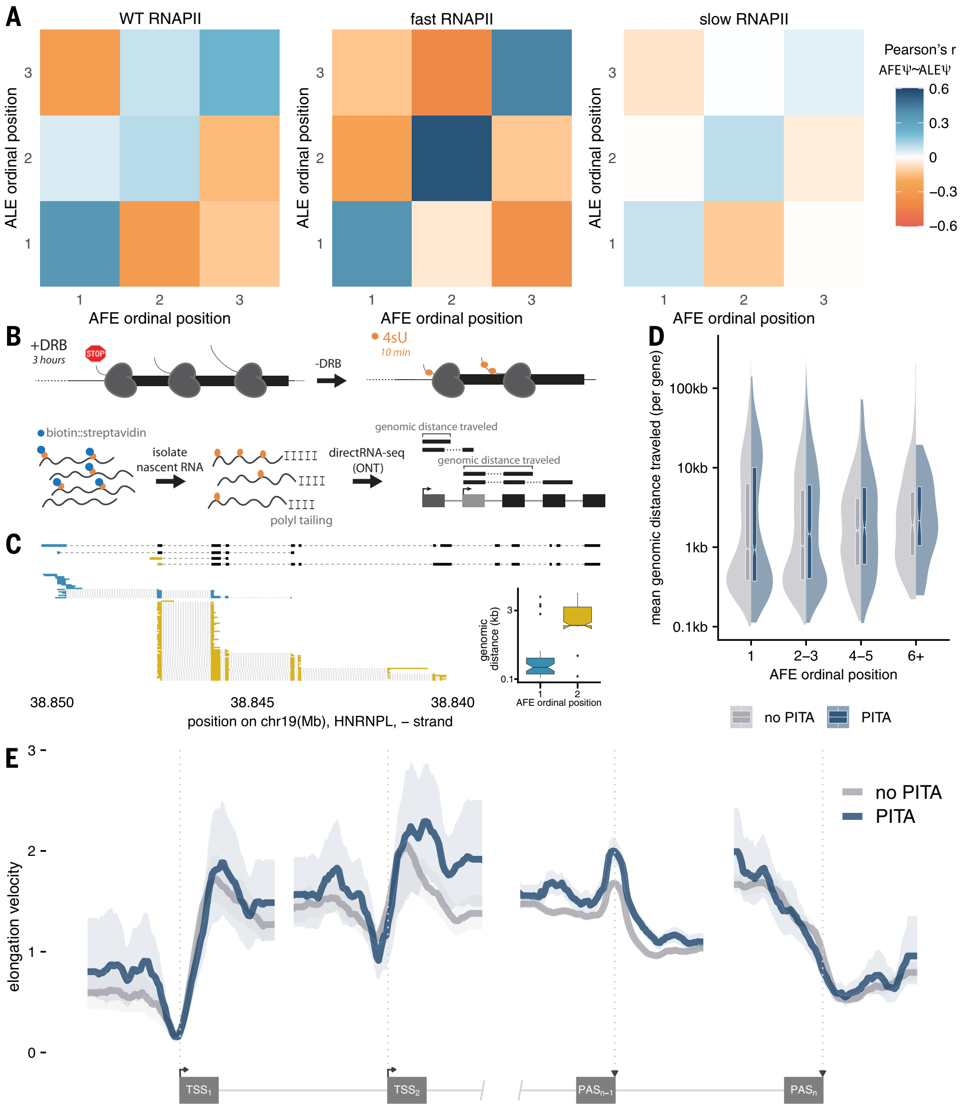


Fig. 5. RNA polymerase II elongation rate is associated with PITA. (**A**) Heatmaps of Pearson's r values for the pairwise correlations between the relative usage (Ψ) of a gene's AFEs and ALEs based on their ordinal position in HEK293T-A2 cells expressing for WT RNAPII (left, $n = 49$), rapidly elongating RNAPII (middle, R749H mutation, $n = 69$) and slowly elongating RNAPII (right, E1126C mutation, $n = 62$). All heatmaps show pairwise correlations for genes expressing exactly three AFEs and three AFEs. (**B**) Schematic of the 4sU-DRB-LRS protocol, which involves transcription blockage and then synchronization with DRB, followed by labeling of nascent RNA with 4sU for 10 min, then long-read library preparation using polyI tailing to facilitate direct RNA LRS. The bottom right schematic shows the estimation of genomic distances that the reads map to, which reflects the distance that RNAPII traveled from each TSS in the labeling period. (**C**) Direct RNA-seq for HNRNPL from long-read 4sUDRB-seq data in K562 cells. Shown are annotated isoforms (top, black) and LRS reads (middle, introns represented by thin gray lines) colored by whether they start in the first expressed AFE (blue) or the second expressed AFE (yellow). Inset shows the distributions of genomic distances across reads for each AFE. (**D**) Distribution of mean genomic distances for each gene, conditioned on reads starting in AFEs with increasing ordinal positions for no PITA (gray) and PITA genes (blue). Boxes show the median and 5 to 95% CI for each distribution. (**E**) Distribution of elongation velocities around upstream and downstream TSSs (left) and PASs (right) for no PITA (gray) and PITA (blue) genes. Elongation velocities are calculated using 50-nt bins across ± 5-kb windows around each site and smoothed with a sliding window approach for visualization (methods). Background density represents the CIs across genes.

correlated usage of ordinally similar terminal exons relative to the WT cells, the rapidly elongating RNAPII strengthened the enrichment of PITA coupling. This observation was consistent in an edited mouse cell line expressing the slow RNAPII mutation (fig. S11A) ([54]). This suggests that RNAPII elongation rates are crucial for coupling TSS-PAS interactions in long PITA genes.

To estimate gene-specific rates of RNAPII elongation, we used the ratio of transient transcriptome sequencing [(TT-seq), quantifying transcribed RNA] to mammalian native elongating transcript sequencing [(mNET-seq), quantifying actively transcribing RNAPIIs] data in K562 cells to calculate RNAPII elongation velocities, as described before ([55], [56]). As has been observed previously ([57]), we observe that longer genes have faster RNAPII elongation velocities. Although PITA genes have significantly faster overall RNAPII elongation velocities relative to both non-PITA and anti-PITA genes (methods; K-S test P-value $< 1 \times 10^{-15}$, fig. S11B), they are not transcribed faster, on average, than expected based on their length (fig. S11C). When we look across different regions of the gene, however, we observe that PITA genes have increased elongation velocities in the middle of the gene body relative to non-PITA genes (fig. S11D).

These results suggest that long PITA genes may be elongated at an optimal rate that is crucial for enabling sustained RNAPII processivity until the end of the gene. Previous work has suggested that faster elongation rates lead to a global increase in the usage of more downstream PASs ([29], [58], [59]). Since PAS strength progressively increases across a gene body [i.e., upstream PASs are weaker than downstream PASs ([60])], this is consistent with the “window of opportunity” model in which faster synthesis favors usage of stronger sites when weaker and stronger sites are in competition. Downstream PASs are stronger than upstream PASs in both PITA and non-PITA genes (fig. S11E). PITA coupling between TSSs and PASs suggests that alternative TSSs may have different elongation rates and, if the rate of elongation at the beginning of mRNA synthesis remains consistent throughout the gene, this may influence PAS usage at the 3′ end. Specifically, we hypothesized that RNAPII starting at downstream TSSs may elongate faster to promote usage of downstream PASs in PITA genes.

To estimate TSS-specific RNAPII elongation rates, we combined the previously described 4-thiouridine-DRB (4sUDRB-seq) ([61]) approach with long-read sequencing (LRS) of nascent RNA (Fig. 5B and methods). With this approach, we can infer TSS-specific elongation rates by estimating the distance that individual RNAPII molecules travel after transcription initiation using the total genomic distance between the read start and end coordinates. The mean distance that RNAPII molecules elongate from each TSS in the labeling period is highly correlated between two independent replicates (Pearson’s r = 0.76, fig. S12A), providing confidence in the robustness of these measurements. Using these long read estimates, we asked about differences in elongation rate between TSSs within the same gene. For example, we observe that reads starting from the upstream TSS of *HNRNPL*, on average, map to a genomic distance of 821 nt whereas reads starting from the downstream TSS, on average, map to a genomic distance of 2606 nt (Fig. 5C). Globally, we observe that more downstream TSSs are associated with longer average genomic distances per gene and that this effect is more pronounced for PITA genes (Fig. 5D and fig. S12B). These results are consistent when performing LRS of cDNA and conditioning on the presence of a 5′ template switching oligo to obtain reads whose 5′ ends reflect TSSs with higher confidence ([40]) (fig. S12C).

Finally, to further characterize elongation rate differences between alternative sites, we again used elongation velocities across different regions of the gene. Although there is a minimal difference in elongation velocities between PITA and non-PITA genes around upstream TSSs, PITA genes are associated with higher elongation velocities at downstream TSSs (Fig. 5E). Promoter sequences do not appear to explain the faster elongation from downstream sites as these downstream promoters contain fewer core promoter elements ([62]) and are less likely to be predicted as active, especially in PITA genes (fig. S12, D and E). PITA genes also have higher elongation velocities around upstream PASs (Fig. 5E), likely leading to the reduced usage of these weaker PASs when RNAPII begins elongating from downstream TSSs. Our results indicate that the RNAPII elongation rate across a gene might be dependent on where the RNAPII started elongating. Specifically, downstream TSSs and upstream PASs are associated with faster RNAPII elongation rates, especially in PITA genes. Together, our observations indicate that a combination of spatial and kinetic mechanisms contribute to a relationship between pre-mRNA ends and ultimate expression of mRNA isoforms (Fig. 6).

Discussion

We observed widespread coordination between mRNA initiation and termination. A recent report in *Drosophila* tissues and human organoids showed that “dominant promoters” can be associated with usage of one or multiple polyadenylation sites ([7]). We find little overlap between PITA genes and genes with “dominant promoters” (fig. S4, B to D), suggesting that these regulatory paradigms may or may not be mutually exclusive.

Previous work has also shown intriguing associations between TSS, internal exon, and PAS usage ([1], [4], [14], [63]). Biochemical and genetic evidence indicates that components of the splicing and CPA machinery, as well as auxiliary regulatory factors, are loaded onto the C-terminal domain of RNAPII at the time of transcription initiation to facilitate co-transcriptional RNA processing ([11], [64], [65]). Transcription and co-transcriptionally loaded factors can also influence splicing ([66], [67]), the recruitment of transcription factors ([68], [69]), and the usage of premature PASs ([8]). Studies have found connections between skipped exons and either TSSs or PASs over evolutionary timescales ([5], [70]), cancer, cellular transitions ([31]), and across tissues ([7], [14]). Our observations of direct TSS-PAS coupling add to this complex and coordinated regulatory landscape driving isoform usage. Although preliminary evidence points to the possibility of interconnections between skipped exon usage and TSS-PAS coregulation ([7]), it remains to be seen to what extent PITA coupling also drives the composition of internal exons in an isoform.

Despite past and present work, it remains unclear what molecular forces drive these correlations. Previous studies have suggested that transcriptional enhancers and elongation factors contribute to the regulation of alternative PASs by binding to upstream enhancer elements ([71]) or by playing a role in recruiting PAS machinery ([11]). However, these features do not explain the pairing of TSS-PAS sites in an

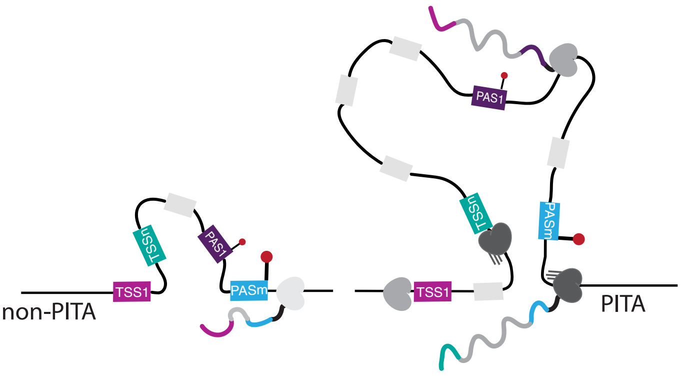


Fig. 6. Spatial organization and kinetic mechanisms drive the ordinal coupling of mRNA initiation and termination in PITA genes. Our results support a model in which longer PITA genes (right) exhibit faster elongation rates when transcription initiates at downstream TSSs. Faster RNAPII trafficking persists through upstream weaker PASs, leading to skipping of these sites in favor of downstream stronger PASs.

ordinal fashion across a gene. Consistent with our observations, previous literature has suggested the involvement of chromatin domains in constraining RNAPII directionality (72), particularly by insulation domains that encompass an entire gene.

We show that the rate of RNAPII elongation is dependent on where transcription begins and that downstream TSSs are associated with faster rates of RNAPII elongation, especially for longer PITA genes. This is consistent with our previous work showing stronger activity in downstream promoters, despite weaker promoter sequences (62). Our results suggest a model by which RNAPII molecules starting at downstream TSSs elongate faster throughout the gene and are more likely to reach a downstream PAS. This aligns with the "window of opportunity" model for RNA processing and suggests the existence of an optimal RNAPII elongation rate for RNA processing events (53). Although the window of opportunity model is classically used to describe the regulation of individual RNA processing sites on a pre-mRNA molecule, PITA coupling suggests that the expression of full-length isoforms may also be governed by differences in elongation rates within and across genes.

Our results also indicate that spatial organization and kinetic mechanisms jointly influence the coupled regulation of mRNA processing events across a gene. It is possible that these relationships exist to maintain efficient transcriptional environments and that PITA coupling is a by-product of these regulatory fluctuations. Insulated subdomains may structure chromatin to place specific regions within cellular subcompartments enriched for transcription and/or splicing machinery, which recent work has indicated may promote RNAPII processivity and recycling across multiple rounds of transcription (33, 73–76). 3D architecture may be especially important for efficient CPA as it would bring RNAPII in contact with kinases that phosphorylate RNAPII to initiate transcription and elongation. The same phosphorylation marks are also necessary for CPA, but it has been seen that phosphorylation levels diminish across a gene (9), potentially setting an intrinsic limit on the productive elongation distance of RNAPII. However, our analyses are unable to infer any causality or directionality between these processes and so it is possible that chromatin conformations influence RNAPII processivity or recycling, or vice versa.

Recent evidence indicates that longer and more highly expressed genes in human cells are more likely to form cohesin-mediated extrusion loops that are bounded by transcription start and end sites (33). Thus, there might be a minimum genomic interval required between TSSs and between PASs to establish paired insulation across TSSs and PASs in mammalian genomes. However, it is unclear whether the first evolutionary determinant of PITA coupling is the regulation of terminal site usage, the distance between terminal sites, or the dynamics of RNAPII in the region. Our results would support the hypothesis that the change in genomic lengths across the gene would be accompanied by (and perhaps driven by) changes in RNAPII elongation rates.

PITA coupling represents paired mRNA 5′ and 3′ end usage that influences ultimate isoform compositions, adding complexity to the promoter-centric view of mRNA isoform regulation. This intricate interplay of gene regulatory mechanisms may be further modulated by tissue-specific factors that increase the specific usage of terminal sites. Understanding these and additional regulators of coupled mRNA terminal site usage will be key to revealing how cells fine-tune isoform expression and define their identities.

Materials and methods

Short-read RNA sequencing analysis

RNA-sequencing data from 17,350 tissue samples spanning 54 tissue sites from GTEx Version 8 were downloaded through dbGaP under study accession number phs000424.v8.p2. RNA-seq data from RNAPII mutant cell lines was downloaded from the NCBI Gene Expression Omnibus (GSE63375). RNA-seq data from mouse data was downloaded from the NCBI Gene Expression Omnibus (GSE127741). A detailed description of short read sequencing samples used can be found in table S1. For all short-read sequencing data, reads were mapped to genome assembly GRCh38.p14 for human or GRCm38.95 for mouse, using STAR v2.7.1a (77). Resulting bam files were processed using the HIT index pipeline to classify and quantify first and last exon usage (35). For exon classification, we utilized conservative parameters on all short read data analyzed, with a HIT_identity_parameters.txt file as follows: *# |HITindex| threshold for calling terminal exons [0.0, 1.0] HITterminal 1.0, # |HITindex| threshold for calling hybrid exons [0.0, 1.0] HIThybrid 0.5, # bootstrapping p-value threshold for HITindex significance [0.0, 1.0], HITpval 1, # confidence interval to use for HITindex significance (none, 0.75, 0.95, 0.95), HIT_CI none, # probability threshold for medium confidence with generative model [0.0, 1.0], prob_med 0.8, # probability threshold for high confidence with generative model [0.0, 1.0], prob_high 0.8.* Given the high sequencing depth in GTEx data, we raised the minimum threshold for the number of reads for confidence in HITindex classification ('—readnum' argument) to 10 for all GTEx samples.

PacBio long read sequencing analysis

Fastq files from PacBio reads were downloaded from the ENCODE Project data portal (https://www.encodeproject.org/) (38), including 113 human samples from 47 tissues and cell types and 130 mouse samples from 9 tissues (detailed in table S1). cDNA-PCR PacBio sequencing data for human organoids was downloaded from the Gene Expression Omnibus (GSE203583). LRS data was mapped to the hg38 human or mm38 reference genome using minimap2 (78) following the developers' recommended parameters: *-ax splice:hq -uf* and *-ax splice* for PacBio and cDNA-PCR reads respectively. Reads were divided into multiple split features to define exons using *bedtools bamtobed (79)* and assigned to overlapping genes. Only primary alignments and reads with all features assigned to the same gene were kept for downstream analyses. A detailed description of the long read sequencing samples can be found in table S1.

To retain full length reads, we then conditioned on reads whose 5′ end was within 20 nt of an empirically determined first exon identified in the full set of GTEx v8 samples (36) for human and a subset of ENCODE samples (80) for mouse, processed by the HITindex pipeline, and whose 3′ end was within 50nt of an experimentally verified human or mouse PAS in the polyASite 2.0 database (81). More than 80% of PacBio reads passed these filtering criteria (40). In addition, fully unspliced reads or reads whose 5′ and 3′ end mapped to the same exon were discarded.

TIF-seq analysis

Processed TIF-seq2 data containing the coordinates and counts of transcript start-end pairs from K562 cells was downloaded from the Gene Expression Omnibus (GSE140912). Each start and end site was intersected with all human genes and pairs which met the following criteria were retained for downstream analyses: (1) both termini aligned to the same gene, (2) the start feature was within 20 nt of an empirically derived first exon (as identified above), and (3) the end feature was within 50nt of an experimentally verified human PASs.

Identifying PITA genes

To identify PITA genes, only genes with at least 10 reads from PacBio Iso-seq and isoforms with at least 2 reads were considered. Genes were classified having dual alternative or solo termini by assessing the total number of first exon and polyA peaks used in each sample. Dual alternative termini genes were defined as those using at least two first exons and at least two polyA peaks, while solo termini genes were using only one first exon or polyA peak (but reads could exhibit variations in start and end coordinates, due to minor technical or biological fluctuations in the exact nucleotide positions defining the read boundaries, fig. S3A). All mapped reads per gene per sample were used to calculate a Spearman's ρ between the start and end coordinates of the reads. Genes were classified as anti-PITA, no PITA or PITA using Spearman's ρ thresholds of $\rho \leq -0.2$, $-0.2 > \rho < 0.2$, or $\rho \geq 0.2$ respectively.

To quantify the enrichment of PITA genes, we first calculated an area under the curve for positive and negative Spearman's ρs using the *density.default* function in R to calculate AUCs using n = 512 and cut = 3. Only samples with more than 1000 genes were considered for the analysis. ΔAUCs were calculated by computing $AUC_{\rho>0} - AUC_{\rho<0}$.

Intertissue PITA genes

To identify intertissue PITA genes, we define tissue-dual alternative termini genes as those classified as solo termini in individual tissues. To avoid tissue-specific gene expression biases, we performed 100 sampling iterations, selecting two reads per isoform per gene per tissue-dual alternative termini, and then calculated the Spearman's ρ between the start and end coordinates of the reads for all the sampled reads across tissues. Genes were then re-classified as solo or dual alternative termini genes based on the number of first exons and polyA peaks used across tissues.

Calculating genomic distances

Genomic distances were calculated using estimated coordinates for first exons and polyA peaks (as described above) for the following intervals: (1) gene length: $PAS_n - TSS_1$, (2) pre-mRNA length of PITA isoforms: $PAS_1 - TSS_1$ and $PAS_n - TSS_n$, (3) TSS interval: $TSS_n - TSS_1$, (4) internal pre-mRNA interval: $PAS_1 - TSS_n$ only calculated when TSS_n is upstream of PAS_1, (5) PAS interval: $PAS_n - PAS_1$.

Identifying dominant promoters with the LATER pipeline

To identify dominant promoters as defined in Alfonso-Gonzalez *et al.* ([7]), the LATER pipeline (https://github.com/hilgers-lab/LATER) was applied to unfiltered human PacBio and cDNA-PCR samples following with recommended parameters ([7]). Genes were classified as having promoter dominance when promoter usage was higher than 0.2 and end dominance when PAS usage was higher than 0.6 ([7]).

Analysis of protein domains

Genomic coordinates for protein domains were obtained through the InterPro portal ([82]). Dual alternative termini genes were overlapped with the annotated protein domains. The proportion of genes in which different isoforms overlap different protein domains was calculated for the pre-mRNA regions of isoforms with support from more than 2 reads.

Conservation analysis

For conservation analyses, we considered the two highest expressed isoforms [most long-read sequencing (LRS) reads] from the tissue with the highest Spearman's ρ. Genes with Spearman's ρ > 0.2 were considered to be PITA genes and all others were used as a control set. The transcript start and end coordinates from the two isoforms were ordered by genomic position and the conservation in a 400 nt genomic region centered around each site was analyzed. Conservation scores were obtained by applying the *bigWigAverageOverBed* tool from *ucsctools* to extract the mean phastCons score for each region using a hg38 phastCons (100 mammal alignment) bigwig file downloaded from the UCSC database ([83], [84]).

CRISPR-a and CRISPR-i experiments

HEK293T-A2 cells were cultured in Dulbecco's modified Eagle's medium (DMEM, Gibco #11965118) containing D-Glucose (4.5 g/L), L-Glutamine and 10% fetal bovine serum (FBS, Gibco #A31406-02) ([85]). Cells were seeded into 6-well plates and transiently transfected ~16 hours later, at ~80% confluency, using lipofectamine3000 (Invitrogen #L3000001). CRISPR activation experiments were carried out by transfecting empty UniSAM (Addgene #99866) vector as a control or UniSAM vector containing an sgRNA targeting the 400 bp region upstream of the intended first exon target ([86]). CRISPR interference experiments were carried out by transfecting a dCas9 vector (Addgene #159086) as a control, or the dCas9 vector containing an sgRNA targeting the 400 bp region downstream of the intended first exon target. All transfections consisted of 3 biological replicates. 48 hours post-transfection, total RNA was extracted using RNeasy Mini Kit (Qiagen #74104) according to the manufacturer's instructions. For *LEPR*, assessment of relative exon usage was performed using qPCR. Briefly, reverse transcription was performed in a reaction mix containing 1 µg of total RNA, using Maxima H Minus First Strand cDNA Synthesis Kit (Thermo Fisher Scientific, #K1652) according to the manufacturer's instructions. Quantitative PCR analyses were performed with SYBR green labeling (Thermo Fisher Scientific Maxima SYBR Green/ROX qPCR Master Mix (2X), #K0222) on the QuantStudio 5 Real-Time PCR System (Applied Biosystems #A28574). For the 7 remaining genes, relative exon usage was assessed from RNA-seq data, which was performed in two batches. For the first batch of genes (*SWI5*, *SYT9*, *TP53*, and control RNA), RNA was sequenced by first doing a polyA enrichment with the NEBNext® Poly(A) mRNA Magnetic Isolation Module (NEB #E7490L) followed by the NEBNext® Ultra II Directional RNA Library Prep Kit for Illumina (#E7760L). Libraries were sequenced on a NextSeq2000. For the second batch (*MAST1*, *ZNF638*, and control RNA), extracted RNA was sent to Novogene Corporation for polyA enriched mRNA library preparation and high-throughput sequencing on a NovaSeq 6000 with 150nt paired-end reads. Raw fastq files were processed as described in the short-read RNA sequencing analysis method section and analyzed using the HITindex pipeline to estimate relative terminal exon usage. Samples were compared to control RNA sequenced within the same batch. All gRNA oligos and qPCR primers for the CRISPR-a or CRISPR-i experiments are listed in table S2.

Estimating polyadenylation site strengths

APARENT2 ([60]) was used to estimate PAS strength for all reported human PAS in the polyASite 2.0 database ([87]) using a 200nt window centered about the middle of each peak and the pre-trained *aparent_all_libs_resnet_no_clinvar_wt_ep_5_var_batch_size_inference_mode_no_drop* model. Logits were calculated as $\ln\left(\frac{APARENT2\ score}{1-APARENT2\ score}\right)$.

Polyadenylation site deletion experiments

We ran APARENT2 as described above across the last exon of *ZNF638* to find the downstream most PAS. We designed two sgRNAs targeting the region flanking the PAS (one upstream and one downstream). Two hundred thousand HEK293T-A2 cells were electroporated using the Neon Transfection System (Thermo-Fisher #MPK5000) with 60pmol of sgRNAs plus 20pmol of SpyCas9. After three days, the cells were diluted to a limiting dilution of 0.25 cells per well to isolate individual clones. Each individual clone was validated using PCR and looking at the size of resulting products using capillary electrophoresis (fig. S6D). We selected the clone with the highest proportion of deleted alleles. Sequences for the sgRNAs and primers are available in table S2.

Analyses of micro-C and epigenetics data

Tier 1 micro-C ultradeep datasets for HFFc6 4DNFI9FVHJZQ and H1-ESC 4DNFI9GMP2J8 were downloaded from the 4DN Portal ([48], [88]) as multi-resolution cooler files ([89]) and used for analysis of interaction landscapes around PITA and no PITA genes. Using HFFc6 and H1-ESCs PacBio data for genes with at least 10 reads and isoforms with at least 2 reads were considered. We selected the most upstream and downstream FE and polyA peak coordinates for genes using multiple FE and PAS. To ensure all selected sites could be clearly visualized on micro-C heatmaps at 500bp resolution, we only used PITA and non-PITA genes with a minimum distance of 2kb between all selected sites and a gene length of 8kb. We obtained 175 and 57 control and PITA genes, respectively, for H1-ESCs and 133 and 63 control and PITA genes, respectively, for HFF. Average aggregated interaction profiles around genes (rescaled meta-gene pileups) were generated using coolpup.py (90) as follows: regions of interest were extended 100% up- and down-stream (flanking) and corresponding observed/expected matrices were extracted from microC (cooler-files) at 500 bp resolution, rescaled to 200*200 pixels, matrices corresponding to negatively stranded genes were "flipped", and averaged. Insulation profiles were calculated genome-wide for both microC datasets at 250 bp resolution using insulation window size of 2500 bp using *cooltools* (91) and stored as bigwig files. Stackups of relevant epigenetic features/tracks [insulation, ATAC-seq, H3K4me3 ChIP-seq, CTCF ChIP-seq (51)] were generated by extracting bigwig signal binned into 200 bins for each genomic interval ($TSS_{1,n}$, $PAS_{1,n}$ flanked with 7500 bp upstream and downstream) using pyBBI package (92) (https://zenodo.org/records/10382981). Insulation profiles for each genomic interval were normalized by subtracting mean flanking signal (50 upstream most and 50 downstream most bins) as in (50).

Calculation of RNAPII elongation velocities

TT-seq and mNET-seq datasets from K562 cells were downloaded from the Gene Expression Omnibus (GSE148433 and GSE159633, respectively, detailed in table S1). For both TT-seq and mNET-seq datasets, initial read trimming was performed using *cutadapt* (93) with parameters *–minimum-length 25*, *–quality-cutoff 25*, and *–overlap 12*. The adapter sequences -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA and -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT were removed from TT-seq reads and -a TGGAATTCTCGGGTGCCAAGGAACTCCAGTCAC and -A AGATCGTCGGACTGTAGAACTCTGAAC were removed from mNET-seq reads. Both trimmed datasets were then aligned to the human reference genome GRCh38.v43 using STAR (77), with alignment parameters*–quantMode GeneCounts*, *–outSAMtype BAM SortedByCoordinate*, *–outFilterType BySJout*, *–outFilterMultimapNmax 1*, *–alignSJoverhangMin 8*, *–alignSJDBoverhangMin 1*, *–outFilterMismatchNmax 999*, *–outFilterMismatchNoverLmax 0.02*, *–alignIntronMin 20*, *–alignIntronMax 1000000*, and *–alignMatesGapMax 1000000*. Spike-in reads were aligned to the yeast reference genome S.cerevisiae.R64 and ERCC synthetic spiked-in sequences as described in (94), for the purpose of normalization of TT-seq counts and mNET-seq counts, respectively. RNAPII elongation velocities were calculated by dividing the amount of transcribed RNA (TT-seq coverage) by the density of transcribing RNAPII (mNET-seq reads), as described previously (55, 56). For each gene, the mean elongation velocity across replicates was used for downstream analyses. Elongation velocities around alternative TSSs and PASs were estimated by calculating the mean enrichment and confidence intervals of TT-seq and mNET-seq reads in 100 bins across a 10 kb window centered on each TSS/PAS. Genes with (1) mean normalized read counts < 0.05 or > 0.95 and (2) less than 5 kb spacing between upstream and downstream TSSs or PASs were excluded from further analyses. Elongation velocity for each bin was calculated by dividing TT-seq signal by mNET-seq signal, and the distribution of elongation velocity across the region was smoothed using a sliding window approach with 10-bin windows before final visualization.

Estimation of TSS-specific elongation rates

K562 cells (ATCC #CCL-243) were grown at 37°C, 5% CO2 using Gibco RPMI-1640 medium (+ glutamine) (Gibco #11875119) supplemented with 10% heat-inactivated FBS (Thermo Fisher Scientific #A5256801) and Penicillin/Streptomycin (Gibco #15140122). Cells were treated cells with 100uM 5,6-Dichlorobenzimidazole 1-beta-D-ribofuranoside (DRB, Sigma Aldrich #D1916) for 3 hours. During the last 5 min of DRB treatment, 4-thio-uridine (4sU; ThermoFisher #J60679.MD) was added to a final concentration of 1mM (as described in (61)). After washing with PBS (Gibco #10010049), cells were resuspended in media containing 1mM 4sU for 10 min and spun down at 1000 RCF for 2 min in DNA low binding falcon tubes. The pellet was resuspended in Trizol (Thermo Fisher Scientific #15596018) and flash frozen in liquid nitrogen. RNA was extracted using Trizol-Chloroform and the aqueous phase was cleaned up using RNA Clean & Concentrator-100 columns (Zymo Research #R1019). 4sU pulldowns were performed with MTSEA biotin-XX (95), followed by RNAseH-based (Biosearch Technologies #H39500) ribosomal RNA depletion using custom probes targeted to mammalian rRNA sequences (96).

For direct RNA sequencing, polyinosine (polyI) tails were added as follows: 300ng of RNA was heated at 80C for 2 min, then mixed with a solution containing 0.5mM Tris-HCl pH 7 (Fisher Scientific #BP1756-100), 0.5 mM ITP (Sigma Aldrich #I0879-50MG), SUPERase. In 1 unit/ul (Invitrogen #AM2696), 2 ul of yeast polymerase and yeast polyA polymerase buffer (Thermo isher Scientific #74225Z25KU) and incubated at 37C for 40 min (as described in (97). The direct RNA nanopore library protocol (SQK-RNA004) was performed using manufacturer's instructions with one modification to replace the standard reverse transcription adapter (RTA) with a custom poly(C) splinted adapter (97). Libraries were sequenced on a GridION machine (GXB02310) with RNA flow cells (FLO-MIN00RA). rRNA reads were removed from fastq files using SortMeRNA (98) and the remaining reads were mapped to GRCh38.95 using minimap2/2.24 (78) with the following parameters: -ax splice -uf -k14. Uniquely aligned reads whose 5′ end overlapped an annotated AFEs and 3′ end did not overlap an annotated PAS peak were used for downstream analysis since they likely represent complete molecules that had not yet finished transcription. Genes with at least 10 reads and isoforms with at least 2 reads were retained for further analysis, as described above.

For cDNA sequencing, samples followed the same protocol as described above until polyI tailing. After polyI tailing, the RNA was reverse transcribed with the Takara Bio SMARTer® PCR cDNA Synthesis Kit (Takara Bio #634926) using a custom sequence for the 3′ end adapter [as described in (99)]: TGAGTCGGCAGAGAACTGGCGAANNNNNNNNNNNCCCCCCCCCC. PCR was then conducted using the Takara Bio Advantage 2 PCR kit (Takara Bio #639207) and the custom PCR primer TGAGTCGGCAGAGAACTGGCGAA. The cDNA nanopore library protocol (SQK-LSK114) was used to prepare libraries that were sequenced on a GridION (GXB02310) machine using FLO-MIN114 flow cells (R10) (Oxford Nanopore Technologies #FLO-MIN114). Prior to rRNA read removal, fastq files were processed with Porechop (https://github.com/rrwick/Porechop) to identify and trim 5′ and 3′ adapters in the reads. To select for complete molecules, reads were retained only if they had the proper adapter at the 5′ end of the read. rRNA reads were removed as described above and remaining reads were mapped with minimap2/2.24 with these parameters: -ax splice. Downstream analysis steps were conducted as described above.

Promoter prediction, GC content, and CpG island analysis

K562 control RNA-seq data was downloaded from the Gene Expression Omnibus (GSE148433, under accession numbers SRR11521575 and SRR11521576). The data was processed using the HIT index pipeline to characterize multi-promoter genes and assign ordinal position to each promoter. To assess promoter properties, the sequences of expressed promoters were extracted from 275 nt upstream and 50 nt downstream of each transcription start site using the *bedtools getfasta* function. The iProEP promoter prediction algorithm was employed to predict the inherent presence of a promoter in the extracted sequences (100). The presence of CpG islands was determined using a 200 bp window moving across the extracted sequences. A CpG island was defined based on the observed/expected CpG value and a GC content greater than 50%, as described in Kim *et al.* (62).

REFERENCES AND NOTES

1. A. Joglekar *et al.*, Single-cell long-read mRNA isoform regulation is pervasive across mammalian brain regions, cell types, and development. bioRxiv 2023.04.02.535281 [Preprint] (2023); doi: 10.1101/2023.04.02.535281

2. T. W. Nilsen, B. R. Graveley, Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457–463 (2010). doi: 10.1038/nature08909; pmid: 20110989
3. B. Tian, J. L. Manley, Alternative polyadenylation of mRNA precursors. *Nat. Rev.* **18**, 18–30 (2017). doi: 10.1038/nrm.2016.116; pmid: 20110989
4. S. Y. Anvar *et al.*, Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biol.* **19**, 46 (2018). doi: 10.1186/s13059-018-1418-0; pmid: 29598823
5. A. Fiszbein, K. S. Krick, B. E. Begg, C. B. Burge, Exon-Mediated Activation of Transcription Starts. *Cell* **179**, 1551–1565.e17 (2019). doi: 10.1016/j.cell.2019.11.002; pmid: 31787377
6. S. A. Hardwick *et al.*, Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. *Nat. Biotechnol.* **40**, 1082–1092 (2022). doi: 10.1038/s41587-022-01231-3; pmid: 35256815
7. C. Alfonso-Gonzalez *et al.*, Sites of transcription initiation drive mRNA isoform selection. *Cell* **186**, 2438–2455.e22 (2023). doi: 10.1016/j.cell.2023.04.012; pmid: 37178687
8. D. Kaida *et al.*, U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**, 664–668 (2010). doi: 10.1038/nature09479; pmid: 20881964
9. C. Laitem *et al.*, CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II-transcribed genes. *Nat. Struct. Mol. Biol.* **22**, 396–403 (2015). doi: 10.1038/nsmb.3000; pmid: 25849141
10. L. Caizzi *et al.*, Efficient RNA polymerase II pause release requires U2 snRNP function Efficient RNA polymerase II pause release requires U2 snRNP function. *Mol. Cell* **81**, p1920–1934.e9 (2021). doi: 10.1016/j.molcel.2021.02.016; pmid: 33689748
11. T. Nagaike *et al.*, Transcriptional activators enhance polyadenylation of mRNA precursors. *Mol. Cell* **41**, 409–418 (2011). doi: 10.1016/j.molcel.2011.01.022; pmid: 21329879
12. Y. E. Guo *et al.*, Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature* **572**, 543–548 (2019). doi: 10.1038/s41586-019-1464-0; pmid: 31391587
13. S. A. Shabalina, A. Y. Ogurtsov, N. A. Spiridonov, E. V. Koonin, Evolution at protein ends: Major contribution of alternative transcription initiation and termination to the transcriptome and proteome diversity in mammals. *Nucleic Acids Res.* **42**, 7132–7144 (2014). doi: 10.1093/nar/gku342; pmid: 24792168
14. E. T. Wang *et al.*, Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008). doi: 10.1038/nature07509; pmid: 18978772
15. A. R. Forrest *et al.*, A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014). doi: 10.1038/nature13182; pmid: 24670764
16. E. K. Stroup, Z. Ji, Deep learning of human polyadenylation sites at nucleotide resolution reveals molecular determinants of site usage and relevance in disease. *Nat. Commun.* **14**, 7378 (2023). doi: 10.1038/s41467-023-43266-3; pmid: 37968271
17. A. Reyes, W. Huber, Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.* **46**, 582–592 (2018). doi: 10.1093/nar/gkx1165; pmid: 29202200
18. S. Pal *et al.*, Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res.* **21**, 1260–1272 (2011). doi: 10.1101/gr.120535.111; pmid: 21712398
19. A. A. Pai, F. Luca, Environmental influences on RNA processing: Biochemical, molecular and genetic regulators of cellular response. *Wiley Interdiscip. Rev. RNA* **10**, e1503 (2019). doi: 10.1002/wrna.1503; pmid: 30216698
20. J. M. Taliaferro *et al.*, Distal Alternative Last Exons Localize mRNAs to Neural Projections. *Mol. Cell* **61**, 821–833 (2016). doi: 10.1016/j.molcel.2016.01.020; pmid: 26907613
21. S. Lianoglou, V. Garg, J. L. Yang, C. S. Leslie, C. Mayr, Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.* **27**, 2380–2396 (2013). doi: 10.1101/gad.229328.113; pmid: 24145798
22. D. C. Di Giammartino, K. Nishida, J. L. Manley, Mechanisms and consequences of alternative polyadenylation. *Mol. Cell* **43**, 853–866 (2011). doi: 10.1016/j.molcel.2011.08.017; pmid: 21925375
23. P. K. Parua, S. Kalan, B. Benjamin, M. Sansó, R. P. Fisher, Distinct Cdk9-phosphatase switches act at the beginning and end of elongation by RNA polymerase II. *Nat. Commun.* **11**, 4338 (2020). doi: 10.1038/s41467-020-18173-6; pmid: 32859893
24. P. K. Parua *et al.*, A Cdk9-PP1 switch regulates the elongation-termination transition of RNA polymerase II. *Nature* **558**, 460–464 (2018). doi: 10.1038/s41586-018-0214-z; pmid: 29899453
25. C. D. Kaplan, H. Jin, I. L. Zhang, A. Belyanin, Dissection of Pol II trigger loop function and Pol II activity-dependent control of start site selection in vivo. *PLOS Genet.* **8**, e1002627 (2012). doi: 10.1371/journal.pgen.1002627; pmid: 22511879
26. H. Braberg *et al.*, From structure to systems: High-resolution, quantitative genetic analysis of RNA polymerase II. *Cell* **154**, 775–788 (2013). doi: 10.1016/j.cell.2013.07.033; pmid: 23932120
27. C. Qiu *et al.*, Universal promoter scanning by Pol II during transcription initiation in Saccharomyces cerevisiae. *Genome Biol.* **21**, 132 (2020). doi: 10.1186/s13059-020-02040-0; pmid: 32487207
28. J. V. Geisberg, Z. Moqtaderi, K. Struhl, The transcriptional elongation rate regulates alternative polyadenylation in yeast. *eLife* **9**, e59810 (2020). doi: 10.7554/eLife.59810; pmid: 32845240
29. J. V. Geisberg *et al.*, Nucleotide-level linkage of transcriptional elongation and polyadenylation. *eLife* **11**, e83153 (2022). doi: 10.7554/eLife.83153; pmid: 36421680
30. M. A. Cortazar *et al.*, Control of RNA Pol II Speed by PNUTS-PP1 and Spt5 Dephosphorylation Facilitates Termination by a "Sitting Duck Torpedo" Mechanism. *Mol. Cell* **76**, 896–908.e4 (2019). doi: 10.1016/j.molcel.2019.09.031; pmid: 31677974
31. R. Goering *et al.*, LABRAT reveals association of alternative polyadenylation with transcript localization, RNA binding protein expression, transcription speed, and cancer survival. doi: 10.1101/2020.10.05.326702
32. M. J. Rowley *et al.*, Condensin II Counteracts Cohesin and RNA Polymerase II in the Establishment of 3D Chromatin Organization. *Cell Rep.* **26**, 2890–2903.e3 (2019). doi: 10.1016/j.celrep.2019.01.116; pmid: 30865881
33. H. Wu, J. Zhang, L. Tan, X. Sunney Xie, Extruding transcription elongation loops observed in high-resolution single-cell 3D genomes. bioRxiv 2023.02.18.529096 [Preprint] (2023); doi: 10.1101/2023.02.18.529096.
34. S. Leidescher *et al.*, Spatial organization of transcribed eukaryotic genes. *Nat. Cell Biol.* **24**, 327–339 (2022). doi: 10.1038/s41556-022-00847-6; pmid: 35177821
35. A. Fiszbein *et al.*, Widespread occurrence of hybrid internal-terminal exons in human transcriptomes. *Sci Adv.* **8**, eabk1752 (2022). doi: 10.1126/sciadv.abk1752; pmid: 35044812
36. F. Aguet *et al.*, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020). doi: 10.1126/science.aaz1776; pmid: 32913098
37. Y. Katz, E. T. Wang, E. M. Airoldi, C. B. Burge, Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010). doi: 10.1038/nmeth.1528; pmid: 21057496
38. Y. Luo *et al.*, New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020). doi: 10.1093/nar/gkz1062; pmid: 31713622
39. F. Reese *et al.*, The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. bioRxiv 2023.05.15.540865 [Preprint] (2023); doi: 10.1101/2023.05.15.540865.
40. E. Calvo-Roitberg, R. F. Daniels, A. A. Pai, Challenges in identifying mRNA transcript starts and ends from long-read sequencing data. *Genome Res.* **34**, 1719–1734 (2024). doi: 10.1101/2023.07.26.550536; pmid: 37546743
41. X. Dong *et al.*, Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures. *Nat. Methods* **20**, 1810–1821 (2023). doi: 10.1038/s41592-023-02026-3; pmid: 37783886
42. C. Mayr, D. P. Bartel, Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**, 673–684 (2009). doi: 10.1016/j.cell.2009.06.016; pmid: 19703394
43. R. Sandberg, J. R. Neilson, A. Sarma, P. A. Sharp, C. B. Burge, Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science* **320**, 1643–1647 (2008). doi: 10.1126/science.1155390; pmid: 18566288
44. Z. Ji, J. Y. Lee, Z. Pan, B. Jiang, B. Tian, Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 7028–7033 (2009). doi: 10.1073/pnas.0900028106; pmid: 19372383
45. R. Hou, C.-C. Hon, Y. Huang, CamoTSS: analysis of alternative transcription start sites for cellular phenotypes and regulatory patterns from 5' scRNA-seq data. *Nat. Commun.* **14**, 7240 (2023). doi: 10.1038/s41467-023-42636-1; pmid: 37945584
46. A. L. Richards *et al.*, Environmental perturbations lead to extensive directional shifts in RNA processing. *PLOS Genet.* **13**, e1006995 (2017). doi: 10.1371/journal.pgen.1006995; pmid: 29023442
47. M. Tajnik *et al.*, Intergenic Alu exonisation facilitates the evolution of tissue-specific transcript ends. *Nucleic Acids Res.* **43**, 10492–10505 (2015). pmid: 26400176
48. N. Krietenstein *et al.*, Ultrastructural Details of Mammalian Chromosome Architecture. *Mol. Cell* **78**, 554–565.e7 (2020). doi: 10.1016/j.molcel.2020.03.003; pmid: 32213324
49. B. Schwalb *et al.*, TT-seq maps the human transient transcriptome. *Science* **352**, 1225–1228 (2016). doi: 10.1126/science.aad9841; pmid: 27257258
50. A.-L. Valton *et al.*, A cohesin traffic pattern genetically linked to gene regulation. *Nat. Struct. Mol. Biol.* **29**, 1239–1251 (2022). doi: 10.1038/s41594-022-00890-9; pmid: 36482254
51. B. Akgol Oksuz *et al.*, Systematic evaluation of chromosome conformation capture assays. *Nat. Methods* **18**, 1046–1055 (2021). doi: 10.1038/s41592-021-01248-7; pmid: 34480151
52. B. Bonev *et al.*, Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557–572.e24 (2017). doi: 10.1016/j.cell.2017.09.043; pmid: 29053968
53. N. Fong *et al.*, Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev.* **28**, 2663–2676 (2014). doi: 10.1101/gad.252106.114; pmid: 25452276
54. M. M. Maslon *et al.*, A slow transcription rate causes embryonic lethality and perturbs kinetic coupling of neuronal genes. *EMBO J.* **38**, e101244 (2019). doi: 10.15252/embj.2018101244; pmid: 30988016
55. K. Žumer *et al.*, Two distinct mechanisms of RNA polymerase II elongation stimulation in vivo. *Mol. Cell* **81**, 3096–3109.e8 (2021). doi: 10.1016/j.molcel.2021.05.028; pmid: 34146481

56. L. Caizzi *et al.*, Efficient RNA polymerase II pause release requires U2 snRNP function. *Mol. Cell* **81**, 1920–1934.e9 (2021). doi: 10.1016/j.molcel.2021.02.016; pmid: 33689748
57. A. Veloso *et al.*, Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res.* **24**, 896–905 (2014). doi: 10.1101/gr.171405.113; pmid: 24714810
58. A. Khitun *et al.*, Elongation rate of RNA polymerase II affects pausing patterns across 3′ UTRs. *J. Biol. Chem.* **299**, 105289 (2023). doi: 10.1016/j.jbc.2023.105289; pmid: 37748648
59. J. V. Geisberg, Z. Moqtaderi, K. Struhl, Chromatin regulates alternative polyadenylation via the RNA polymerase II elongation rate. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2405827121 (2024). doi: 10.1073/pnas.2405827121; pmid: 38748572
60. J. Linder, S. E. Koplik, A. Kundaje, G. Seelig, Deciphering the impact of genetic variation on human polyadenylation using APARENT2. *Genome Biol.* **23**, 232 (2022). doi: 10.1186/s13059-022-02799-4; pmid: 36335397
61. G. Fuchs *et al.*, 4sUDRB-seq: Measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome Biol.* **15**, R69 (2014). doi: 10.1186/gb-2014-15-5-r69; pmid: 24887486
62. G. Kim, C. L. Carroll, Z. P. Wakefield, M. Tuncay, A. Fiszbein, U1 snRNP regulates alternative promoter activity by inhibiting premature polyadenylation. *Mol. Cell* **85**, 1968–1981.e7 (2025). doi: 10.1016/j.molcel.2025.04.021; pmid: 40378830
63. Z. Zhang, B. Bae, W. H. Cuddleston, P. Miura, Coordination of alternative splicing and alternative polyadenylation revealed by targeted long read sequencing. *Nat. Commun.* **14**, 5506 (2023). doi: 10.1038/s41467-023-41207-8; pmid: 37679364
64. K. Glover-Cutter, S. Kim, J. Espinosa, D. L. Bentley, RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. *Nat. Struct. Mol. Biol.* **15**, 71–78 (2008). doi: 10.1038/nsmb1352; pmid: 18157150
65. H. Takahashi *et al.*, The role of Mediator and Little Elongation Complex in transcription termination. *Nat. Commun.* **11**, 1063 (2020). doi: 10.1038/s41467-020-14849-1; pmid: 32102997
66. F. Ullah, S. Jabeen, M. Salton, A. S. N. Reddy, A. Ben-Hur, Evidence for the role of transcription factors in the co-transcriptional regulation of intron retention. *Genome Biol.* **24**, 53 (2023). doi: 10.1186/s13059-023-02885-1; pmid: 36949544
67. M. Thompson *et al.*, Splicing in a single neuron is coordinately controlled by RNA binding proteins and transcription factors. *eLife* **8**, e46726 (2019). doi: 10.7554/eLife.46726; pmid: 31322498
68. K. Y. Kwek *et al.*, U1 snRNA associates with TFIIH and regulates transcriptional initiation. *Nat. Struct. Biol.* **9**, 800–805 (2002). doi: 10.1038/nsb862; pmid: 12389039
69. M. Uriostegui-Arcos, S. T. Mick, Z. Shi, R. Rahman, A. Fiszbein, Splicing activates transcription from weak promoters upstream of alternative exons. *Nat. Commun.* **14**, 3435 (2023). doi: 10.1038/s41467-023-39200-2; pmid: 37301863
70. A. Bergfort, K. M. Neugebauer, The promoter as a trip navigator: Guiding alternative polyadenylation site destinations. *Mol. Cell* **83**, 2395–2397 (2023). doi: 10.1016/j.molcel.2023.06.022; pmid: 37478824
71. B. Kwon *et al.*, Enhancers regulate 3′ end processing activity to control expression of alternative 3'UTR isoforms. *Nat. Commun.* **13**, 2709 (2022). doi: 10.1038/s41467-022-30525-y; pmid: 35581194
72. M. M. Ibrahim *et al.*, Determinants of promoter and enhancer transcription directionality in metazoans. *Nat. Commun.* **9**, 4472 (2018). doi: 10.1038/s41467-018-06962-z; pmid: 30367057
73. L. Han *et al.*, Concentration and length dependence of DNA looping in transcriptional regulation. *PLOS ONE* **4**, e5621 (2009). doi: 10.1371/journal.pone.0005621; pmid: 19479049
74. L. Chen *et al.*, R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters. *Mol. Cell* **68**, 745–757.e5 (2017). doi: 10.1016/j.molcel.2017.10.008; pmid: 29104020
75. A. M. Moabbi, N. Agarwal, B. El Kaderi, A. Ansari, Role for gene looping in intron-mediated enhancement of transcription. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 8505–8510 (2012). doi: 10.1073/pnas.1112400109; pmid: 22586116
76. N. Agarwal, A. Ansari, Enhancement of Transcription by a Splicing-Competent Intron Is Dependent on Promoter Directionality. *PLOS Genet.* **12**, e1006047 (2016). doi: 10.1371/journal.pgen.1006047; pmid: 27152651
77. A. Dobin *et al.*, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013). doi: 10.1093/bioinformatics/bts635; pmid: 23104886
78. H. Li, New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021). doi: 10.1093/bioinformatics/btab705; pmid: 34623391
79. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). doi: 10.1093/bioinformatics/btq033; pmid: 20110278
80. S. Lin *et al.*, Comparison of the transcriptional landscapes between human and mouse tissues. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 17224–17229 (2014). doi: 10.1073/pnas.1413624111; pmid: 25413365
81. C. J. Herrmann *et al.*, PolyASite 2.0: A consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Res.* **48**, D174–D179 (2020). doi: 10.1093/nar/gkz918; pmid: 31617559
82. T. Paysan-Lafosse *et al.*, InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427 (2023). doi: 10.1093/nar/gkac993; pmid: 36350672
83. A. Siepel *et al.*, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005). doi: 10.1101/gr.3715005; pmid: 16024819
84. B. J. Raney *et al.*, The UCSC Genome Browser database: 2024 update. *Nucleic Acids Res.* **52**, D1082–D1088 (2024). doi: 10.1093/nar/gkad987; pmid: 37953330
85. P. Khandelia, K. Yap, E. V. Makeyev, Streamlined platform for short hairpin RNA interference and transgenesis in cultured mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 12799–12804 (2011). doi: 10.1073/pnas.1103532108; pmid: 21768390
86. A. Fidanza *et al.*, An all-in-one UniSam vector system for efficient gene activation. *Sci. Rep.* **7**, 6394 (2017). doi: 10.1038/s41598-017-06468-6; pmid: 28743878
87. C. J. Herrmann *et al.*, PolyASite 2.0: A Consolidated Atlas of Polyadenylation Sites from 3′ End Sequencing. *Nucleic Acids Res* **8**, D174–D179 (2020). doi: 10.1093/nar/gkz918; pmid: 31617559
88. S. B. Reiff *et al.*, The 4D Nucleome Data Portal as a resource for searching and visualizing curated nucleomics data. *Nat. Commun.* **13**, 2365 (2022). doi: 10.1038/s41467-022-29697-4; pmid: 35501320
89. N. Abdennur, L. A. Mirny, Cooler: Scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311–316 (2020). doi: 10.1093/bioinformatics/btz540; pmid: 31290943
90. I. M. Flyamer, R. S. Illingworth, W. A. Bickmore, Coolpup.py: Versatile pile-up analysis of Hi-C data. *Bioinformatics* **36**, 2980–2985 (2020). doi: 10.1093/bioinformatics/btaa073; pmid: 32003791
91. Open2C *et al.*, Cooltools: enabling high-resolution Hi-C analysis in Python. bioRxiv 2022.10.31.514564 [Preprint] (2022); doi: 10.1101/2022.10.31.514564v1.
92. W. J. Kent, A. S. Zweig, G. Barber, A. S. Hinrichs, D. Karolchik, BigWig and BigBed: Enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010). doi: 10.1093/bioinformatics/btq351; pmid: 20639541
93. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011). doi: 10.14806/ej.17.1.200
94. L. Wachutka, L. Caizzi, J. Gagneur, P. Cramer, Global donor and acceptor splicing site kinetics in human cells. *eLife* **8**, e45056 (2019). doi: 10.7554/eLife.45056; pmid: 31025937
95. E. E. Duffy *et al.*, Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry. *Mol. Cell* **59**, 858–866 (2015). doi: 10.1016/j.molcel.2015.07.023; pmid: 26340425
96. Z. Zhang, W. E. Theurkauf, Z. Weng, P. D. Zamore, Strand-specific libraries for high throughput RNA sequencing (RNA-Seq) prepared without poly(A) selection. *Silence* **3**, 9 (2012). doi: 10.1186/1758-907X-3-9; pmid: 23273270
97. H. L. Drexler *et al.*, Revealing nascent RNA processing dynamics with nano-COP. *Nat. Protoc.* **16**, 1343–1375 (2021). doi: 10.1038/s41596-020-00469-y; pmid: 33514943
98. E. Kopylova, L. Noé, H. Touzet, SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012). doi: 10.1093/bioinformatics/bts611; pmid: 23071270
99. I. Legnini, J. Alles, N. Karaiskos, S. Ayoub, N. Rajewsky, FLAM-seq: Full-length mRNA sequencing reveals principles of poly(A) tail length control. *Nat. Methods* **16**, 879–886 (2019). doi: 10.1038/s41592-019-0503-y; pmid: 31384046
100. H.-Y. Lai *et al.*, IProEP: A computational predictor for predicting promoter. *Mol. Ther. Nucleic Acids* **17**, 337–346 (2019). doi: 10.1016/j.omtn.2019.05.028; pmid: 31299595
101. E. Calvo-Roitberg *et al.*, mRNA initiation and termination are spatially coordinated, Version v2, Zenodo (2025); https://doi.org/10.5281/zenodo.15837772.

ACKNOWLEDGMENTS

We thank Z.P. Wakefield for providing a curated list of human transcription start sites; H. Cheng and E. Sontheimer for discussions on CRISPR deletion experiments and the gift of SpyCas9 RNP; M. Walhout for suggesting the interspecies comparative analyses; P. Zamore for discussions; and the Fiszbein and Pai labs for helpful discussions and comments. **Funding:** This work was funded by the following: National Institutes of Health grants R35GM133762 and R01HG012967 (to A.A.P.); National Institutes of Health grant R35GM147254 and National Science Foundation grant CAREER2339464 (to A.F.); National Institutes of Health grants R01HG003143 and UM1HG011536 (to J.D.) **Author contributions:** Conceptualization: A.F. and A.A.P.; Methodology: E.C.-R., C.L.C., V.S., A.F., and A.A.P.; Formal Analysis: E.C.-R., C.L.C., G.Y.K., V.S., S.V.V., and S.T.M.; Investigation: E.C.-R., C.L.C., V.S., J.D.P., and M.U.-A.; Visualization: E.C.-R., C.L.C., G.Y.K., V.S., S.V.V., A.F., and A.A.P.; Funding acquisition: J.D., A.F., and A.A.P.; Project administration: A.F. and A.A.P.; Supervision: J.D., A.F., and A.A.P.; Writing – original draft: A.F. and A.A.P.; Writing – review & editing: E.C.-R., C.L.C., J.D., A.F., and A.A.P.; **Competing interests:** J.D. is an investigator of the Howard Hughes Medical Institute. J.D. is also an inventor on patent application US 12,146,186 B2, held by the University of Massachusetts Chan Medical School, Harvard College, the Whitehead Institute for Biomedical Research, and the Massachusetts Institute of Technology, which covers Hi-C technology. **Data and materials availability:** All code used for analysis in this manuscript is archived at Zenodo (*101*). Analyses of short and long-read sequencing datasets is detailed at https://github.com/thepailab/mrna_initiation_termination. Code for microC-related analyses is available at https://github.com/dekkerlab/pita_project. Sequencing data for CRISPRa/CRISPRi and DRB-4sU-LRS are available under the GSE294979 and GSE294980 accession numbers, respectively. **License information:** Copyright © 2025 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/content/page/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.ado8279
Materials and Methods; Figs. S1 to S12; Tables S1 and S2; MDAR Reproducibility Checklist

Submitted 22 February 2024; resubmitted 25 April 2025; accepted 9 July 2025

10.1126/science.ado8279