nature structural & molecular biology

TECHNICAL REPORT

https://doi.org/10.1038/s41594-022-00785-9

Screening thousands of transcribed coding and non-coding regions reveals sequence determinants of RNA polymerase II elongation potential

Hanneke Vlaming ✉, Claudia A. Mimoso, Andrew R. Field, Benjamin J. E. Martin and Karen Adelman ✉

Precise regulation of transcription by RNA polymerase II (RNAPII) is critical for organismal growth and development. However, what determines whether an engaged RNAPII will synthesize a full-length transcript or terminate prematurely is poorly understood. Notably, RNAPII is far more susceptible to termination when transcribing non-coding RNAs than when synthesizing protein-coding mRNAs, but the mechanisms underlying this are unclear. To investigate the impact of transcribed sequence on elongation potential, we developed a method to screen the effects of thousands of INtegrated Sequences on Expression of RNA and Translation using high-throughput sequencing (INSERT-seq). We found that higher AT content in non-coding RNAs, rather than specific sequence motifs, drives RNAPII termination. Further, we demonstrate that 5′ splice sites autonomously stimulate processive transcription, even in the absence of polyadenylation signals. Our results reveal a potent role for the transcribed sequence in dictating gene output and demonstrate the power of INSERT-seq toward illuminating these contributions.

At divergent mammalian promoters, RNAPII transcribes a protein-coding messenger RNA (mRNA) and non-coding upstream antisense RNA (uaRNA) in close proximity to one another. Whereas mRNAs are long, precisely processed and stable, uaRNAs are generally <2 kb in length, unspliced, non-polyadenylated and unstable[1]. RNAPII initiates from indistinguishable sequence elements on either side of the shared nucleosome-depleted region and the same transcription machinery is present at both sense and antisense promoters[2], making the contrast in transcript fate particularly striking. This dichotomy suggests that differences arise during RNA elongation and lie within the transcribed sequences.

While our understanding of sequence effects on RNAPII transcription elongation is incomplete, some important sequence elements have been described. Best known is the polyadenylation signal (PAS), which consists of a PAS hexamer (canonically AAUAAA[3]) and several auxiliary elements that are recognized by the cleavage and polyadenylation (CPA) machinery[4]. PASs mediate the cleavage of most full-length mRNAs, and influence a fraction of premature termination in mRNAs and non-coding RNAs[5–8]. Splice site motifs have also been suggested to affect transcription, with various reports implicating splicing in stimulation of transcription initiation, pause-release, or the suppression of premature termination[9–14]. Notably, a splicing-independent role for the 5′ splice site (5′SS) has been suggested[6,8,10,15–17], and the U1 snRNP (U1) binds the 5′SS in nascent RNA shortly after its synthesis, raising the possibility that association of U1 with elongating RNAPII can impact transcription elongation[18]. However, prior approaches to dissect the function(s) of 5′ SSs have investigated a small number of introns[10–12,17] or employed global splicing inhibitors[6,8,13,15,16], which cause cellular toxicity and lead to indirect or off-target effects. Moreover, most analyses have focused narrowly on whether the presence of a 5′SS suppresses premature termination at cryptic PASs in a process called 'telescripting'[15,16], leaving open the question of a more general role. Thus, whether RNA splicing and/or 5′SSs broadly promote transcription remains to be systematically evaluated.

Results

We developed INSERT-seq to decipher how transcribed sequence affects gene expression. A fluorescent reporter sequence was integrated at a non-coding uaRNA locus in mouse embryonic stem cells (mESCs, Fig. 1a), and a library of thousands of sequences was introduced between the transcription start site (TSS) and reporter (Fig. 1b). The library comprised a repertoire of sequences from coding and non-coding transcripts and synthetic sequences (Supplementary Tables 1 and 2) designed to fully explore the relationship between sequence composition and expression.

Effects of inserted sequences on RNA and protein levels were read out in high-throughput assays (Fig. 1b and Supplementary Table 3). In RNA assays, the relative enrichment of each insert in cDNA was determined, as compared to genomic DNA. cDNA was generated using a reverse primer that anneals 221 nt downstream of the inserted sequences, such that processive transcription through an insert is required for the insert to be counted by sequencing. Initiation events occurring within inserts will not lead to inserts being counted. Importantly, both steady-state RNA and nascent RNA (described below, Fig. 2c) can be evaluated in this assay. Effects of inserted sequences on expression of the reporter protein encoded downstream were measured using Sort-seq[19,20]: the pool of cells was sorted into six bins on the basis of the level of red fluorescence relative to a control fluorophore (see Methods) and the inserts present in genomic DNA from each bin were sequenced. For each insert, the distribution over the bins was used to calculate a Sort-seq score representing low (1) to high (6) protein abundance. Both screening methods showed good agreement between biological replicates (Extended Data Fig. 1a).

Department of Biological Chemistry and Molecular Pharmacology, Blavatnik Institute, Harvard Medical School, Boston, MA, USA.
✉e-mail: Hanneke_Vlaming@hms.harvard.edu; Karen_Adelman@hms.harvard.edu

5′-end sequence contributes to transcript expression. First, we asked whether the composition of the transcribed sequence affects its own transcription. To this end, the library included TSS-proximal regions from different transcript classes. As a negative control, mRNA termination regions were included, which exhibited low abundance (Fig. 1c,d), as expected. TSS-proximal regions from mRNAs supported the highest levels of RNA and protein, followed by long intergenic non-coding RNAs (lincRNAs), uaRNAs, and enhancer RNAs (eRNAs) (Fig. 1c,d). This result is consistent with the endogenous elongation potential of these transcript classes and suggests that the composition of the TSS-proximal sequence contributes to transcription regulation. Nonetheless, despite significant differences between transcript classes (Fig. 1c,d), the distributions of RNA levels overlap, with some highly abundant eRNA sequences observed and, conversely, some mRNA sequences detected at low levels. These findings are consistent with the many similarities observed for transcription of coding and non-coding RNAs[21]. Next, we tested whether functional sequence elements were limited to the most TSS-proximal regions or could be found equally in TSS-distal regions (160–333 bp from the TSS) (Fig. 1e and Extended Data Fig. 1b). Interestingly, TSS-proximal regions supported higher abundance than their distal counterparts. Thus, we suggest that the mammalian genome has evolved to contain signals within the initially transcribed region that directly influence elongation potential.

We specifically evaluated sequences from typical enhancers (TE) and super enhancers (SE)[22], because it has been proposed that transcription termination is particularly important within SEs to prevent DNA damage caused by collisions between closely spaced, convergently transcribing RNAPIIs[23]. In agreement with this idea, we observed that eRNA sequences from SEs are consistently more negative than those in TEs (Fig. 1f and Extended Data Fig. 1c), suggesting that eRNA sequences themselves contribute to efficient termination observed in SEs.

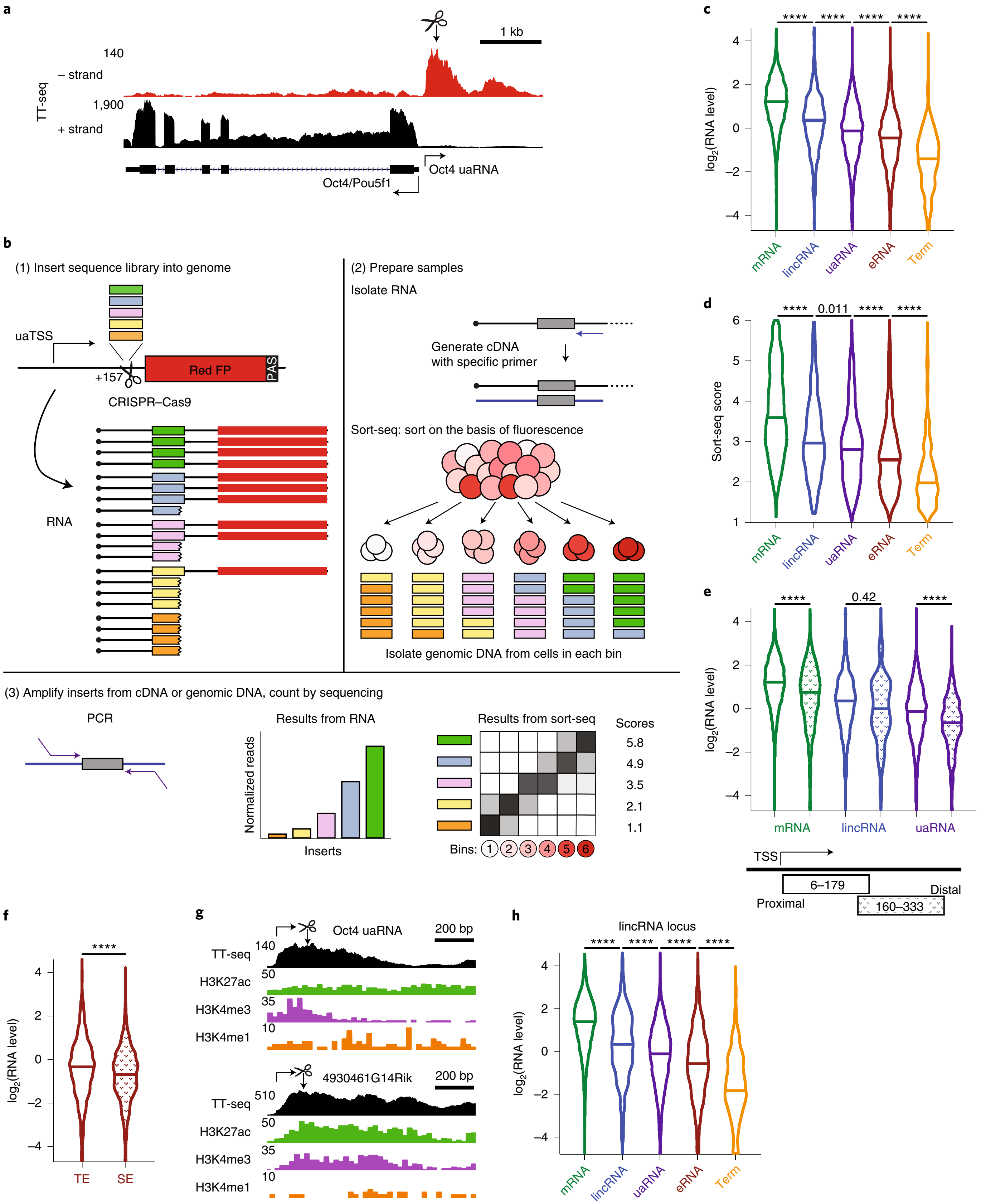
Effect of the transcribed sequence is independent of context. To test whether these conclusions were generalizable, we repeated the RNA-level screen after integrating the sequence library downstream of the 4930461G14Rik lincRNA TSS. We selected this locus because this lincRNA is many kb long, highly transcribed and, in comparison with the uaRNA locus, displays higher levels of histone H3 acetylated at K27 (H3K27ac), a larger domain of H3 trimethylated at K4 (H3K4me3), and less H3K4me1 (Fig. 1g). Screening the library at this lincRNA locus, without introduction of a fluorescent reporter, revealed a strong correlation with the data obtained at the uaRNA locus (Fig. 1h, Extended Data Fig. 1d, and Supplementary Table 4). We conclude that the signals contained in transcribed sequences can be conveyed independent of chromatin context or promoter identity.

Transcribed sequence directly affects transcription levels. Since uaRNAs and eRNAs have been reported to be unstable due to rapid degradation by the exosome complex[2,24–26], we tested whether the observed differences in RNA abundance could be explained by RNA stability. First, we repeated the RNA screen with the library at the uaRNA locus after knockdown of the exosome subunit EXOSC3 (RRP40; Extended Data Fig. 2a). EXOSC3 depletion resulted in a general stabilization of transcripts from the uaRNA reporter locus (Extended Data Fig. 2b). However, using internal normalization of INSERT-seq data, we found that the relative levels of inserts were unchanged after EXOSC3 knockdown (Fig. 2a,b), with mRNA regions remaining the most positive and eRNAs the most negative (Fig. 2b and Supplementary Table 5). Second, we isolated nascent RNA from the library-containing pool of cells using a modified version of PRO-seq[27] and used this in an RNA-based screen. Consistently, we found the pattern of mRNA > lincRNA > uaRNA > eRNA in insert abundance in the nascent RNA (Fig. 2c), and a good correlation between nascent and steady-state RNA levels (Extended Data Fig. 2c). Agreement was also observed when performing INSERT-seq using chromatin-associated RNA (Extended Data Fig. 2d,e), which is enriched in nascent and recently synthesized transcripts and thus is less sensitive to effects of RNA stability. We conclude that the observed effects are largely independent of RNA stability and that the initially transcribed sequence directly affects transcription. Our data support that this regulation occurs predominantly at the level of transcription elongation, because INSERT-seq selectively measures productive transcript formation; however, contributions of other mechanisms cannot be excluded.

GC content inherently affects transcriptional output. mRNAs are more GC-rich than average genomic regions, especially at the 5′ ends, which often overlap with CpG islands[1]. The GC content of non-coding RNAs is closer to the genome average of 42% G/C in mice[28] (Extended Data Fig. 3a). To test whether GC content influences expression, we assessed the effects of ~1,000 synthetic controls that were present in the screened library of insert sequences. These synthetic controls were generated over a range of GC contents and, notably, devoid of the most common PAS hexamers present in canonical termination sequences. We observed that inserts with higher GC content had a higher abundance in all assays (Fig. 3a and Extended Data Fig. 3b,c), indicating a striking impact of GC content on RNA production. We tested whether this relation was due to enrichment of CpG dinucleotides specifically[29], but

Fig. 1 | INSERT-seq demonstrates the role of transcribed sequences in gene regulation. a, Transient transcriptome sequencing (TT-seq) data at the *Oct4* (*Pou5f1)* locus in mESCs, measuring newly synthesized RNA. As described in the Methods, a red fluorescent reporter was integrated into one allele, 172 bp from the TSS, as indicated by the arrow and scissors. **b**, A schematic of INSERT-seq. At the reporter locus, a library of inserts with 173-bp variable regions was introduced by CRISPR–Cas9. To measure the effects of each insert on RNA abundance, RNA (steady-state or nascent) was isolated from the pool of cells, inserts were amplified from cDNA (made with a gene-specific primer) and counted by next-generation sequencing. The abundance of inserts in RNA was normalized to abundance in genomic DNA. Protein expression was measured by sorting cells on the basis of their level of red fluorescence (see Methods) and the abundance of inserts in genomic DNA from each fluorescence bin was used to calculate a Sort-seq score. **c**,**d**, Steady-state RNA (**c**) and Sort-seq (protein; **d**) results of inserts containing TSS-proximal genomic regions and mRNA terminators. Violins show frequency distribution and median. mRNAs $n=3{,}867$, lincRNAs $n=342$, uaRNAs $n=1{,}743$, eRNAs $n=2{,}106$, mRNA terminators (Term) $n=416$. Comparisons between neighbors by Kruskal–Wallis test. **e**, Steady-state RNA levels of inserts containing TSS-proximal (same as **c**) and TSS-distal genomic regions of indicated RNA classes. For TSS-distal regions, mRNAs $n=944$, lincRNAs $n=98$, uaRNAs $n=557$. Comparisons by Kruskal–Wallis test. **f**, Steady-state RNA levels of inserts containing TSS-proximal regions from typical enhancers (TE, not overlapping super enhancers, $n=1{,}506$) and super enhancers (SE, defined by ref. [22], $n=600$), compared with two-sided Mann–Whitney test. **g**, Snapshots of TT-seq and ChIP–seq data at the *Oct4* uaRNA and 4930461G14Rik lincRNA loci. See Methods for data sources. Transcription start sites and cut/integration sites are indicated; at the lincRNA locus the library was integrated 102 bp downstream of the TSS. **h**, Steady-state RNA levels after library integration at the 4930461G14Rik lincRNA locus, showing the same inserts classes as in **c**. mRNAs $n=3{,}987$, lincRNAs $n=381$, uaRNAs $n=1{,}853$, eRNAs $n=2{,}290$, mRNA terminators $n=458$. All comparisons between neighbors are significant by Kruskal–Wallis test. For all panels, ****$P<0.0001$ and all higher P values are indicated in the plots.

found that CpG content was not a better predictor of RNA level than overall GC content (Pearson correlation coefficient of 0.47 in both cases, Fig. 3a and Extended Data Fig. 3d). We therefore conclude that early-transcribed regions enriched in G and C nucleotides support higher levels of processive transcription than sequences with high AT content.



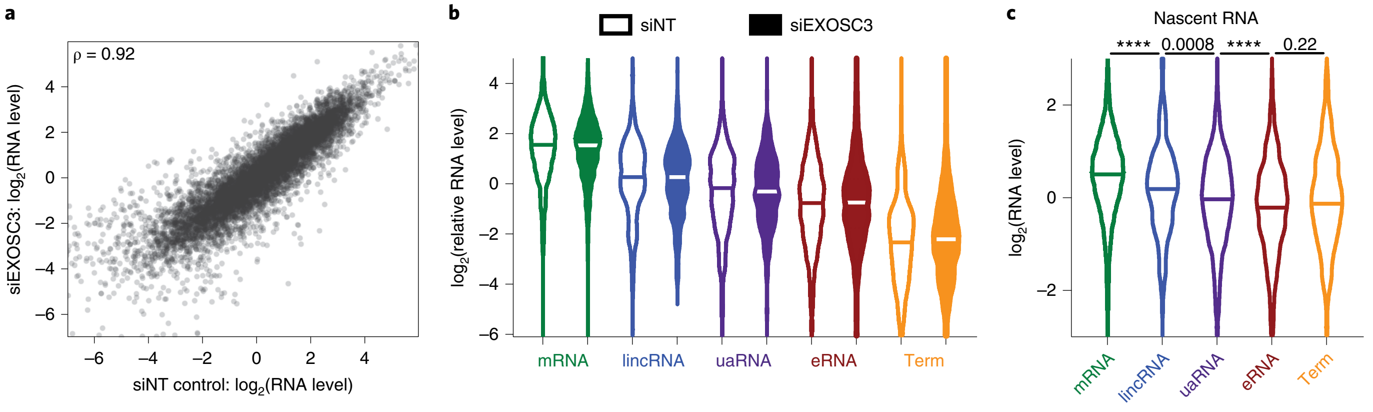


Fig. 2 | Transcribed sequence directly affects transcription levels. **a**, Correlation between relative abundance of inserts in steady-state RNA from cells 48 hours after transfection with a non-targeting control siRNA (siNT) or siRNAs targeting the exosome subunit EXOSC3. Plotted are all the insert classes used for Fig. 1, as well as synthetic control sequences ($n = 9,539$ total). Each data point is the mean of three replicate experiments. **b**, Steady-state RNA results with library at uaRNA locus, comparing control siRNA (siNT, open) to exosome depletion (siEXOSC3, filled). Data sets were internally normalized to control sequences (see Methods), such that this plot shows a lack of relative changes in RNA levels. mRNAs $n = 3,356$, lincRNAs $n = 304$, uaRNAs $n = 1,480$, eRNAs $n = 1,749$, mRNA terminators $n = 333$. **c**, Nascent RNA results with library at uaRNA locus. Nascent RNA was isolated after biotin-ribonucleoside triphosphate (rNTP) run-on (Methods). Groups are as in Fig. 1c. Neighbors were compared by Kruskal–Wallis test; ****$P < 0.0001$, and higher P values are indicated in the plot.

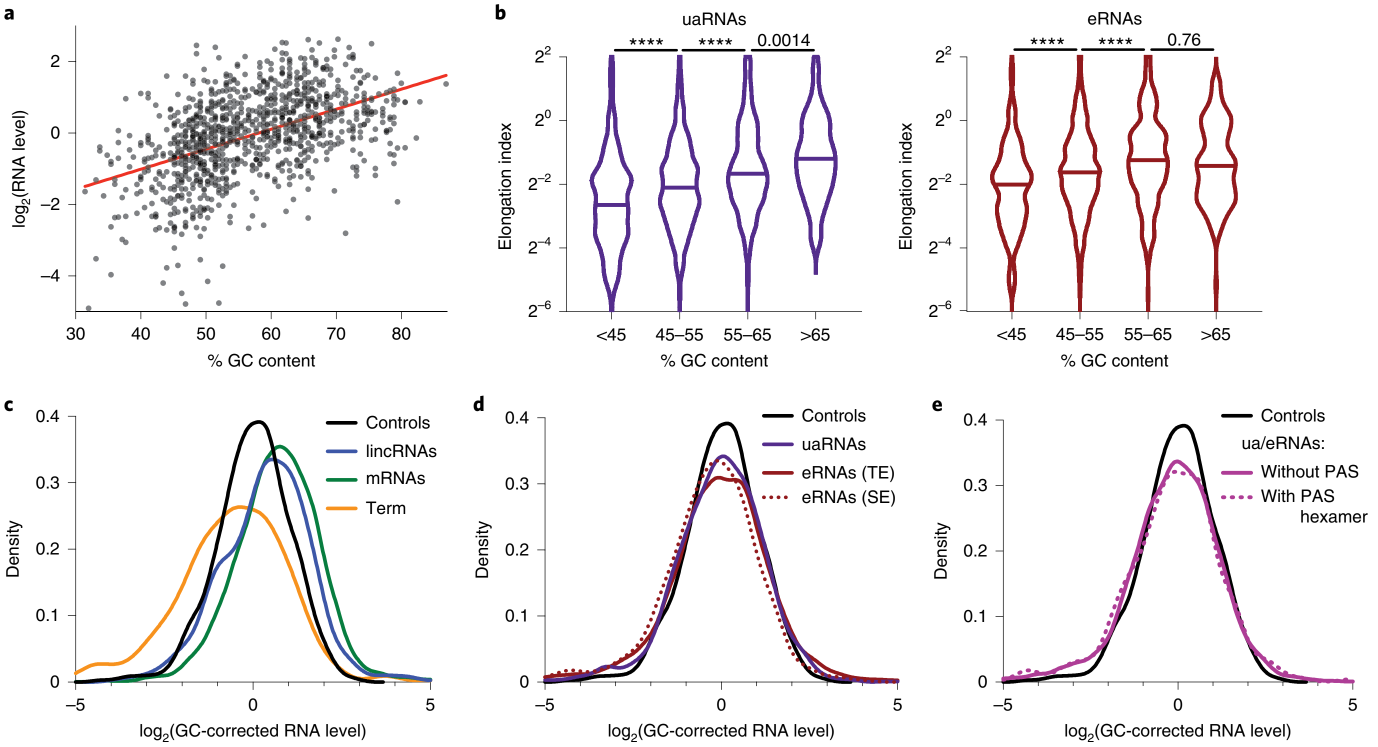


Fig. 3 | GC content inherently affects transcriptional output. **a**, Relation between GC content of synthetic control sequences and their steady-state RNA levels ($n = 1,059$). The red line is the best linear fit through the data. Pearson $r = 0.47$, $P < 0.0001$. **b**, Elongation indices from mESC PRO-seq data at endogenous genomic locations of uaRNA (left) and eRNA (right) sequences tested in INSERT-seq. The elongation index was calculated as the ratio of PRO-seq density corresponding to elongating RNAPII (window from +50 to +250 downstream of the TSS) divided by the promoter-proximal RNAPII density (from the TSS to +50). Higher values indicate more efficient elongation. Loci were grouped by the GC content of their TSS-proximal regions (TSS + 6 to +179), as included in the library, comparisons by Kruskal–Wallis test, **** indicates $P < 0.0001$, higher P values are indicated in the plot. **c**, Density plot of steady-state RNA abundance levels corrected for the predicted abundance based on the GC content (based on the fit line in panel a) by subtraction. Groups as in Fig. 1c. All three groups differ from the controls by Kruskal–Wallis test, $P < 0.0001$. **d**, Like b, but showing TSS-proximal regions from uaRNAs ($n = 1,743$) and eRNAs from typical enhancers (TE; $n = 1,506$) and super enhancers (SE; $n = 600$) alongside controls. Only SE eRNA regions were significantly different from the controls by Kruskal–Wallis test, $P < 0.0001$. **e**, Density plot of GC-corrected steady-state RNA levels of all TSS-proximal and TSS-distal uaRNA and eRNA regions, grouped by the presence of a canonical PAS hexamer (AWTAAA, where W = A/T). Controls $n = 1,059$, without PAS $n = 3,895$, with PAS $n = 511$. The groups with and without PAS hexamer are not significantly different by Mann–Whitney test ($P = 0.69$).

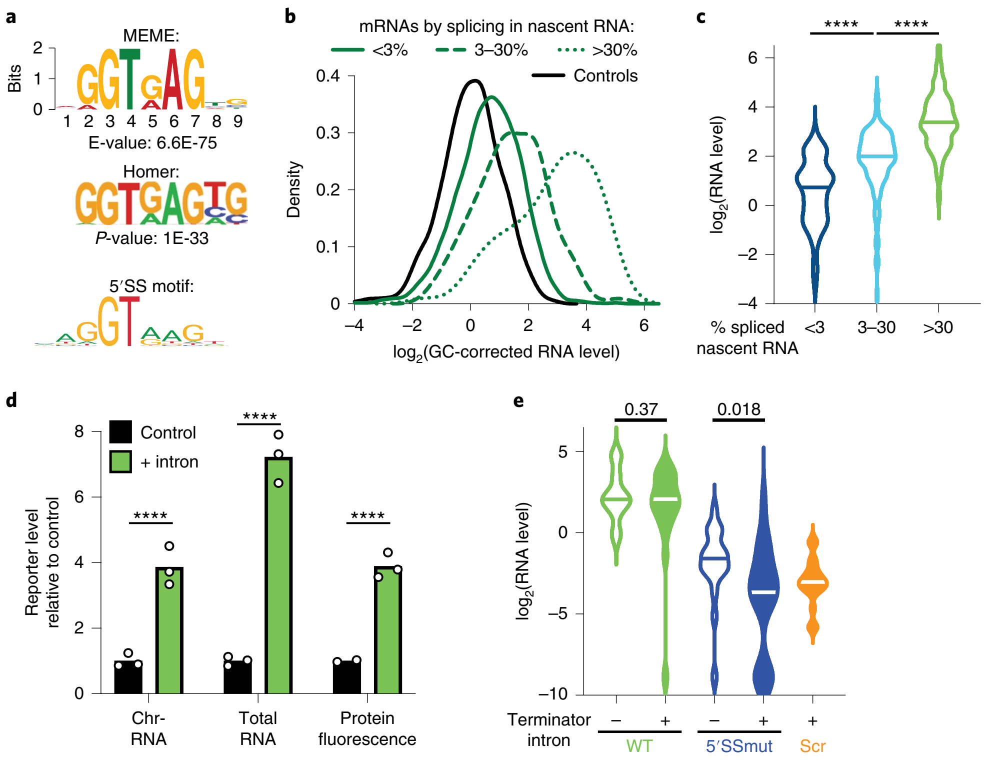


Fig. 4 | Co-transcriptionally spliced introns boost transcription. a, De novo motif identified by MEME and Homer algorithms to be most significantly enriched in the top 10% over the bottom 50% of TSS-proximal mRNA regions by GC-corrected steady-state RNA level. 5′SS motif from the JASPAR database (SD000.1) is shown for comparison. **b**, Density plot of GC-corrected steady-state RNA levels of TSS-proximal mRNA regions grouped by splicing efficiency (spliced/total). Controls $n = 1{,}059$, <3% spliced $n = 3{,}687$, 3-30% spliced $n = 102$, >30% spliced $n = 78$. To focus on co-transcriptional splicing, splicing efficiency for **b** and **c** was calculated using nascent RNA data. All groups differ significantly from each other by Kruskal–Wallis test, $P < 0.0001$. **c**, Steady-state RNA levels of inserts containing wild-type introns (unbarcoded) grouped by splicing efficiency measured using the nascent RNA screen data. <3% spliced $n = 76$, 3-30% spliced $n = 107$, >30% spliced $n = 198$, significance tested by Kruskal–Wallis test. **d**, Reporter level measured in clonal mESC lines with an intron from the mouse *Atp5a1* gene inserted upstream of the fluorescent reporter at the uaRNA locus, compared with clonal control lines (CRISPR clones without intron integration). Effects in chromatin-associated (Chr) and steady-state (total) RNA were determined by RT–qPCR and normalized to an internal control gene (*TBP*). Fluorescence was measured by flow cytometry. Comparisons were made using two-way ANOVA and corrected with the Šidák method. $n = 3$ independently generated cell lines per genotype, though fluorescence was measured for only 2 of the control lines. **e**, Steady-state RNA levels of variations of a subset of introns. Introns without (open) and with (filled) a strong PAS terminator[40] replacing part of the intronic sequence were compared in otherwise wild-type introns and 5′SS mutant (5′SSmut) introns (each $n = 19$). For reference, strong PAS terminators were inserted in a background of scrambled (scr) introns ($n = 18$). Comparisons by two-sided Wilcoxon tests. For all panels, ****$P < 0.0001$, and all higher P values are indicated in the plots.

We corroborated these results using PRO-seq data from wild-type mESCs to look at transcriptionally engaged RNAPII over the sequences included in the INSERT-seq library. For uaRNA and eRNA loci included in our screen, we found that transcribed regions with higher GC content exhibited elevated levels of RNAPII (Extended Data Fig. 3e). Furthermore, with increasing GC content, these loci display a higher elongation index (downstream over upstream PRO-seq signal), indicative of more efficient transcription elongation (Fig. 3b). Thus, even in the genomic context, where numerous factors regulate transcriptional output, we find that a high GC content in the initially transcribed region promotes processive transcription.

To probe to what extent GC content explained the behavior of genomic sequences in our library, we used the relation between GC content and abundance to 'correct' the abundance of all inserts. After GC-correction, mRNA and lincRNA 5′ sequences remained positive and mRNA 3′ ends remained negative compared with controls (Fig. 3c). However, the RNA levels of uaRNAs and typical enhancer (TE) eRNAs were indistinguishable from controls after GC-content correction (Fig. 3d). Only eRNAs from SEs showed a small but significant shift towards lower abundance (Fig. 3d), supporting the idea that these sequences have evolved to induce more efficient transcription termination.

Notably, our data imply that early termination of eRNAs and uaRNAs does not rely on specific sequence motifs. In fact, the presence of PAS hexamers in uaRNAs and eRNAs caused no additional reduction in RNA level in our screen (Fig. 3e). We conclude that the high AT content of the genomic regions from which most uaRNAs and eRNAs arise inherently causes RNAPII to be termination-prone, and they have not evolved to enrich for specific motifs. Indeed, high AT content increases RNAPII pausing and backtracking[30], because AT-rich RNAs fail to form stable secondary structures that prevent RNAPII backtracking[31,32] and AT-rich RNA–DNA hybrids destabilize the ternary elongation complex. Consequently, T-rich sequences promote termination by several RNA polymerases[33–36]. Thus, we propose that frequent RNAPII stalling and backtracking within AT-rich non-coding regions increases the chance of early termination, as it does downstream of a PAS at mRNA 3′ ends[37].

Co-transcriptionally spliced introns boost processive transcription. We then focused on mRNA TSS-proximal regions. Searches for sequence motifs that could convey positive signals in the most

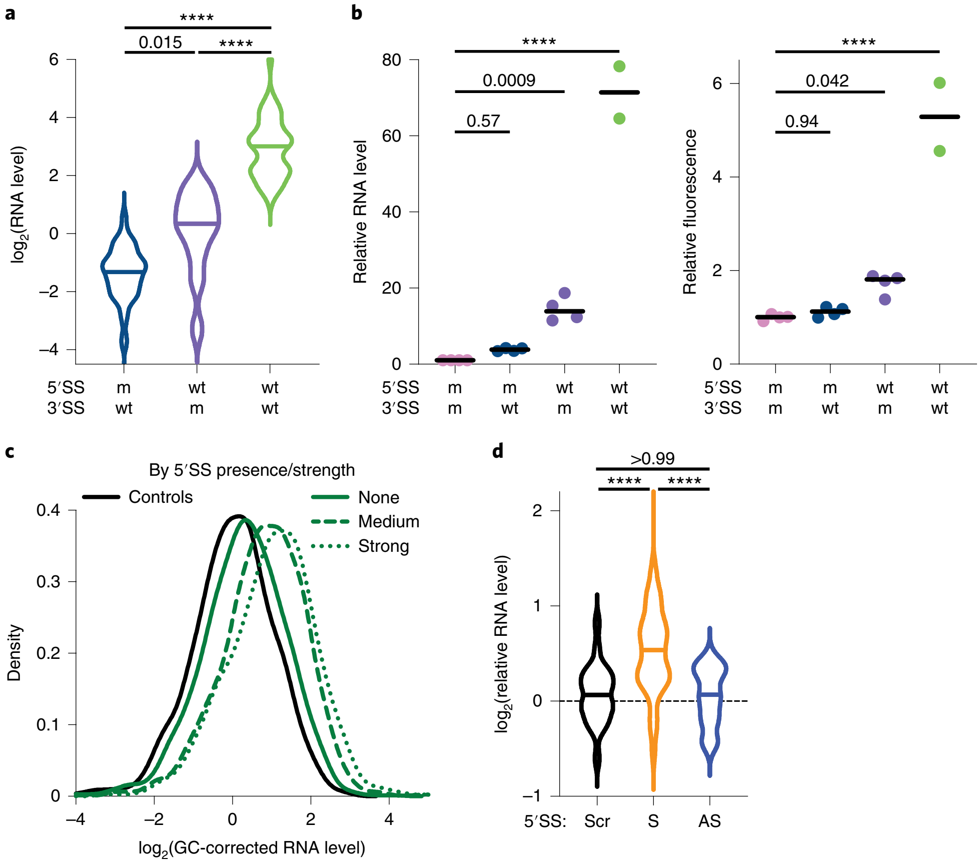


Fig. 5 | Splicing-dependent and splicing-independent role of the 5′SS. a, Steady-state RNA levels of intron-containing inserts with wild-type (wt) or mutant (m) splice sites. Only introns are shown of which the wild-type version was >30% spliced and mutants <3% spliced in nascent RNA. 5′SS mutants $n = 51$, 3′SS mutants $n = 23$, wt $n = 52$, comparisons by Kruskal–Wallis test. **b**, Reporter RNA level (*Tbp*-normalized RT–qPCR; left) and fluorescence measured by flow (right) in different clonal cell lines containing integrated versions of an intron from the *Smc1* gene. Splice sites were either wild-type (wt) or mutated (m), and results were normalized to the double mutant. Comparisons made by one-way ANOVA. **c**, Density plot of GC-corrected steady-state RNA levels of unspliced TSS-proximal mRNA regions grouped by the presence and strength (MaxEnt score[41]) of a 5′SS motif (see Methods). None $n = 1{,}392$, medium (MaxEnt 5–10) $n = 1{,}792$, strong (MaxEnt 10+) $n = 503$. Both 5′SS-containing groups are significantly different from the inserts without a 5′SS motif ($P < 0.0001$) and from each other ($P = 0.0009$) by Kruskal–Wallis test. **d**, Relative steady-state RNA levels of 10-nt annotated 5′SSs with a MaxEnt score of >5, embedded into several background sequences. Only unspliced inserts (<3% spliced in nascent RNA) were considered. Scrambled (scr, $n = 24$) and antisense (AS, $n = 24$) versions of 5′SSs were compared with sense (S) 5′SSs ($n = 50$) by Kruskal–Wallis test. For all panels, ****$P < 0.0001$, and all higher P values are indicated in the plots.

abundant TSS-proximal mRNA inserts identified the same top hit using MEME[38] and Homer[39]: a 5′ splice site (5′SS) motif (Fig. 4a). Splicing, and the 5′SS in particular, have been suggested to stimulate RNA production[8–17], and the mechanisms underlying this activity are a subject of great interest[18]. Thus, we used INSERT-seq to systematically assess the effect of intron sequences on RNA production. We first measured the co-transcriptional splicing efficiency of inserts (using nascent RNA, see Methods) and found that, among mRNA TSS-proximal regions, efficiently spliced inserts were most abundant (Fig. 4b). Further, in inserts containing annotated mouse introns, splicing efficiency also correlated with abundance (Fig. 4c). This result was consistent across all screen read-outs (Extended Data Fig. 4 and Supplementary Table 6). Moreover, clonal cell lines containing an intron at this locus displayed significantly elevated RNA and protein expression as compared with intron-less controls (Fig. 4d). Importantly, as described below, analysis of splice site mutants excludes the possibility that this stimulation is caused by intronic enhancers. We note that the locus where the sequences were integrated contains no PAS hexamers, thus the stimulation caused by spliced introns does not rely on the suppression of PAS-mediated transcript cleavage or termination.

Previous work nonetheless indicates that a 5′SS suppresses the recognition of a downstream cryptic PAS[6–8,15–17], and prior study of a strong PAS-containing terminator in a plasmid context showed that PAS recognition can be inhibited inside an intron[40]. To rigorously test this model, we inserted the strong 49-nt PAS-containing terminator[40] inside a variety of introns. If the terminator leads to RNA cleavage inside an insert, that will prevent amplification and lead to low abundance in the sequencing library. While addition of the terminator sequence into wild-type splicing-competent introns had no effect on RNA levels (Fig. 4e), mutation of the 5′SS enabled the terminator to significantly reduce transcript abundance. In fact, RNA levels dropped to the same level as that of inserts with a terminator in a scrambled sequence background (Fig. 4e). Thus, while the PAS terminator can robustly induce termination at this locus, this property is fully masked when the terminator is within a spliced intron. To our knowledge, these data are the first to demonstrate in a genomic context that introns can broadly suppress the usage of a strong, canonical PAS terminator.

Distinguishing splicing-dependent and splicing-independent effects of 5′ splice sites. Because splicing efficiency correlated with insert abundance, we sought to establish a causal link using splice-site mutants that abrogated splicing in efficiently spliced wild-type introns (see Methods). Small (1- and 3-nt) mutations in either the 5′SS or 3′SS greatly reduced RNA and protein levels (Fig. 5a and Extended Data Fig. 5a), indicating that it is the process of splicing and not the sequence composition inside introns that promotes transcription.

Notably, 5′SS mutations had greater effects on RNA abundance than 3′SS mutations (Fig. 5a), consistent with the 5′SS having a splicing-independent role in stimulating transcription. The stronger effect of the 5′SS mutation was preserved when considering only inserts that did not contain PAS hexamers (Extended Data Fig. 5b), demonstrating that the splicing-independent 5′SS function does not rely on suppressing PAS-mediated transcript cleavage or termination.

We confirmed the effects of both 5′SS and 3′SS mutations in clonal cell lines that had a wild-type or mutant intron integrated at the reporter locus. Compared with cell lines in which both SSs were mutated, restoring just the 3′SS had no effect on RNA or protein abundance (Fig. 5b). Restoring the 5′SS alone significantly increased the reporter level, while the effect of restoring both splice sites, and thereby splicing, was much larger (Fig. 5b and Extended Data Fig. 5c). This experiment supports the conclusion that the 5′SS has a splicing-independent role in stimulating transcription. Importantly, the double mutant behaves similarly to the 5′SS mutant, implying that the 3′SS provides a positive signal only in the context of a spliced intron and that the process of splicing stimulates transcription. We note that the process of splicing could strengthen the splicing-independent role of the 5′SS, or the splicing-dependent and splicing-independent stimulation could function through independent mechanisms.

An autonomous role for the 5′ splice site sequence. If the 5′SS promotes processive transcription in a splicing-independent manner, it should do so in other sequence contexts. To test this, we searched screened TSS-proximal sequences for matches to the 5′SS consensus sequence (see Methods). Considering only unspliced or very lowly spliced inserts, we found that regions containing a strong 5′SS (predicted by the MaxEntScan algorithm[41]) are more abundant than those without a 5′SS in the context of both mRNAs and uaRNA/eRNAs (Fig. 5c, Extended Data Fig. 5d, and Supplementary Table 8). This shows that a 5′SS sequence can stimulate transcription even in a context where it is not used as a splice site. More than one 5′SS motif within our 173-bp variable regions, however, did not strongly stimulate transcription further (Extended Data Fig. 5e).

To further test the independence from sequence context, we determined whether 5′SSs could stimulate processive transcription in a random background sequence. To this end, strong 5′SS sequences (10 bp) from 50 introns were embedded in 5 random background sequences. The presence of a 5′SS in the sense orientation led to significant, consistent increases in expression, while the same sequences in the antisense orientation did not (Fig. 5d and Extended Data Fig. 5f). Notably, the orientation dependence points to the 5′SS sequence being recognized in the RNA rather than DNA context, likely by U1. We conclude that regardless of surrounding sequence, recognition of a 5′SS in the initially transcribed RNA by U1 snRNP promotes RNAPII elongation potential[18].

Discussion

We developed INSERT-seq to compare the effects of thousands of integrated sequences on nascent RNA, steady-state RNA and protein levels. This elucidated several features of transcription regulation by the initially transcribed sequence. By inserting all sequences at one genomic locus, we isolated the effects of sequence from confounding factors, such as variable promoter strengths and chromatin contexts. This allowed us to study causal relationships in a manner that is impossible at endogenous loci. From our results, we draw several central conclusions.

First, the GC content of the initially transcribed sequence inherently affects transcriptional output. In randomly generated sequences, higher GC content yields higher transcription levels and elevated RNA abundance. This likely results from high rates of RNAPII pausing and backtracking on less thermodynamically stable AT-rich sequences[31,32]. We propose that RNAPII is particularly susceptible to pausing and termination during early elongation, before association of elongation factors that make RNAPII optimally processive. While it is known that GC content can influence steady-state RNA accumulation[42] and is different between coding and non-coding regions, to our knowledge GC-richness in the initially transcribed region has not been previously proposed to contribute to differences in transcription levels among RNA species.

Second, there is no evidence that uaRNA or eRNA sequences have evolved to contain specific signals that govern RNAPII elongation. The effects of uaRNA and eRNA sequences can be predicted solely on the basis of their high AT content (with some exceptions, for example in super enhancers). Thus, we envision RNAPII being generally termination-prone when transcribing AT-rich intergenic space, rather than recognizing a specific termination signal. Importantly, this model predicts termination over a diffuse region, which is consistent with the observation that both eRNAs and uaRNAs have highly heterogeneous 3′ ends[2].

Third, mRNA 5′ ends contain positive signals, and the strongest of these are related to splicing. 5′SSs stimulate transcription through both splicing-dependent and splicing-independent mechanisms. Remarkably, a 5′SS can suppress the usage of strong PAS terminators, validating and extending the telescripting model[15]. However, the effect of 5′SSs goes beyond this role, to stimulate transcription more broadly. We propose that changes in RNAPII conformation that occur upon U1 binding make it intrinsically more processive and/or prevent association of other termination machineries, such as Integrator[2,43] and WDR82/ZC3H4 (refs. [44,45]).

From an evolutionary perspective, it is interesting to note that features favoring processive elongation by RNAPII are highly enriched at the 5′ ends of mRNAs. Protein-coding genes often include a 5′SS near the TSS (~50% have a 5′SS in the first 179 nt), which would promote rapid loading of U1. Furthermore, mammalian promoters are often embedded within CpG islands (79% in mouse, 92% in human), conferring a high GC content to the 5′ ends. Accordingly, we found that TSS-proximal sequences (TSS + 6 to +179) were more positive than sequences slightly downstream (TSS + 160 to +333; Fig. 1e), particularly for mRNAs. We suggest that these characteristics contribute to the processive nature of RNAPII at mRNAs and the ability to counteract termination machineries. Providing a molecular mechanism for such sequence-mediated transcription stimulation and identifying additional sequence elements that regulate processive transcription will be important future work. Our newly developed methodology, INSERT-seq, can be leveraged toward these goals.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41594-022-00785-9.

Received: 4 July 2021; Accepted: 28 April 2022;
Published online: 9 June 2022

References

1. Lykke-Andersen, S. et al. Integrator is a genome-wide attenuator of non-productive transcription. *Mol. Cell* **81**, 514–529.e6 (2021).

2. Scruggs, B. S. et al. Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Mol. Cell* **58**, 1101–1112 (2015).
3. Tian, B., Hu, J., Zhang, H. & Lutz, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* **33**, 201–212 (2005).
4. Shi, Y. & Manley, J. L. The end of the message: multiple protein-RNA interactions define the mRNA polyadenylation site. *Genes Dev.* **29**, 889–897 (2015).
5. Ntini, E. et al. Polyadenylation site–induced decay of upstream transcripts enforces promoter directionality. *Nat. Struct. Mol. Biol.* **20**, 923–928 (2013).
6. Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360–363 (2013).
7. Core, L. J. et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).
8. Chiu, A. C. et al. Transcriptional pause sites delineate stable nucleosome-associated premature polyadenylation suppressed by U1 snRNP. *Mol. Cell* **69**, 648–663 (2018).
9. Le Hir, H., Nott, A. & Moore, M. J. How introns influence and enhance eukaryotic gene expression. *Trends Biochem. Sci.* **28**, 215–220 (2003).
10. Damgaard, C. K. et al. A 5′ splice site enhances the recruitment of basal transcription initiation factors in vivo. *Mol. Cell* **29**, 271–278 (2008).
11. Bieberstein, N. I., Carrillo Oesterreich, F., Straube, K. & Neugebauer, K. M. First exon length controls active chromatin signatures and transcription. *Cell Rep.* **2**, 62–68 (2012).
12. Fiszbein, A., Krick, K. S., Begg, B. E. & Burge, C. B. Exon-mediated activation of transcription starts. *Cell* **179**, 1551–1565(2019).
13. Sousa-Luís, R. et al. POINT technology illuminates the processing of polymerase-associated intact nascent transcripts. *Mol. Cell* **81**, 1935–19502021).
14. Caizzi, L. et al. Efficient RNA polymerase II pause release requires U2 snRNP function. *Mol. Cell* **81**, 1920–1934.e9 (2021).
15. Kaida, D. et al. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**, 664–668 (2010).
16. Berg, M. G. et al. U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150**, 53–64 (2012).
17. Andersen, P. K., Lykke-Andersen, S. & Jensen, T. H. Promoter-proximal polyadenylation sites reduce transcription activity. *Genes Dev.* **26**, 2169–2179 (2012).
18. Zhang, S. et al. Structure of a transcribing RNA polymerase II–U1 snRNP complex. *Science* **371**, 305–309 (2021).
19. Kinney, J. B., Murugan, A., Callan, C. G. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl Acad. Sci. USA* **107**, 9158–9163 (2010).
20. Sharon, E. et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521–530 (2012).
21. Field, A. & Adelman, K. Evaluating enhancer function and transcription. *Annu. Rev. Biochem.* **89**, 213–234 (2020).
22. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
23. Flynn, R. A. et al. 7SK–BAF axis controls pervasive transcription at enhancers. *Nat. Struct. Mol. Biol.* **23**, 231–238 (2016).
24. Preker, P. et al. RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322**, 1851–1854 (2008).
25. Seila, A. C. et al. Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008).
26. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
27. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013).
28. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
29. Krinner, S. et al. CpG domains downstream of TSSs promote high levels of gene expression. *Nucleic Acids Res.* **42**, 3551–3564 (2014).
30. Noe Gonzalez, M., Blears, D. & Svejstrup, J. Q. Causes and consequences of RNA polymerase II stalling during transcript elongation. *Nat. Rev. Mol. Cell Biol.* **22**, 3–21 (2021).
31. Zamft, B., Bintu, L., Ishibashi, T. & Bustamante, C. Nascent RNA structure modulates the transcriptional dynamics of RNA polymerases. *Proc. Natl Acad. Sci.* **109**, 8948–8953 (2012).
32. Turowski, T. W. et al. Nascent transcript folding plays a major role in determining RNA polymerase elongation rates. *Mol. Cell* **79**, 488–503(2020).
33. Roberts, J. W. Mechanisms of bacterial transcription termination. *J. Mol. Biol.* **431**, 4030–4039 (2019).
34. Mishra, S. & Maraia, R. J. RNA polymerase III subunits C37/53 modulate rU:dA hybrid 3′ end dynamics during transcription termination. *Nucleic Acids Res.* **47**, 310–327 (2019).
35. Fouqueau, T. et al. The cutting edge of archaeal transcription. *Emerg. Top. Life Sci.* **2**, 517–533 (2018).
36. Davidson, L., Francis, L., Eaton, J. D. & West, S. Integrator-dependent and allosteric/intrinsic mechanisms ensure efficient termination of snRNA transcription. *Cell Rep.* **33**, 108319 (2020).
37. White, E., Kamieniarz-Gdula, K., Dye, M. J. & Proudfoot, N. J. AT-rich sequence elements promote nascent transcript cleavage leading to RNA polymerase II termination. *Nucleic Acids Res.* **41**, 1797–1806 (2013).
38. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
39. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–89 (2010).
40. Levitt, N., Briggs, D., Gil, A. & Proudfoot, N. J. Definition of an efficient synthetic poly(A) site. *Genes Dev.* **3**, 1019–25 (1989).
41. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).
42. Mordstein, C. et al. Codon usage and splicing jointly influence mrna localization. *Cell Syst.* **10**, 351–362.e8 (2020).
43. Elrod, N. D. et al. The integrator complex attenuates promoter-proximal transcription at protein-coding genes. *Mol. Cell* **76**, 738–752 (2019).
44. Austenaa, L. M. I. et al. A first exon termination checkpoint preferentially suppresses extragenic transcription. *Nat. Struct. Mol. Biol.* **28**, 337–346 (2021).
45. Estell, C., Davidson, L., Steketee, P. C., Monier, A. & West, S. ZC3H4 restricts non-coding transcription in human cells. *eLife* **10**, e67305 (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Mouse embryonic stem cell lines. All experiments were performed in the F121-9 line, a Castx129 female mouse hybrid ES cell line[46]. T2A-TagBFP2 was integrated using CRISPR–Cas9 at the location of the Oct4 stop codon by co-transfecting plasmids pHV123 and pKW09 (for all plasmids used in this study see Supplementary Table 8; for primers see Supplementary Table 9). All transfections in this study were done using suspension transfections with Lipofectamine 2000 (Thermo Fisher). After sorting for GFP$^+$ and TagBFP2$^+$ cells, a clonal line was selected with integration at only the 129 allele of the Oct4 C terminus, which was validated by Sanger sequencing single-nucleotide polymorphisms just outside the homology arms. In this line, mKate2 was inserted in the *Oct4* uaRNA, 172 bp downstream of the uaTSS, by co-transfecting pHV111 and pHV109. After sorting for mKate2$^+$ cells, a clonal line was selected with integration at only the 129 allele. These cell lines maintained expression of the red and blue reporter proteins, as well as good mESC morphology, over many passages.

Clonal lines with the ninth intron of the mouse *Atp5a1* gene (ENSMUST00000026495.15) inserted just upstream of the mKate2 reporter, 172 bp downstream of the *Oct4* uaRNA TSS, were generated by co-transfecting pHV114 and a linear double-stranded repair template. The repair template was generated with two steps of PCR from a DNA fragment (ordered from Twist Bioscience) containing the intron (including 3 nt each of the flanking exons) and homology to the uaRNA reporter locus. The first PCR introduced more homology to the locus, the second PCR introduced phosphorothioate bonds on both ends of the amplicon (see Supplementary Table 9 for primers). After transfection, GFP$^+$ cells were sorted, and three intron-containing clonal lines were identified by genotyping. Clones that went through the same procedure but had only a 1-bp insertion at the cut site were used as control lines. Wild-type and mutant versions of the 14th intron of *Smc1a* (ENSMUST00000045312.6) with a small part of the flanking exons were integrated upstream of mKate2 by co-transfecting pHV137 with pHV163/164/165/166, sorting for GFP$^+$ cells and genotyping individual clonal lines.

All mESC lines used for INSERT-seq were cultured in KO-DMEM medium (Gibco) supplemented with 15% KO Serum replacement (Gibco), GlutaMAX, penicillin–streptomycin, non-essential amino acids, beta-mercaptoethanol, 1000 U/ml LIF (Cell Guidance Systems), 1 µM MEK inhibitor (PD0325901; Stemgent), and 3 µM GSK3 inhibitor (CHIR99021; Stemgent). Cells used for ChIP–seq and TT-seq were grown in serum-free embryonic stem cell (SFES) medium with the same concentrations of LIF and inhibitors. SFES was composed of 50/50 NeuroBasal medium (Gibco) and DMEM/F12 medium (Gibco), supplemented with 0.5x B-27 (Gibco), 0.5× N-2 (Gibco), 2 mM L-glutamine (Gibco), 0.05% bovine albumin fraction V (Gibco) and 1.5×10^{-4} M monothioglycerol (Sigma).

Design of insert library. A library of 16,461 173-nt sequences was designed, primarily consisting of sequences derived from specific regions in the mouse genome (mm10). The largest class (10,774 sequences) were sequences just downstream of TSSs as defined by Start-seq data (GSE43390)[47], using GENCODE M20 annotations (removing confidence level 4/5 transcripts). TSS locations were refined from annotation using TSScall with a Start-seq 5′ read count threshold of 16, search window of 250 bp, and joining distance of 500 bp (https://github.com/lavenderca/TSScall). Using these TSSs, TSS-proximal (+6–+179) and TSS-distal regions (+160–+333) from different transcript classes were selected as described in Supplementary Table 1. In addition, 503 regions of annotated protein-coding transcript ends were included as described in Supplementary Table 1.

High-confidence introns were defined using GENCODE M20, selected to be between 50 and 158 nt in size, and filtered on the basis of evidence of splicing in mESCs (more details in Supplementary Table 1). The 173-nt sequences included in the library start 12 nt upstream of the 5′SS. For a subset of introns, multiple barcoded wild-type and mutant versions were added to the library. One-nucleotide and 3-nt mutations were made for both the 5′SS and 3′SS. A python script (https://github.com/audy/barcode-generator) was used to generate a set of 7-nt barcodes with a minimum editing distance of 5 and a maximum stretch of 3. These barcodes replaced the first 7 nt of each sequence.

Shorter sequences were placed in the context of the same five background sequences, see Supplementary Table 1. In these backgrounds, nucleotides 10 to 19 were replaced by 10-nt 5′SSs (−3 to +7). A subset of 5′SSs was also inserted in the antisense direction, or was scrambled (no GGT remained after scrambling).

As a control, a set of 1,100 completely randomly generated sequences was included, which covers a range of GC content that was comparable to the regions taken from the genome. Sequences containing AWTAAA, AGGTR, or an out-of-frame ATG were excluded.

Library cloning. The 173-nt sequences described above were flanked by common regions to allow for amplification of the whole library and sequencing using common Illumina sequencing primers: 5′-TTTCCCTACACGACGCTCTTCCGATCTA[N_{173}] TAGATCGGAAGAGCACACGTCTGAACTCC-3′. The reverse complement of these sequences was ordered as an oligonucleotide pool from Agilent Technologies, and was amplified for 10 cycles using NEBNext Ultra II Q5 (New England Biolabs). pHV141 (*Oct4* uaRNA) and pHV152 (lincRNA) were linearized using AflII, and the amplified library was inserted using Gibson assembly. The resulting DNA was concentrated and cleaned up using AMPure XP beads and electroporated into E. cloni 10 G ELITE electrocompetent cells (Lucigen). A small fraction of electroporated cells was plated to estimate the number of transformants (~1.5 M or more), the remainder was grown up overnight in liquid LB + Amp cultures, and plasmids were isolated.

Library integration. For integration at the *Oct4* uaRNA allele, first a gRNA was designed to target a site +157 nt downstream of the uaTSS, which overlays a SNP between 129 and CAST. Thus, this guide should only target the 129 allele in the hybrid mESCs. Using TIDE[48], the efficiency of this gRNA (pHV137) was measured to be ~75% on the 129 allele and negligible on the CAST allele. The gRNA plasmid (pHV137) and library-containing plasmid pool (pHV142) were co-transfected into the clonal cell line with the integrated TagBFP2 and mKate2 reporters at a large scale. At 48 hours after transfection, 6 million GFP$^+$ cells were sorted and propagated further. Genotyping suggested that integration had occurred in ~25% of this pool of cells. During propagation, care was taken to maintain library representation by always passaging >12 million cells.

For integration at the 4930461G14Rik lincRNA allele, the same approach was used, but this locus did not include a fluorescent reporter, and the gRNA in pHV149 targeted TSS + 102 specifically at the CAST allele. pHV149 and the library-containing plasmid pool (pHV154) were co-transfected into wild-type F121-9 cells. At 48 hours post-transfection, 5.5 million GFP$^+$ cells were sorted and propagated further. Genotyping suggested that integration had occurred in ~33% of this pool of cells. During propagation, care was taken to maintain library representation by always passaging >12 million cells.

RNA-based INSERT-seq screens. For steady-state RNA read-out, >5 million library-containing cells were pelleted and resuspended in TRIzol (Ambion). RNA was cleaned up using Direct-zol RNA miniprep columns (Zymo Research), treated with RQ1 DNase (Promega), then cleaned up again using a total RNA purification kit (Norgen Biotek Corp).

For nascent RNA, library-containing cells were permeabilized and run on largely as described[49], with the following changes: for permeabilization, buffers W and P were supplemented with 1 mM EGTA, and buffer F was supplemented with 1 mM EDTA. Cells were resuspended in 500 µL Buffer W and 10 mL Buffer P was added, and this was incubated on a Nutator for 1 minute, then put on ice for 6 minutes. Cells were washed once with 10 mL buffer W, before being resuspended in Buffer F. Run-on was performed on 30 million permeabilized cells, and in the 2× run-on reaction mix, $MgCl_2$ was present at 10 mM, SUPERase–In at 0.4 U/µL, and (biotin-)rNTPs at 50 µM each. Run-on was performed at 37 °C for 15 minutes, and biotin-11-C/UTP (Perkin-Elmer) was mixed with unlabeled rGTP and rATP (New England BioLabs). Immediately after RNA isolation using the Total RNA Purification Kit (Norgen Biotek Corp), biotinylated RNAs were bound to Streptavidin M-280 magnetic beads (Thermo Fisher) and binding and washing was done as described.

Chromatin-associated RNA (Chr-RNA) was isolated as published previously[50], but 0.1% Triton X-100 was added after pellets were resuspended in buffer B. Incubations were all performed on ice, 8 minutes in buffer A and two incubations in buffer B for 15 minutes. The washed chromatin pellets were resuspended in Trizol, RNA was extracted using Direct-zol RNA miniprep columns (Zymo Research), treated with RQ1 DNase (Promega), then cleaned up again using a total RNA purification kit (Norgen Biotek Corp).

Purified steady-state and chromatin-associated RNA and nascent RNA bound to beads was reverse transcribed with Superscript IV (Thermo Fisher) using a gene-specific primer. For experiments with the library integrated at the Oct4 uaRNA reporter locus, the RT primer annealed in mKate2, 221 nt downstream of the variable region. For experiments with the library integrated at the 4930461G14Rik lincRNA locus, the RT primer annealed 287 nt downstream of the variable region. cDNA from steady-state RNA was RNase-treated, and cDNA from nascent RNA was removed from the beads with heat. cDNA was concentrated using 2× RNA XP beads (Beckman Coulter) to limit the number of parallel PCR reactions that had to be set up.

Sequencing libraries were generated through 17–20 PCR cycles using NEBNext Ultra II Q5 (New England Biolabs) and NEBNext Multiplex oligonucleotides, uniquely indexing each sample. Amplicons were cleaned up using 1.5× AMPure XP beads (Beckman Coulter) and concentrations were quantified using the NEBNext Library Quant Kit for Illumina, before mixing and checking the size distribution by TapeStation (Agilent). Paired-end sequencing was performed on MiSeq, NextSeq500, HiSeq4000, or Novaseq6000 (Illumina, 110 + 50 cycles or 150 + 150 cycles).

For normalization purposes, genomic DNA was extracted using QuickExtract (Lucigen) in parallel with RNA extraction, and sequencing libraries were generated from genomic DNA as described below for the Sort-seq protocol. In these experiments, either the *Oct4* uaRNA or the 4930461G14Rik lincRNA locus was amplified in the first round of PCR, depending on where the library was integrated.

Steady-state RNA screens after EXOSC3 depletion or non-targeting control siRNA treatment were performed similarly to other steady-state RNA screens, except that 48 hours prior to harvesting, cells were transfected with 40 nM siRNAs, using the RNAiMAX reagent (Invitrogen) in a suspension transfection.

For EXOSC3 depletion, a mix of four siGENOME siRNAs was used (catalog ID MQ-064537-01-0002, Horizon Discovery). As a control, the siGENOME Non-Targeting Control siRNA no. 2 was used.

Using these methods, four replicates of steady-state RNA and two replicates of nascent RNA and chromatin-associated RNA at the Oct4 uaRNA reporter locus were performed, as well as three replicates each of steady-state RNA after siRNA treatment (siEXOSC3 and control) and with the library integrated at the 4930461G14Rik lincRNA locus. We note that the steady-state RNA results inherently have a larger dynamic range than the results with nascent and chromatin-associated RNA: using RT–qPCR and comparing to a standard of in vitro transcribed RNA, we have estimated that there are ~15 copies of the reporter transcript in steady-state RNA per cell. In comparison, on the ~650 bp between the annealing site of the RT primer and the PAS downstream of the reporter, we would not expect more than 1 RNAPII.

Sort-seq screen. Single-cell suspensions of library-containing cells were made in PBS with 1% FBS and 1 mM EDTA. PI (0.1 μg/mL final concentration) was added and PI-negative cells were sorted into six bins on the basis of the ratio between red (594 nm, 610/20) over blue (405 nm, 450/40) fluorescence using a FACSAria cell sorter (BD Biosciences). Sorting on the ratio between integrated fluorescent proteins reduces noise introduced by cell-to-cell variation in cell size and cell cycle stage. Gates were set so that the middle four bins were of equal width in the blue × red scatter plot, while the outer two bins were wider to include even cells with larger changes in protein expression. Compared with control cells in which the pHV137 Cas9+gRNA plasmid had been co-transfected with a repair template that made only a point mutation to prevent re-cleavage, the population of library-containing cells had the same median red/blue ratio, but a wider distribution.

Cells collected in each fluorescence bin, as well as unsorted cells, were pelleted and genomic DNA extracted using QuickExtract (Lucigen). Genomic DNA was also extracted from a pellet of unsorted cells. In a first round of PCR, the *Oct4* uaRNA locus was amplified from these samples with NEBNext Ultra II Q5, using primers that would not amplify remaining plasmid or non-specific integrations at other loci. Amplicons were cleaned up using 0.9× AMPure XP beads and a small fraction used for a genotyping PCR. This confirmed an enrichment of insert-containing alleles in the outer bins, whereas alleles that did not integrate an insert were mostly found in the middle two bins.

The amplicons from the first round of PCR were used as a template in a PCR with the NEBNext multiplex oligos to generate the sequencing library. This PCR and all subsequent steps were done as described for the RNA screen, with two changes: samples were amplified for ~22 cycles total between the first and second PCR step, and after pooling the sample was gel-extracted from a 1% agarose gel to size-select. Two replicate Sort-seq experiments were performed and had strong agreement (Spearman's rho of 0.91).

Counting reads per insert. To map inserts amplified from genomic DNA, first Cutadapt v1.14 (ref. [51]) was used to trim off adapters and low-quality ends, then bowtie2 (version 2.3.4.3)[52] was used to map against an index made of the library of 16,462 insert sequences. For bowtie2 mapping, the following parameters were set:–no-discordant–no-mixed–rdg 20,5–rfg 20,5–score-min L,0,−0.11. This basically disallowed gaps and allowed up to three mismatches on high-quality bases, with the editing distance between any 2 sequences in the library being >5.

For the cDNA-derived samples, splice junctions were called with STAR (version 2.7.0)[53], using as input a concatenated fastq file with the trimmed reads of replicates 2–4 (which had much higher coverage than replicate 1) of steady-state RNA-derived libraries from the untreated *Oct4* uaRNA reporter locus screen. The following flags were used to run STAR:–outSJfilterOverhangMin 10 4 4 4–peOverlapNbasesMin 10–peOverlapMMp 0.08–alignEndsType EndToEnd–outFilterMismatchNmax 5–scoreDelOpen −10–scoreInsOpen −10. The detected splice junctions were filtered to include only junctions on the + strand with >10 read counts and >25% of the maximum counts for other junctions detected in the same insert. A new reference genome was generated that contained both unspliced and spliced versions of the full library of sequences. Bowtie2 was run with the same settings as above to map the cDNA libraries to a bowtie2 index of this new reference.

Samtools v1.9[54] was used to output the number of reads mapped to each sequence in the 'reference genome'.

Analysis of INSERT-seq RNA data. Data of steady-state RNA, nascent RNA and corresponding genomic DNA (gDNA) were first normalized by setting the total of mapped reads to 1. Abundances of all unspliced and spliced versions of an insert were added up for a total abundance per insert per sample. A cut-off was applied to filter out inserts with very low abundance in gDNA, for which no reliable RNA/gDNA ratio could be calculated. RNA data were first normalized to abundance in the gDNA data, and then normalized to controls, so that the median RNA/gDNA ratio of ~1,100 synthetic (randomly generated) control sequences was 1. For each insert, a splicing efficiency was calculated by calculating the spliced/total ratio, where 'spliced' is the summed abundance of all spliced versions (0 if no spliced versions were detected). After averaging replicates and before plotting data on a $\log_2$ scale, inserts with an RNA/gDNA ratio of 0 were set to the minimum non-zero ratio for that dataset.

Sort-seq analysis. A cut-off was applied to filter out very lowly abundant inserts for which no reliable distribution could be calculated. To calculate a distribution over the bins for each insert, it has to be considered that the percentage of insert-containing cells can vary between bins and that sequencing yields no data on this percentage, necessitating a different method to calculate the scaling factor for each bin. After setting the total of mapped reads to one in each sequencing sample, a linear model was fit to predict the composition of the unsorted sample using the composition of each of the six bins. For the first replicate, two unsorted samples were sequenced, the replicate starting from more input material was weighed 75%, the lower input replicate was weighed 25%. The coefficients of this model (which fit the data well, with a Pearson correlation of >0.9 between predicted and observed) were used to scale bins and calculate the percentage of each insert found in each of the bins. From this, the Sort-seq score was calculated by assigning each bin a score of one through six and calculating a weighted average per insert.

Analysis of sequences embedded in constant background regions. Short sequences were all evaluated in the context of the same five background sequences. For each background, the median signal of 30 scrambled 10-nt 5′SSs was used to normalize all sequences in that background. For RNA assays, normalization involved dividing by the median of scrambled 10-nt 5′SSs, for Sort-seq this median was subtracted. The mean effect of an embedded sequence across all backgrounds was then calculated.

Analysis of barcoded wild-type and mutated introns. Original introns were included in the library with three barcodes, and 1-bp and 3-bp mutations of the 5′SS and 3′SS with two barcodes each. Barcodes were randomly matched up with mutant versions, allowing us to determine that none of the barcodes affected the results. Since the 1-bp and 3-bp mutations per splice site behaved very similarly, they were considered together as simply 5′SS mutant or 3′SS mutant. The median of all barcoded replicates across all replicate experiments was calculated as the average value for each intron version (original, 5′SS mutant, 3′SS mutant). To assess the effects of splice site mutations, introns were considered only if mutations reduced splicing efficiency to <3% in nascent RNA from >30% in original introns.

Motif searching algorithms. To find sequence motifs that may convey positive signals, the MEME[38] and Homer[39] algorithms were used to identify motifs enriched in the top 10% versus bottom 50% of TSS-proximal mRNA regions. Searches were limited to the given strand, corresponding to the sequence of the RNA. For MEME, the ZOOPS Motif Site Distribution was expected, and motifs were 6–12 nt long. The Homer search was limited to motifs of 6, 8, or 10 nt.

Finding 5′SS motifs using MOODS and MaxEnt. MOODS v1.9.3 (ref. [55]) was used to find all 5′SS motif matches (motif SD0001.1 from the JASPAR database) in the insert sequences using a lenient cut-off ($P < 0.05$). Only motif matches on the given strand (RNA sequence) were considered, and the MaxEnt score[41] for the sequence of each motif match was obtained. Each insert was assigned the maximum MaxEnt score of any of its 5′SS motif matches. Sequences with a MaxEnt score of >5 were considered a medium-strength 5′SS, and sequences with a MaxEnt score of >10 were considered strong 5′SSs.

PRO-seq. Cell permeabilization and PRO-seq library construction was performed as described before[49], with the following modifications: 1 million permeabilized mESCs were spiked with 5% permeabilized *Drosophila* S2 cells. PRO-seq libraries were generated from four biological replicates by the Nascent Transcriptomics Core at Harvard Medical School. Then, 2× nuclear run-on buffer comprised of 10 mM Tris (pH 8), 10 mM $MgCl_2$, 1 mM DTT, 300 mM KCl, 20 μM/ea biotin-11-NTPs (Perkin-Elmer), 0.8U/μL SUPERase–In (Thermo), and 1% sarkosyl. The run-on reaction was performed at 37 °C. The NEB 5′DNA adenylation kit was used to adenylate the 3′ adapter. Adenylated 3′ adapter was ligated overnight at 16 °C by T4 RNA ligase 2, truncated KQ (NEB), per the manufacturer's instructions with 15% PEG-8000 final. Following the 3′ adapter ligation and before the second bead binding, samples were incubated in 180 μL betaine blocking buffer (1.25 M betaine in binding buffer with 0.6 μM blocking oligonucleotide (TCCGACGATCCCAC GTTCCCGTGG/3InvdT/)) for 5 min at 65 °C and then 2 min on ice. Following treatment with T4 polynucleotide kinase (NEB), beads were washed once each in high salt, low salt, and blocking oligonucleotide wash (0.25× T4 RNA ligase buffer (NEB), 0.3 μM blocking oligonucleotide) buffers. Beads were then resuspended in 5′ adapter mix (10 pmol 5′ adapter, 30 pmol blocking oligo, water). For the 5′ adapter ligation, 15% PEG-8000 was used. Eluted cDNA products were amplified for 9–11 cycles of PCR with NEBNext Ultra II Q5 master mix (NEB) and Illumina TruSeq PCR primers RP-1 and RPI-X, per the manufacturer's instructions. Final libraries were pooled and sequenced on the Illumina NovaSeq.

PRO-seq data preprocessing was performed as described elsewhere[49]. PRO-seq around uaRNA and eRNA TSSs included in the INSERT-seq library were investigated by generating count matrices from the PRO-seq bedgraph files using the make_heatmap script (https://github.com/AdelmanLab/NIH_scripts/tree/

main/make_heatmap). Loci with <15 promoter read counts (PRO-seq reads of which the 5′ ends map in the TSS ± 10-nt window) were filtered out, and groups were divided up on the basis of GC content. Metagene profiles show the mean counts per 25-nt bin for each group. One outlier in group 2 of the eRNAs was removed. The elongation index was calculated as the ratio of PRO-seq density in the window from +50 to +250 downstream of the TSS (representing elongating pol II), divided by the density from the TSS to +50 (representing promoter-proximal pol II).

RT–qPCR. Steady-state RNA was isolated using either TRIzol (Ambion) and Direct-zol columns (Zymo Research) or RNeasy columns (Qiagen). Chromatin-associated RNA was isolated as described above for INSERT-seq screens. All RNA was DNase-treated and then purified using a total RNA purification kit (Norgen Biotek). cDNA was generated using Superscript IV (Thermo Fisher) with random hexamer oligonucleotides. qPCRs were performed using a home-made SYBR mastermix on a BioRad CFX384 qPCR instrument.

Immunoblotting. Single-cell suspensions were pelleted, and pellets were resuspended in 1× Laemmli sample buffer (BioRad) with 1:40 beta-mercaptoethanol. Samples were boiled for 10 minutes and spun down at 13,000*g* for 5 minutes. The lysates of 100,000 cells were run on 4–20% Mini-Protean TGX Precast Protein Gels (BioRad), according to the manufacturer's instructions. Proteins were transferred to nitrocellulose membranes for 70 minutes at 300 mA. Membranes were blocked using 5% dried milk in PBS, and incubated with anti-EXOSC3 (1:2,000, A303-909A, Bethyl Labs) and anti-actin (1:2,000, sc-1616, SantaCruz) in 5% dried milk in TBS-T, followed by HRP-conjugated secondary antibodies (1:10,000, 111-035-144 and 705-035-147, Jackson ImmunoResearch). After incubation with SuperSignal West Pico PLUS Chemiluminescent Substrate (BioRad) according to manufacturer's instructions, blots were imaged using the BioRad ChemiDoc, using the ImageLab software.

Flow analysis on clonal lines. Single-cell suspensions of library-containing cells were made in PBS with 1% FBS and 1 mM EDTA. Red fluorescence of *Atp5a1* intron-containing lines was measured on a FACSAria cell sorter (BD Biosciences), with an excitation laser of 594 nm and 610/20 nm detector. Red fluorescence of *Smc1* intron-containing lines was measured on a HS800S cell sorter (Sony), with an excitation laser of 561 nm and 600/60 nm detector.

ChIP–seq genome profile snapshots. Genomic profile snapshots were generated using the fluff software package[56].

Published H3K4me1 data[57] were downloaded and converted to FASTQ from the SRA using the SRA toolkit (accessions SRR1202461 and SRR1202462). H3K4me3 and H3K27ac ChIP–seq data were generated as follows. mESCs were cross-linked for 5 (H3K4me3) or 10 (H3K27ac) minutes with 1% formaldehyde and were quenched with 0.25 µM glycine. Chromatin fragmentation was performed in sonication buffer (20 mM Tris pH 8.0, 2 mM EDTA, 0.5 mM EGTA, 0.5% SDS, 0.5 mM PMSF) with a Qsonica Q800R bath sonicator at 70% amplitude pulsing with 15 second 'ON' and 45 seconds 'OFF' cycles, for 30 (H3K4me3) or 20 (H3K27ac) minutes total sonication time. Fragment sizes were checked by agarose gel. Fragmented chromatin from 7.5 M (H3K4me3) or 2.5 M (H3K27ac) mESCs was used as input for immunoprecipitation. For the H3K27ac ChIP–seq, mESCs were spiked with sonicated chromatin from 0.5 M S2 cells. ChIP material was diluted with 1 mL of IP buffer (20 mM Tris pH 8.0, 2 mM EDTA, 0.5% Triton X-100, 150 mM NaCl, 10% glycerol, 5% BSA) and pre-cleared with 10 µL DynaBeads Protein A (Thermo Fisher) for 1 hour at 4 °C. The supernatant was then further diluted with 250 µL IP buffer and incubated with 16 µL H3K4me3 antibody (EpiCypher cat. no. 13-0028, lot no. 18303001) or 10 µL H3K27ac antibody (Active Motif cat. no. 39133, lot no. 22618011) overnight at 4 °C. Bound chromatin was isolated with 150 µL (H3K4me3) or 50 µL (H3K27ac) of pre-blocked DynaBeads Protein A for 2 hours at 4 °C. Beads were washed 1x in low salt (20 mM Tris pH 8.0, 2 mM EDTA, 1% Triton X-100, 150 mM NaCl, 0.1% SDS), 3× in high salt (20 mM Tris pH 8.0, 2 mM EDTA, 1% Triton X-100, 500 mM NaCl, 0.1% SDS), 1× in LiCl (20 mM Tris pH 8.0, 2 mM EDTA, 250 mM LiCl, 1% IGEPAL, 1% (wt/vol) sodium deoxycholate), and 2× in TE (10 mM Tris pH 8.0, 0.1 mM EDTA) prior to elution with 2× 250 µL of elution buffer (10% SDS, 100 mM sodium bicarbonate). Then, 0.2 M NaCl was added and crosslinks were reversed by incubation at 65 °C overnight. The IP mixture was treated with Proteinase K (Invitrogen) for 1 hour at 50 °C. DNA was extracted with phenol–chloroform–isoamyl alcohol and precipitated with ethanol. Sequencing libraries were prepared using the NEBNext Ultra II DNA Kit (New England Biolabs) as per the manufacturer's protocol. H3K4me3 and H3K27ac ChIP–seq libraries were sequenced on an Illumina NextSeq using a paired-end 160 (read1: 80, index: 6, read2: 80) and 80 cycle (read1: 43, index: 6, read2: 42) kits respectively.

Data were mapped to mm10 (GRCm38) using Bowtie v1.2.2 (ref. [58]), using the parameters -v2 -X1000 –best and -k1 (for H3K4me1/me3) or -m1 (for H3K27ac). Using the extract_fragments script (https://github.com/AdelmanLab/NIH_scripts/tree/main/extract_fragments), reads were filtered for insert fragment sizes of 75–500 bp, duplicate reads were removed using Samtools v1.3, and bedgraphs mapping the midpoints of deduplicated fragments were generated. For H3K4me1 data, which were sequenced as single-read, fragment lengths of 150 bp were assumed. Outputted bedgraphs were at 1-bp resolution for H3K4me1/me3 data and in 25-bp bins for H3K27ac data. Bedgraphs from replicate experiments were merged using the bedgraph2stdbedgraph script (https://github.com/AdelmanLab/NIH_scripts/tree/main/bedgraphs2stdBedGraph), and 1-bp merged bedgraphs were rebinned to 25-bp bins using binBedGraph (https://github.com/benjaminmartin02/binBedGraph).

TT-seq. TT-seq was performed essentially as described[59], with some minor modifications. Briefly, mESCs, treated with 0.03 % DMSO for 4 hours, were labeled with 500 µM 4-thiouridine (Sigma) for 20 minutes. Cells were counted and collected in TRIzol, to which was added 5% S2 cells, labeled with 4sU for 2 hours. RNA was chloroform-extracted, DNase-treated, and chloroform-extracted. To 60 µg total RNA was added 2× fragmentation buffer (final concentration: 75 mM Tris Cl, pH 8.3, 112.5 mM KCl, and 4.5 mM $MgCl_2$) and heated to 95 °C for 5 minutes. Fragmentation was stopped by adding EDTA to 50 mM and placing samples on ice. RNA was ethanol-precipitated and resuspended in H_2O. Fragment sizes were checked on the TapeStation (peak size of ~800 nt). The biotinylation reaction was performed with 0.025 mg/mL MTSEA-Biotin XX (Biotum) in reaction buffer (20% *N*,*N*-dimethylformamide, 1 mM EDTA, 10 mM HEPES pH 8) for 45 minutes in the dark. Labeled RNA was chloroform-extracted and ethanol-precipitated, and resuspended in 90 µL H_2O. Seventy-five microliters of DynabeadsTM M-280 streptavidin (Thermo Fisher) were prewashed with decon solution (0.1 M NaOH + 50 mM NaCl), 2× 100 mM NaCl, 2× high salt buffer (100 mM Tris Cl pH 7.4, 10 mM EDTA pH 8, 1 M NaCl, 0.05% (vol/vol) Tween 20), and resuspended in high salt buffer. Labeled RNA was denatured by heating to 65 °C for 5 minutes and placing on ice for 2 minutes. Ten microliters of high salt buffer was added to labeled RNA. Prewashed beads were placed on magnet, supernatant discarded, and then beads were resuspended in the labeled RNA. Beads and RNA were rotated for 30 minutes in the dark. Beads were then washed 4× for 1 minute with high salt buffer. Labeled RNA was eluted 2× with 100 mM DTT (1,4-dithiothreitol). RNA was cleaned up with microelute columns (Norgen). Sequencing libraries were prepared from 250 ng labeled RNA using the TruSeq Stranded Total RNA sequencing kit with RiboZero rRNA depletion, shortening the fragmentation step to 3 minutes and using 8 cycles of PCR amplification.

Reads were trimmed for a minimum quality score of 20 using a custom script (https://github.com/AdelmanLab/NIH_scripts/tree/main/trim_and_filter_PE), adapter sequences were removed using cutadapt version 1.14 (ref. [51]). Reads were first mapped to dm6 using bowtie[58] before aligning to mm10 using STAR (version 2.7.3a)[53]. Strand-specific bedgraphs were generated from deduplicated BAM files using STAR.

Plotting and statistical tests. All violin plots were made in Graphpad Prism v9.1 with smoothing set to medium. Lines of density plots were calculated in R and then plotted in Graphpad Prism.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw and processed data files of all INSERT-seq experiments, PRO-seq, H3K4me3 ChIP–seq, and TT-seq are available at the Gene Expression Omnibus, accession no. GSE178230. H3K27ac ChIP–seq data are available through the 4DN data portal (https://data.4dnucleome.org/), ExperimentSet accession no. 4DNESQ33L4G7. H3K4me1 mESC ChIP–seq data were downloaded from the Gene Expression Omnibus, accession no. GSE56138. Reference genome mm10 (GRCm38) can be downloaded using RefSeq assembly accession number GCF_000001635.20. Supplementary Tables 3–7 provide all normalized and averaged data from INSERT-seq experiments, as well as which inserts are included in which plot. Uncropped image files and processed data shown in each plot are provided as source data. Source data are provided with this paper.

Code availability

All scripts used for analysis of INSERT-seq data can be found on Github: https://github.com/AdelmanLab/Vlaming2021_INSERT-seq_paper. URLs for all custom scripts used for PRO-seq, TT-seq and ChIP–seq analysis are provided in the Methods; these can be found at https://github.com/AdelmanLab/NIH_scripts/ and https://github.com/benjaminmartin02/binBedGraph.

References

46. Rivera-Mulia, J. C. et al. Allele-specific control of replication timing and genome organization during development. *Genome Res.* **28**, 800–811 (2018).
47. Williams, L. H. et al. Pausing of RNA polymerase II regulates mammalian developmental potential through control of signaling networks. *Mol. Cell* **58**, 311–322 (2015).

48. Brinkman, E. K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* **42**, e168–e168 (2014).
49. Reimer, K. A., Mimoso, C. A., Adelman, K. & Neugebauer, K. M. Co-transcriptional splicing regulates 3′ end cleavage during mammalian erythropoiesis. *Mol. Cell* **81**, 998–1012.e7 (2021).
50. Henriques, T. et al. Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. *Mol. Cell* **52**, 517–528 (2013).
51. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10 (2011).
52. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
53. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
54. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
55. Korhonen, J. H., Palin, K., Taipale, J. & Ukkonen, E. Fast motif matching revisited: high-order PWMs, SNPs and indels. *Bioinformatics* **33**, 514–521 (2016).
56. Georgiou, G. & van Heeringen, S. J. fluff: exploratory analysis and visualization of high-throughput sequencing data. *PeerJ* **4**, e2209 (2016).
57. Buecker, C. et al. Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell Stem Cell* **14**, 838–853 (2014).
58. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
59. Schwalb, B. et al. TT-seq maps the human transient transcriptome. *Science* **352**, 1225–1228 (2016).

Acknowledgements

We thank K. Sasaki for her help in optimizing the run-on protocol for screening purposes and E. Kaye for discussions on library design. We thank S. Buratowski for useful discussions on the project, and D. Shlyueva and T. H. Jensen for feedback on the manuscript. We are also grateful to the Flow Cytometry Facility at the HMS Department of Immunology for cell sorting help and advice, the HMS Nascent Transcriptomics Core for PRO-seq library construction, and the HMS Biopolymers Facility and The Bauer Core Facility at Harvard University for next-generation sequencing. This research was supported by the European Molecular Biology Organization (ALTF 531-2017 to H. V.), Human Frontier Science Program (LT000651/2018-L to H. V.), the National Institutes of Health (NIH R01 GM139960 to K. A.), startup funding from Harvard Medical School to K. A, the National Science Foundation Graduate Research Fellowship (DGE1745303 to C. A. M.) and the Canadian Institutes of Health Research (Banting fellowship to B. J. E. M.).

Author contributions

H. V. and K. A. conceived the study and designed experiments. H. V. performed experiments and analyzed data. C. A. M. performed PRO-seq data analysis, helped generate intron-containing clonal cell lines, and optimized the run-on assay and knockdown conditions. B. J. E. M. and A. R. F. performed ChIP–seq and TT-seq experiments. K. A. supervised the study. H. V. and K. A. wrote the manuscript with input from all co-authors.

Competing interests

K. A. is a consultant for Syros Pharmaceuticals, is on the scientific advisory board of CAMP4 Therapeutics, and receives research funding from Novartis unrelated to this work. The remaining authors declare no competing interests.

Additional information

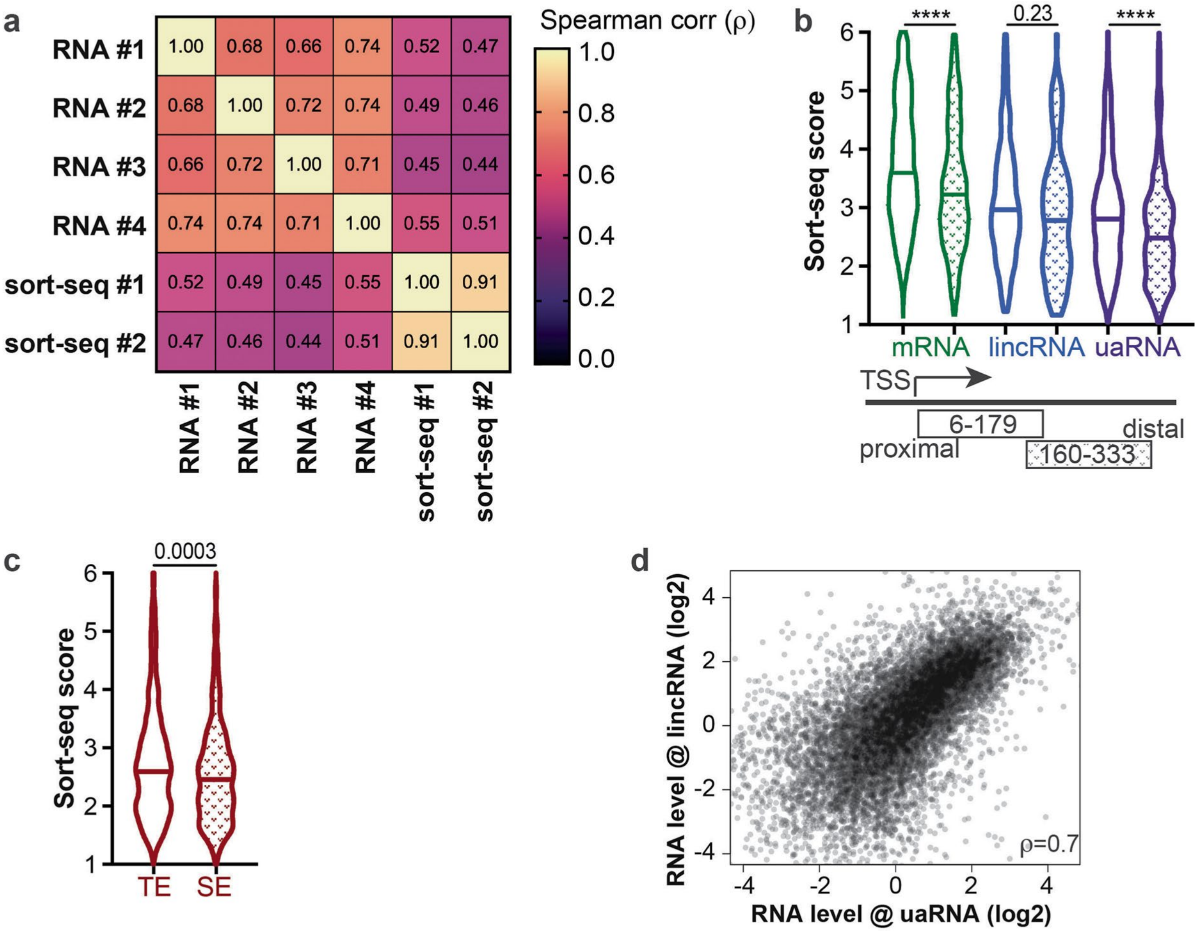
Extended data is available for this paper at https://doi.org/10.1038/s41594-022-00785-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41594-022-00785-9.

Correspondence and requests for materials should be addressed to Hanneke Vlaming or Karen Adelman.

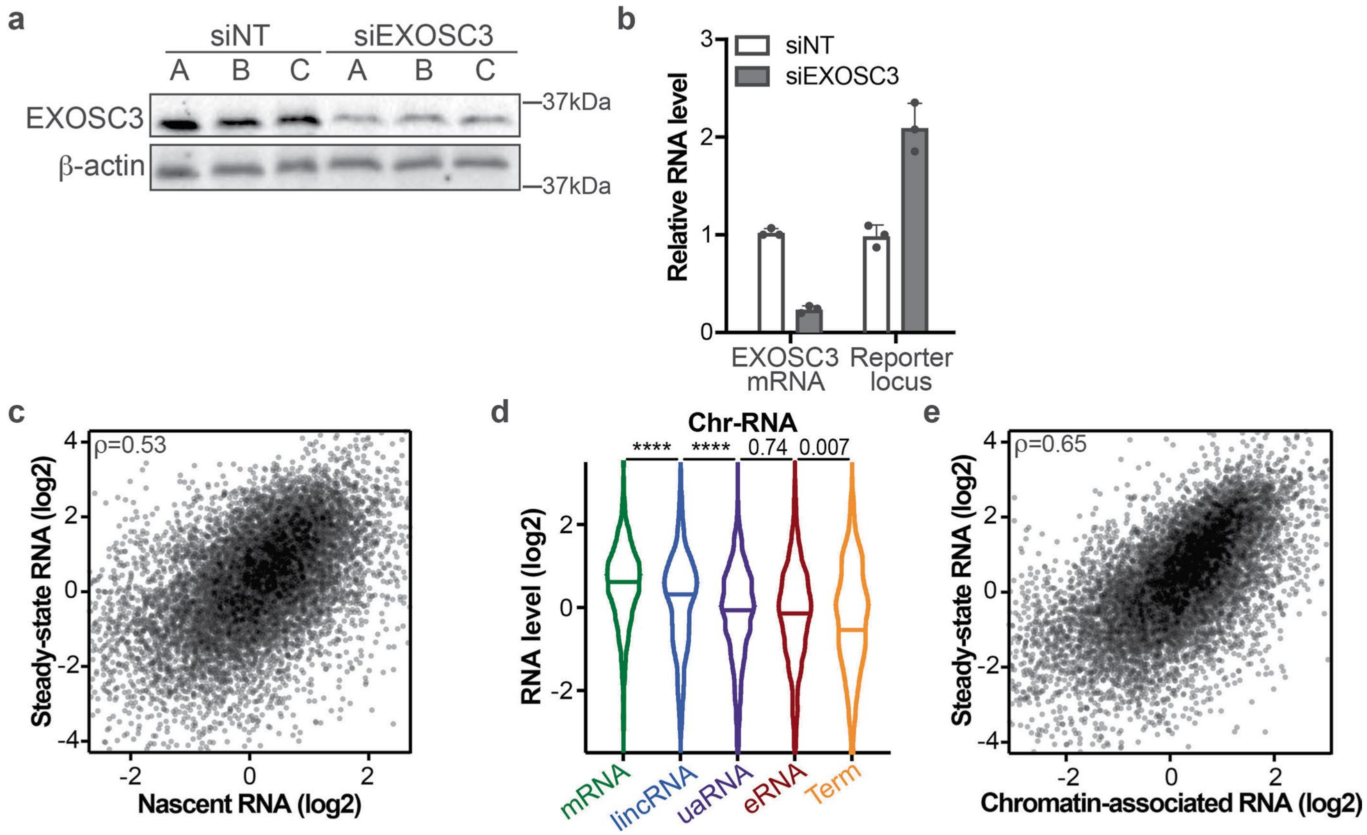
Peer review information *Nature Structural and Molecular Biology* thanks Yongsheng Shi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Carolina Perdigoto, in collaboration with the Nature Structural & Molecular Biology team. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

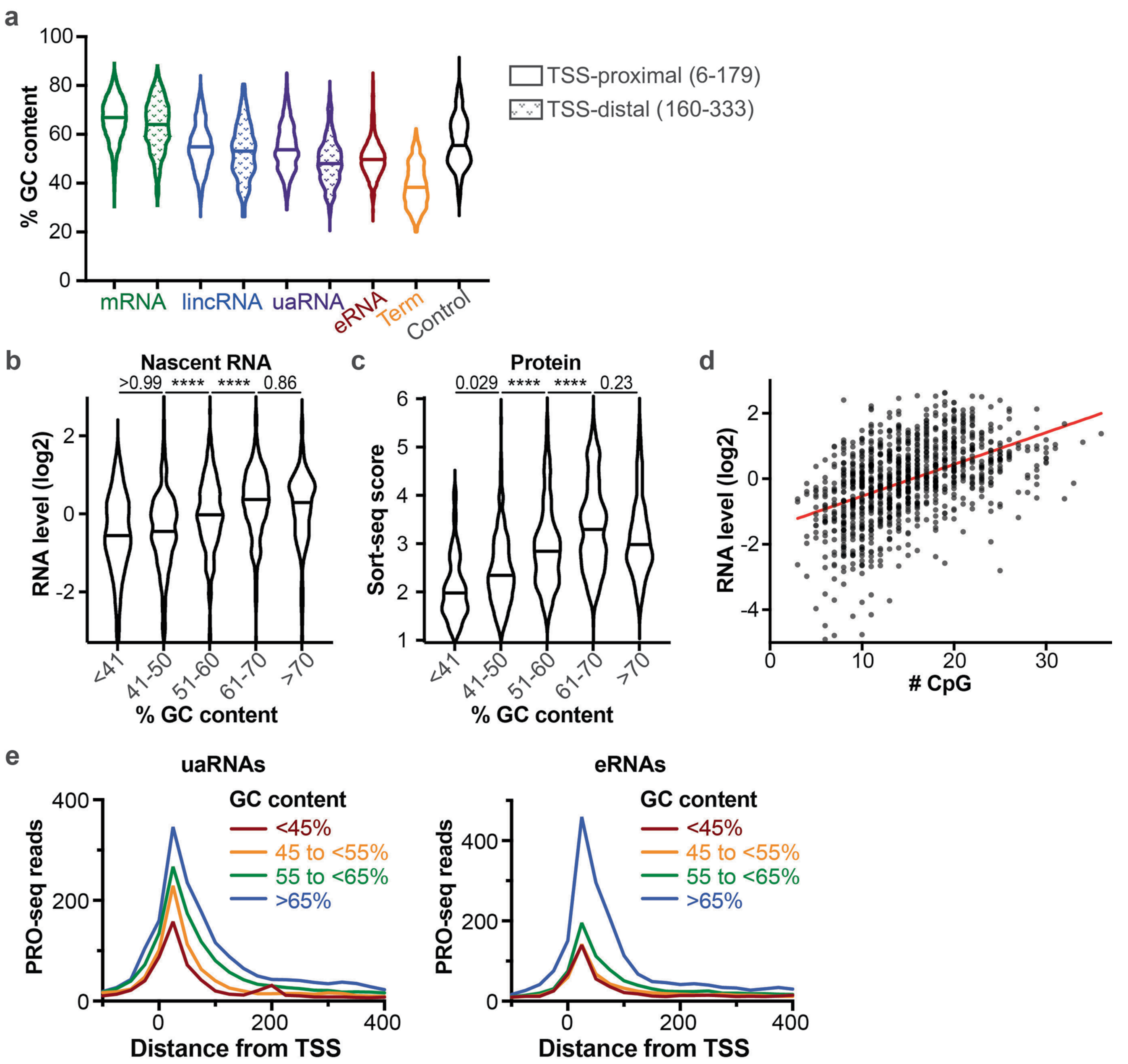


	RNA #1	RNA #2	RNA #3	RNA #4	sort-seq #1	sort-seq #2
RNA #1	1.00	0.68	0.66	0.74	0.52	0.47
RNA #2	0.68	1.00	0.72	0.74	0.49	0.46
RNA #3	0.66	0.72	1.00	0.71	0.45	0.44
RNA #4	0.74	0.74	0.71	1.00	0.55	0.51
sort-seq #1	0.52	0.49	0.45	0.55	1.00	0.91
sort-seq #2	0.47	0.46	0.44	0.51	0.91	1.00

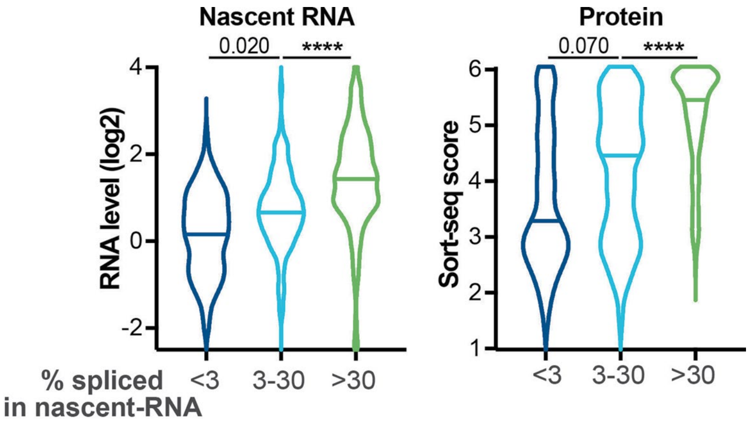
Extended Data Fig. 1 | Correlations between INSERT-seq experiments. a, Spearman correlation coefficients between steady-state RNA and Sort-seq experiments, using all inserts for which data was obtained in each of the six experiments (n = 12,090). **b**, Sort-seq scores of inserts containing TSS-proximal and TSS-distal genomic regions of indicated RNA classes. Same groups as in Fig. 1e. Comparisons between proximal and distal regions by Kruskal-Wallis test, **** indicates $P < 0.0001$. **c**, Sort-seq scores of inserts containing TSS-proximal regions from typical enhancers (TE, $n = 1,506$) and super enhancers (SE[22], $n = 600$), compared by Mann–Whitney test. **d**, Correlation between steady-state RNA levels at the *Oct4* uaRNA locus (average of 4 replicates) and *4930461G14Rik* lincRNA locus (average of 3 replicates). Plotted are all inserts used for Fig. 1, as well as synthetic controls sequences (Fig. 3), for which data was obtained at the lincRNA locus ($n = 11,600$).



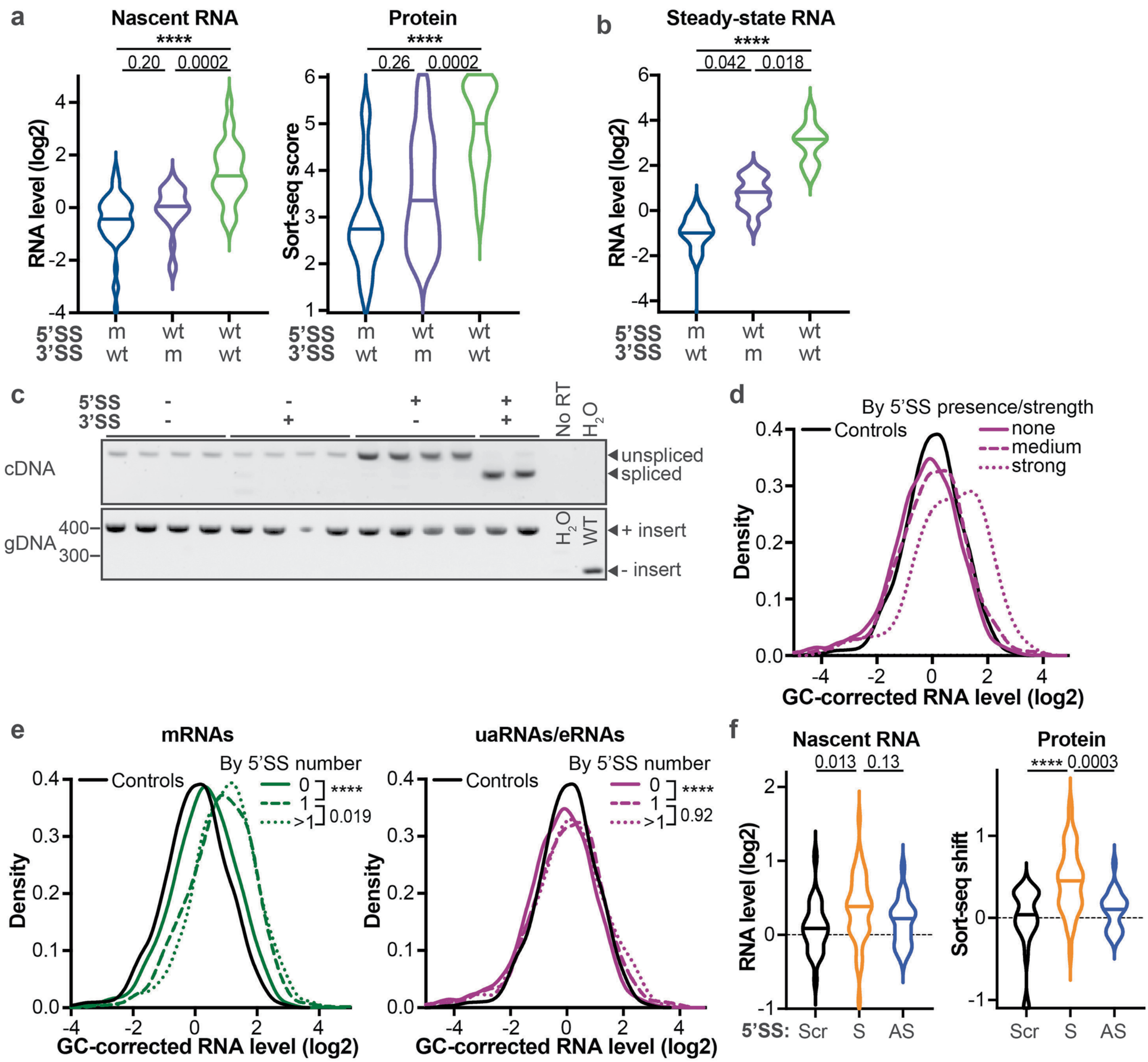
Extended Data Fig. 2 | EXOSC3 knockdown validation and correlation between nascent RNA and steady-state RNA results. a, Immunoblot showing EXOSC3 protein level in control and siEXOSC3 conditions, harvested from the same experiment as the screen in Fig. 2a, b. **b**, RT-qPCR on steady-state RNA samples with which the screen was performed, showing levels of the *EXOSC3* mRNA and the reporter transcript, just downstream of the library integration site, both internally normalized to TBP. Bars show mean, whiskers indicate standard deviation, $n=3$ biologically independent experiments. **c**, Correlation between nascent RNA (average of 2 replicates) and steady-state RNA (average of 4 replicates) levels, showing all inserts used for Fig. 1, as well as synthetic controls sequences (Fig. 3), $n=11{,}132$. **d**, Chromatin-associated RNA (Chr-RNA) results with library at uaRNA locus. mRNAs $n=3{,}832$, lincRNAs $n=339$, uaRNAs $n=1{,}730$, eRNAs $n=2074$, mRNA terminators $n=414$. Neighbors were compared by Kruskal-Wallis test, **** indicates $P<0.0001$, higher P values are indicated in the panel. **e**, Correlation between Chr-RNA (average of 2 replicates) and steady-state RNA (average of 4 replicates) levels, all inserts from panel c for which Chr-RNA data was obtained ($n=11{,}029$).



Extended Data Fig. 3 | GC content in genomic regions and its effect on expression. a, Distribution of GC contents in inserts of the indicated classes included in the library. Open violins show TSS-proximal regions, patterned violins show TSS-distal regions. **b**,**c**, Nascent RNA abundance (b) and sort-seq scores (c) of control sequences grouped by GC content percentage. *N* = 39/281/330/292/117 for <41/41-50/51-60/61-70/>70%, respectively. Neighbors were compared by Kruskal-Wallis test, **** indicates $P < 0.0001$, higher P values are indicated in the panel. **d**, Relation between the number of CpG dinucleotides in synthetic control sequences and their steady-state RNA levels (*n* = 1,059). The red line is the best linear fit through the data. Pearson r = 0.47, $P < 0.0001$. **e**, Metagene representations of PRO-seq signal around TSSs of uaRNAs (left) or eRNAs (right), grouped by GC content of the transcribed sequence from +6 to +179 downstream of the TSS (the region included in our screening library). Data shown are from endogenous genomic locations of sequences included in the INSERT-seq screen. Read counts were summed into 25nt bins.



Extended Data Fig. 4 | Co-transcriptionally spliced introns boost transcription and protein expression. Nascent RNA levels (left) and Sort-seq scores (right) of inserts containing wild-type introns (unbarcoded) grouped by splicing efficiency measured using the nascent RNA screen data. <3% spliced *n* = 76, 3-30% spliced *n* = 107, >30% spliced *n* = 198, significance tested by Kruskal-Wallis test. **** indicates *P* < 0.0001, higher P values are indicated in the figure.



Extended Data Fig. 5 | Effects of splice site mutants and 5′SS insertion in INSERT-seq and clonal lines. a, Nascent RNA levels (left) and Sort-seq scores (right) of intron-containing inserts with wild-type (wt) or mutant (m) splice sites. As in Fig. 5a, only introns are shown of which the wild-type version was >30% spliced in nascent RNA and mutants were <3% spliced. 5′SS mutants $n=51$, 3′SS mutants $n=23$, WT $n=52$, comparisons by Kruskal-Wallis test. The differences between 5′SS and 3′SS mutants was not significant in these analyses, but the pattern of the 3′SS mutants being more abundant on average was consistent with the steady-state RNA result (Fig. 5a). **b**, Steady-state RNA levels of intron-containing inserts with wild-type (+) and mutant (-) splice sites as in Fig. 5a, but showing only inserts that do not contain a PAS hexamer (any of the top-10 PASs in mouse[3]). 5′SS mutants $n=19$, 3′SS mutants $n=10$, WT $n=20$, comparisons by Kruskal-Wallis test. **c**, Characterization of all clonal cell lines shown in Fig. 5b, where versions of the 14th intron of the Smc1 gene with wild-type (+) or mutant (−) splice sites were integrated at the *Oct4* uaRNA reporter locus. Top shows RT-PCR, bottom shows PCR on genomic DNA. All clonal lines show genomic integration of the same size in the genomic DNA, but only lines where the intron is flanked by two wild-type splice sites show evidence of splicing. Note that lanes should not be quantitatively compared to each other, as amounts of template material were not controlled. **d**, Density plot of GC-corrected steady-state RNA levels of unspliced TSS-proximal/distal uaRNA/eRNA regions grouped by the presence and strength (MaxEnt score[41]) of a 5′SS motif (see Methods). None $n=2{,}632$, medium (MaxEnt 5–10) $n=1{,}554$, strong (MaxEnt10+) $n=106$. All groups are significantly different from each other ($P<0.0001$) by Kruskal-Wallis test. **e**, Density plot of GC-corrected steady-state RNA levels of unspliced TSS-proximal mRNA regions (left) and TSS-proximal/distal uaRNA/eRNA regions (right) grouped by the number of 5′SS motifs (MaxEnt score >5). mRNAs: none $n=1{,}392$, 1 $n=1{,}604$, >1 $n=691$. uaRNA/eRNAs: none $n=2{,}632$, 1 $n=1{,}232$, >1 $n=428$, comparisons by Kruskal-Wallis test. **f**, Relative nascent RNA levels (left) and sort-seq scores (right) of 10nt annotated 5′SSs with a MaxEnt score of >5, embedded into several background sequences. Only unspliced inserts (<3% spliced in nascent-RNA) were considered. Same groups as in Fig. 5d: scrambled (Scr, $n=24$) and antisense (AS, $n=24$) versions of 5′SSs were compared to sense (S) 5′SSs ($n=50$) by Kruskal=Wallis test. In all panels, **** indicates $P<0.0001$, higher P values are indicated in each plot.

nature portfolio

Corresponding author(s): Hanneke Vlaming

Last updated by author(s): 2022/04/07

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	- Electrophoresis gels: BioRad ChemiDoc with ImageLab version 6.0.0 build 26 - NGS: Illumina platforms as indicated for each sample on GEO with appropriate Illumina Realtime Analysis, base calling and demultiplexing software as used by The Bauer Core Facility at Harvard University, the HMS Biopolymers core or Novogene. - FACS: BD Biosciences FACSAria (Fig 4d) or Sony HS800S (Fig 5b) - qPCR: Bio-Rad CFX384 instrument

Data analysis

For INSERT-seq data mapping: Cutadapt v1.14, bowtie2 (version 2.3.4.3), STAR (version 2.7.0), Samtools v1.9, R (version 3.5.1) with package Biostrings_2.50.2.
Further data normalization outputting tables: R version 3.5.2 with the following packages: scales_1.0.0, stringr_1.4.0, Biostrings_2.50.2.
Motif searches: Homer version 4.10.3 with perl version 5.30.0, or MEME version 5.3.3 through https://meme-suite.org/meme/tools/meme
To run MOODS: python version 2.7.12, MOODS version 1.9.3. Followed by MaxEnt score calculation through the webtool http://hollywood.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html.
The data tables generated in R were copied into Graphpad Prism v9.1 for plotting and statistical tests.

Genomic profile snapshots were generated using the fluff (v.3.0.4, https://github.com/simonvh/fluff), using Conda (4.2.13) and Python (3.6.0).

PRO-seq, TT-seq and ChIP-seq data analysis: Bowtie v1.2.2, cutadapt v1.14, STAR v2.7.3a, Samtools v1.3.1.

All code is available on GitHub:
https://github.com/AdelmanLab/Vlaming2021_INSERT-seq_paper
https://github.com/AdelmanLab/NIH_scripts/tree/main/make_heatmap, https://github.com/AdelmanLab/NIH_scripts/tree/main/extract_fragments, https://github.com/AdelmanLab/NIH_scripts/tree/main/bedgraphs2stdBedGraph, https://github.com/benjaminmartin02/binBedGraph, https://github.com/AdelmanLab/NIH_scripts/tree/main/trim_and_filter_PE.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Raw and processed data files of all INSERT-seq experiments, PRO-seq, H3K4me3 ChIP-seq and TT-seq are available at Gene Expression Omnibus, accession no. GSE178230. H3K27ac ChIP-seq data is available through the 4DN data portal (https://data.4dnucleome.org/), ExperimentSet accession no. 4DNESQ33L4G7. H3K4me1 mESC ChIP-seq data was downloaded from Gene Expression Omnibus, accession no. GSE56138. Reference genome mm10 (GRCm38) can be downloaded using RefSeq assembly accession number GCF_000001635.20. Supplementary Tables 3-7 provide all normalized and averaged data from INSERT-seq experiments, as well as which inserts are included in which plot.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Large numbers of different inserts were tested in each screen, the number of inserts per class was limited by the total library size we could feasably screen.
2-4 biological replicates were performed for INSERT-seq screens. 4 biological replicates were performed for PRO-seq and 2 biological replicates for TT-seq and ChIP-seq experiments. No statistical methods were used to predetermine the sample size. The sample sizes as are those typically used in the field.
For measurements on clonal cell lines, we aimed to generate 4 independent clonal cell lines per genotype and used all validated cell lines that were obtained (2-4).

Data exclusions

No data was excluded, all data filtering is described in the methods and figure legends.

Replication

2 biological replicates were performed for INSERT-seq experiment based on nascent-RNA and chromatin and using the sort-seq readout. 4 biological replicates were performed for INSERT-seq with steady-state RNA readout in untreated cells and 3 replicates for siRNA-treated cells. Biological replicates showed agreement, as shown in the figures, and all generated data sets could be used in the paper.
For PRO-seq, 4 biological replicates were performed and checked for agreement before pooling the data. The same was done for the 2 biological replicates performed for TT-seq and ChIP-seq.

Randomization

Randomization and covariates were not applicable in this study. Proper controls were used and both the control and treatment were prepared in the same manner, and processed in parallel to eliminate batch effects.

Blinding

Blinding was not used in this study. Blinding was not required because the results of sequencing of nucleic acid libraries and other physical measurements of biomolecules is not affected by the experimenter's knowledge of sample identity.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
□	☒	Antibodies
□	☒	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
☒	□	Human research participants
☒	□	Clinical data
☒	□	Dual use research of concern

Methods

n/a	Involved in the study	
□	☒	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Antibodies

Antibodies used

EXOSC3 polyclonal (A303-909A, Bethyl Labs, 1:2,000 in western blot)
Actin polyclonal (sc-1616, SantaCruz, 1:2,000 in western blot)
goat anti-rabbit secondary: Peroxidase AffiniPure Goat Anti-Rabbit IgG (H+L) (Jackson ImmunoResearch, 111-035-144, 1:10,000)
donkey anti-goat secondary: Peroxidase AffiniPure Donkey Anti-Goat IgG (H+L) (Jackson ImmunoResearch, 705-035-147, 1:10,000)
H3K4me3 monoclonal (EpiCypher Cat#13-0028, lot #18303001, 16 uL for ChIP)
H3K27ac polyclonal (Active Motif cat#39133, lot#22618011, 10 uL for ChIP)

Validation

EXOSC3 signal decreases upon siEXOSC3 treatment (ED Fig 2a)
The Actin antibody has been used as a loading control in many different studies. Per www.citeab.com, this antibody has been used in over 4,000 publications, about a third of which in mouse cells.
Validation of the H3K4me3 antibody is available on manufacturer's website (https://www.epicypher.com/products/antibodies/snap-chip-certified-antibodies/histone-h3k4me3-antibody-snap-chip-certified) and in Suppl Fig 2 of Lam et al, Nature Comm, 2019 (doi: 10.1038/s41467-019-11820-7)
Validation of the H3K27ac antibody is available on manufacturer's website (https://www.activemotif.com/catalog/details/39133)

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

F121-9 hybrid mESC lines were obtained from the Gilbert lab through the 4DN consortium. Other cell lines were derivatives of this original cell line.
S2 cells were used for spike-in normalization in PRO-seq, and were obtained from the Drosophila Genomics Resource Center.

Authentication

Sequencing of SNPs confirmed that these were hybrid mESCs. Newly made clonal cell lines were checked by genotyping PCR and Sanger sequencing of the relevant allele.
S2 cells had been used in other transcriptomics studies that confirmed their identity, spike-in reads could be mapped to the Drosophila Melanogaster reference genome.

Mycoplasma contamination

All cell lines were tested negative for mycoplasma 3-4 times a year.

Commonly misidentified lines
(See ICLAC register)

No commonly misidentified lines were used in this study.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

□ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

H3K4me3 ChIP-seq: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178230 (this link also contains all INSERT-seq, PRO-seq and TT-seq data).
H3K27ac ChIP-seq data: https://data.4dnucleome.org/experiment-set-replicates/4DNESQ33L4G7/

Files in database submission

For H3K4me3 ChIP-seq 4 raw data files are provided per biological replicate: Read 1 + read 2 from two separate runs.
LIB040448_CHS00151390_R1.fastq.bz2
LIB040448_CHS00151390_R2.fastq.bz2
LIB041462_CHS00156008_R1.fastq.bz2
LIB041462_CHS00156008_R2.fastq.bz2

	LIB040448_CHS00151391_R1.fastq.bz2 LIB040448_CHS00151391_R2.fastq.bz2 LIB041462_CHS00156009_R1.fastq.bz2 LIB041462_CHS00156009_R2.fastq.bz2 For H3K27ac ChIP-seq, two fastq files per biological replicate are provided (read 1 + read 2) 4DNFIZMTL3II.fastq.gz 4DNFIE3KPPYK.fastq.gz 4DNFIO49NSRI.fastq.gz 4DNFILM2RTL3.fastq.gz The following Bigwig file is provided: mESC_F121-9.H3K4me3.ChIPseq.50bp.bins.bw
Genome browser session (e.g. UCSC)	Not provided, not relevant for INSERT-seq experiments.

Methodology

Replicates	Two biological replicates were performed for each ChIP-seq experiment. Agreement between the two data sets was checked before merging them, see data quality.
Sequencing depth	H3K4me3: PE 160-cycle kit (read1:80, index:6, read2:80) rep 1: 41830488 uniquely mapped reads, 9.27 % duplicates rep 2: 49169600 uniquely mapped, 8.925 % duplicates H3K27ac: PE 80-cycle kit (read1:43, index:6, read2:42) rep 1: total reads: 42160852; mapped reads (PE): 25044592; after frag filter: 24067803; %dups: 0.084281% rep 2: total reads: 35481506; mapped reads (PE): 22198607; after frag filter: 21328400; %dups: 0.068345%
Antibodies	H3K4me3 monoclonal (EpiCypher Cat#13-0028, lot #18303001, 16 uL for ChIP) H3K27ac polyclonal (Active Motif cat#39133, lot#22618011, 10 uL for ChIP)
Peak calling parameters	No peaks were called.
Data quality	Raw sequencing quality was assessed with FASTQC. To check reproducibility, correlations were calculated for counts over regions of TSS (gencode v99) +/- 1kb. For H3K4me3: spearman rho and pearson r both 0.99. For H3K27ac: spearman rho = 0.88; pearson r = 0.97.
Software	Custom code provided on Github: https://github.com/AdelmanLab/NIH_scripts/tree/main/extract_fragments, https://github.com/AdelmanLab/NIH_scripts/tree/main/bedgraphs2stdBedGraph, https://github.com/benjaminmartin02/binBedGraph, https://github.com/AdelmanLab/NIH_scripts/tree/main/trim_and_filter_PE.