



Supplementary Materials for

TT-seq maps the human transient transcriptome

Björn Schwalb,* Margaux Michel,* Benedikt Zacher,* Katja Frühauf, Carina Demel,
Achim Tresch, Julien Gagneur,† Patrick Cramer†

*These authors contributed equally to this work.

†Corresponding author. Email: gagneur@in.tum.de (J.G.); patrick.cramer@mpibpc.mpg.de (P.C.)

Published 3 June 2016, *Science* **352**, 1225 (2016)

DOI: 10.1126/science.aad9841

This PDF file includes:

Materials and Methods

Figs. S1 to S17

Tables S1 to S4

Full Reference List

Materials and Methods:

RNA-seq. K562 cells were acquired from DSMZ (Braunschweig, Germany). Cells were grown in RPMI 1640 medium (Gibco, Carlsbad, CA, USA) supplemented with 10% heat-inactivated FBS (Gibco) and 1% Penicillin/Streptomycin (100x, PAA now GE Healthcare, Chalfont St. Giles, UK) at 37°C under 5% CO₂. Cells were labeled in media for 5 min with 500 μM 4-thiouridine (4sU, Sigma-Aldrich, St. Louis, MO, USA.) and harvested through centrifugation for 2 min at 3,000 rpm. RNA extraction was performed with TRIzol (Life Technologies, Carlsbad, CA, USA) following the manufacturers' instructions except for the addition of an RNA spike-in mix together with TRIzol. The purified RNA was split in two samples and one of the two samples was fragmented at 240 ng/μl on a BioRuptor Next Gen (Diagenode, Seraing, Belgium) at high power for one cycle of 30''/30'' ON/OFF. Fragmented and non-fragmented samples were subjected to labeled RNA purification as previously described (26). Labeled fragmented (TT-seq), labeled (4sU-seq), total (RNA-seq) and total fragmented (RNA-seq with fragmentation) RNA were treated with 2 units of DNase Turbo (Life Technologies) and sequencing libraries were prepared with the Ovation Human Blood RNA-seq library kit (NuGEN, Carlos, CA, USA) following the manufacturers' instructions. All samples were sequenced on an HiSeq 1500 sequencer (Illumina, San Diego, CA, USA).

RNA spike-in mix preparation. Six spike-ins (ERCC-00043, ERCC-00170, ERCC-00136, ERCC-00145, ERCC-00092 and ERCC-00002) from the ERCC RNA spike-in mix (Life Technologies) were chosen as to have the same nucleotide length and U numbers, but with different GC content (40 to 60%). Spike-ins were amplified through PCR with the forward primer containing a T7 promoter sequence. Each spike-in was subjected to *in vitro* transcription with the Megascript T7 Transcription Kit (Life Technologies) with either 1:10 4sUTP:UTP ratio for spike-ins ERCC-00043, ERCC-00136 and ERCC-00092 or only UTP for spike-ins ERCC-00170, ERCC-00145, ERCC-00002; resulting in labeled and non-labeled spike-ins, respectively. All spike-ins were purified with AMPure XP beads (Beckman-Coulter, Fullerton, CA, USA) and quantified with Nanodrop 2000 (Thermo Scientific, Boston, MA, USA), agarose gel and Qubit 3.0 Fluorometer (Life Technologies). Spike-ins were then mixed in equal amount to a final concentration of 6 ng/μl.

Sequencing data processing. Paired-end 50 base reads with additional 6 base reads of barcodes were obtained for each of the samples, i.e. total RNA (RNA-seq), total fragmented RNA (RNA-seq with fragmentation), labeled RNA (4sU-seq) and labeled fragmented RNA (TT-seq). Reads were demultiplexed and mapped with STAR 2.3.0 (for RNA-seq) (27) and Bowtie 2.1.0. (28) (for ChIP-seq) to the hg20/hg38 (GRCh38) genome assembly (Human Genome Reference Consortium). Samtools (29) was used to quality filter SAM files, whereby alignments with MAPQ smaller than 7 (-q 7) were skipped and only proper pairs (-f99, -f147, -f83, -f163) were selected. Further data processing was carried out using the R/Bioconductor environment. For visualization purposes piled-up read counts for every genomic position were summed up over replicates, size factors for each condition were calculated as described (30) and used to correct for library size and sequencing depth variations.

Calculation of observed intron over exon ratio. Read counts for all GENCODE annotated exons and introns were calculated with HTSeq (31) from RNA-seq, 4sU-seq and TT-seq data. Comparison of intron read counts with respect to flanking exon read counts (both with an RPK of at least 10) yielded an average proportion of intron coverage to exon coverage of 0.08, 0.23, 0.6 for RNA-seq, 4sU-seq and TT-seq.

Calculation of theoretical exon, intron and mixture densities. Densities were calculated as relative average read frequencies around the first TSS, 3'SS and last pA site of a transcript in the following manner. Reads mapping the respective regions were piled up with the use of the GENCODE transcript annotation (version 22) under the assumption that every single exonic

nucleotide (exon density) or intronic nucleotide (intron density) has been sequenced exactly once. Mixture densities for 0.08, 0.23, 0.6 and 1 intron over exon ratio were obtained as relative average read frequencies of a combination of the 1.0x exon read frequencies with 0.08, 0.23, 0.6 and 1.0x the intron read frequencies. The transcript density represents the combination of 1.0x exon read frequencies with 1.0x intron read frequency, reflecting the theoretical situation where a transcript has been completely and uniformly sequenced only once. Accumulation of these mixed read frequencies in a metagene-wise fashion aligned at the designated loci results in the depicted exon, intron and mixture densities (Fig. S1A).

Chromatin state annotation. ENCODE ChIP-seq data was mapped to the hg20/hg38 (GRCh38) genome assembly (Human Genome Reference Consortium) using Bowtie 2.1.0 (28). Samtools (29) was used to quality filter SAM files, whereby alignments with MAPQ smaller than 7 (-q 7) were skipped. These data were single-end reads yielding a shift between the plus and the minus strand coverage. To obtain midpoint positions of the ChIP-seq fragments, the read ends were shifted in the appropriate direction by half the average fragment length as estimated by strand coverage cross-correlation. Next, ChIP-seq tracks were summarized by the number of fragment midpoints in consecutive bins of 200 bp width. The R/Bioconductor software package STAN (11) was used to learn a standard hidden Markov model with multivariate, discrete emission function of independent Poisson-Log-Normal distributions (one for each data track) using the Baum-Welch algorithm. 18 states were chosen because this provided a high resolution of chromatin states without overfitting. Note that model selection criteria like BIC or AIC do generally not apply in a genomic context since their objective is predictive performance rather than interpretability and thus they suggest an excessively high number of states. State labels (1_Low, 2_Weak_enhancers, etc.) were provided by inspection of the state emission probabilities.

Transcription Unit (TU) annotation. Genome-wide coverage was calculated from TT-seq fragment midpoints in consecutive 200 bp bins throughout the genome. Binning reduced the number of uncovered positions within expressed transcripts and increased the sensitivity for detection of lowly synthesized transcripts. A two-state hidden Markov model with a Poisson-Log-Normal emission distribution was learned in order to segment the genome into “transcribed” and “untranscribed” states, which yielded an initial prediction of 86,676 TUs. In order to filter out spurious predictions, we defined a threshold for minimal expression (RPK) based on TUs overlapping with annotated GRO-cap TSS (5). The threshold was optimized based on the Jaccard-Index, which resulted in 39,811 TUs with a minimal RPK of 15.5 (Fig. S4A). To further filter these, we required each TU to overlap with an annotated GRO-cap TSS, an annotated GENCODE transcript (version 22), or that the TSS of the TU overlaps with a prediction of an active promoter state (5_Weak_Promoter or 11_Strong_Promoter) or enhancer state (2_Weak_Enhancer or 15_Strong_Enhancer) from our chromatin state segmentation. 21,874 TUs were supported by at least one of these external data sets (Fig. S4B). Subsequently, TUs start and end sites were refined to nucleotide precision by finding borders of abrupt coverage increase or decrease between two consecutive segments in the two 200 bp bins located around the initially assigned start and stop sites via fitting a piecewise constant curve to the coverage profiles (whole fragments) for both replicates using the segmentation method from the R/Bioconductor package "tilingArray" (32).

Transcript sorting. We sorted each TU into one of the following seven classes: eRNA, sincRNA, asRNA, conRNA, uaRNA, lincRNA and mRNA. First, TUs reciprocally overlapping by at least 50% in the same strand a GENCODE annotation (version 22) were classified as mRNAs and lincRNAs. Next, TUs located on the opposite strand of either a mRNA or lincRNA were classified as asRNA – if the TSS was located > 1 kb downstream of the sense TSS – as uaRNA if its TSS was located < 1 kb upstream of the sense TSS – and as conRNA if its TSS was located < 1 kb downstream of the TSS. Each of the remaining TUs

did not overlap with GENCODE annotation and was classified into eRNA – if its TSS fell into an enhancer state – or as sincRNA – if its TSS fell into a promoter states. This resulted in 19,219 non-ambiguously classified RNAs on which the rest of the analysis was focused.

Estimation of RNA synthesis rates and half-lives. For all 19,219 classified TUs isoform-independent exonic regions were determined using a model for constitutive exons (33). Read counts for all features were calculated using HTSeq (31). To estimate rates of RNA synthesis and degradation, we used a statistical model that describes the read counts k_{ij} for gene i in sample j (TT-seq or (total cellular) RNA-seq samples) fragmented or total RNA-seq samples) by gene-specific amounts of labeled and unlabeled RNA amounts α_i , β_i . The model also includes a parameter L_i for the length of the respective feature, scaling factors σ_j that account for variations in sequencing depth, and cross-contamination rate ϵ_j that models the proportion of unlabeled reads purified in the TT-seq sample. The expectation of the number of reads k_{ij} was modeled as:

$$E(k_{ij}) = L_i \cdot \sigma_j \cdot (\alpha_i + \epsilon_j \beta_i)$$

Note that ϵ_j is set to 1 for the (total cellular) RNA samples. Sequencing depth σ_j and cross-contamination rate ϵ_j were calculated using the spike-ins data, setting $\alpha_i = 0$ and $\beta_i = 1$ for the unlabeled spike-ins and setting $\alpha_i = 1$ and $\beta_i = 0$ for the labeled spike-ins. The model was fitted by maximum likelihood assuming negative binomial distribution with dispersion parameters as calculated by DESeq2 (34). Having sequencing depth σ_j and cross-contamination rate ϵ_j estimated, the same model was applied to all TUs to provide estimates of the labeled and unlabeled RNA amounts α_i , β_i . These in turn were converted into synthesis and degradation rates (μ_i , λ_i) assuming first-order kinetics as in (35) using the following equations:

$$\alpha_i = \frac{\mu_i}{\lambda_i} \cdot (1 - e^{-\lambda_i t})$$

$$\alpha_i + \beta_i = \frac{\mu_i}{\lambda_i}$$

where $t = 5$ minutes. And therefore:

$$\lambda_i(t) = -\frac{1}{t} \cdot \log\left(\frac{\beta_i}{\alpha_i + \beta_i}\right)$$

$$\mu_i(t) = (\alpha_i + \beta_i) \cdot \lambda_i(t)$$

Note that this approach is customized for TT-seq and introduces a conceptually new interpretation of transcript stability because TT-seq involves the fragmentation of labeled RNAs prior to the purification of their labeled parts. Labeling and modeling approaches that were used so far quantify RNAs as newly synthesized despite the fact that they carry a non-negligible part of pre-existing RNA. This is introducing a bias especially towards longer genes given our short labeling pulse of 5 minutes. Thus this approach can be applied to estimate the local synthesis and degradation rates at any genomic position. When applied to a complete TU, it estimates the typical synthesis rate and half-life of nucleotide bonds within the TU. We think this is necessary given the complexity of the human genes with regard to the vast number of transcript isoforms and the elaborate nature of splicing events that all influence the per gene estimation of synthesis and decay. Note that TT-seq data did not exhibit the so-called labeling bias (Fig. S3C).

RNA structure and U1 motifs. The first 1000 nt from the RNA 5'-end (TSS) of each transcript was divided into 100 nt bins, where two successive bins were shifted by 50 nt. The free folding energy of each of these bins was calculated using RNAfold from the ViennaRNA

package (36) and the bin with the minimal free energy was selected for plotting as a measure for the most stable local structure within the region (Fig. S8C). Predicted structured RNAs in the human genome were selected from (37) and overlap with the TT-seq transcript annotation and half-lives were plotted (Fig. S9B, C). To analyze RNAs for the occurrence of U1 motifs, the 5'-most 1000 nt of each transcript were screened for occurrences of the consensus sequence of the U1 binding site (GGUAAG) and for those of the 5'-splice site GGUGAG and GUGAGU. Transcripts were then divided into 'zero' or 'one or more' occurrences and transcript lengths and half-lives were plotted.

Detection and analysis of transcription termination sites (TTSs). Two replicate measurements of the TT-seq sample were subjected to VST (variance stabilization transformation) to yield approximately homoscedastic data (34, 38). Size factors for each replicate were calculated as described (30) and used to correct for library size and sequencing depth variations. For each gene, the genomic region for site determination (potential termination window, PTW) was set from the last annotated pA site to the subsequent annotated TSS. PTWs exceeding 10 kb were trimmed down to 10 kb. 7,658 genes were selected for TTS determination based on two criteria, (i) the accumulated average coverage from whole fragments of the two replicates in the first 50 bp of the PTW was at least 2.5, and (ii) the PTW contained a consecutive stretch of at least 100 bp where no mapped fragment could be detected in both replicates. For each of the 7,658 TUs, putative TTSs were estimated by finding borders of abrupt coverage drops of at least 5-fold between consecutive segments via fitting a piecewise constant curve to the coverage profiles (whole fragments) for both replicates using the segmentation method from the R/Bioconductor package "tilingArray"(32). This yielded TTSs estimates for 6,977 TUs. The ultimate TTS was set to be the border dividing two segments where the mean coverage dropped to 0 at least 500 bp downstream of the last pA-site and could be derived for all 6,977 TUs. The putative strong TTS was set to be the border dividing two segments where the mean coverage dropped with the maximum difference at least 500 bp downstream of the last pA-site and could be derived for all 5,113 TUs.

TTS sequence motifs. The mean melting temperature for each window of +/-100 bp around the estimated TTSs was calculated as the gene-wise position based estimate of the melting temperature of 8-base pair DNA-RNA hybrids (39). K-mer enrichment analysis was performed comparing the frequency of all possible k-mers in a window of +/- 5 bp around the estimated TTS for $k = 3, \dots, 9$ against the background frequency of the respective k-mer obtained from 10 different windows of the same size shifted 25 bp to 125 bp in steps of 10 bp downstream, using Fisher's exact test, yielding virtually the same consensus sequence as enriched. Testing was also performed for fixed positions (Fig. S10E). K-mer enrichment analysis was additionally performed comparing the frequency of all possible GC-containing k-mers versus all possible AT-containing k-mers in a window of +/- 5 bp around the estimated TTS as described above. The enriched sequence motif $(C/G)_{(2-6)}A$ followed by $(T/A)_{(3-6)}$ is not found when the termination window length is scrambled (Fig. S14).

In vivo termination experiment. Control (AGTATCGAAGTCAGCAACTG), short/ $C_3AN_8T_4$ (AGTACCCAAGTCAGCATTTT) and long/ $C_7AN_8T_4$ (CCCCCCAAGTCAGCATTTT) sequences were inserted 4 times into the pCMV-GLuc 2 (NEB, Ipswich, MA, USA) plasmid in the 585 bp region after the polyA site of the *Gaussia* luciferase gene and before the SV40 promoter of the neomycin resistance gene using PCR (see Table 1 for cloning primers). The polyA site was removed in a similar fashion (Table 1). 2 μ g plasmid were transfected into 1×10^6 K652 cells using the SF Cell line 4D-Nucleofector X kit and unit from Lonza (Basel, Switzerland). RNA was isolated 4h post transfection using the RNeasy Plus Mini Kit (Qiagen, Hilden, Germany), reverse-transcribed (Maxima, Thermo Scientific) and qPCR was performed in technical triplicates using SYBR Select master mix (applied biosystems, Carlsbad, CA, USA) and appropriate primers (see Table 2 for qPCR primers). Ct-values of the

termination read-through readout were normalized for transfection efficiency using the qPCR readout assessing the luciferase transcript amount. Additionally, we controlled for a day-specific effect apparent on the termination read-through readout by setting day-specific control to 0 and normalizing respective values accordingly. Ct values were transformed to absolute numbers by taking it to the power of 2.

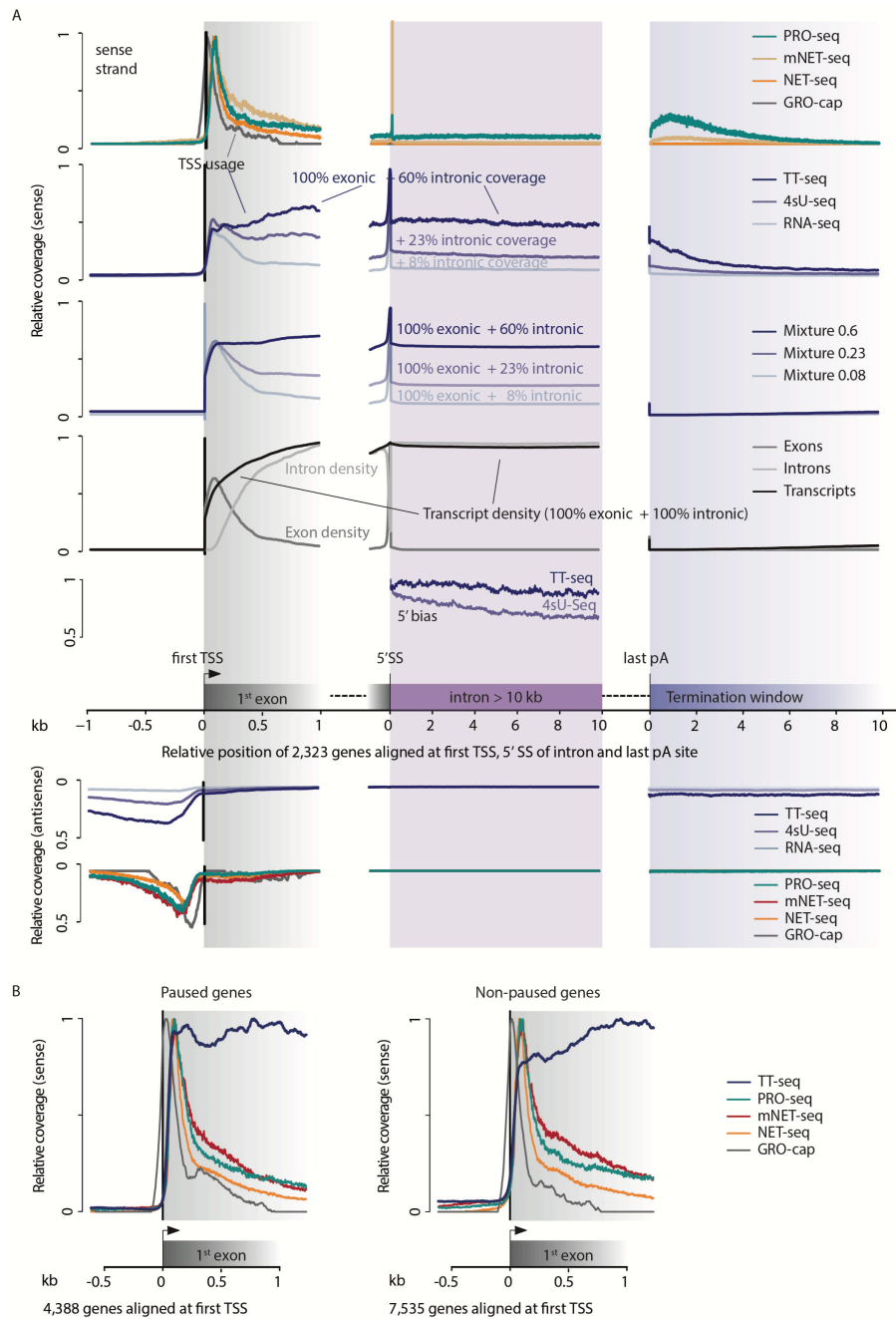


Fig. S1. TT-seq maps entire transcripts. **(A)** Metagenome analysis comparing TT-seq of 2,323 RNAs to different experimental methods and to theoretical exon, intron and mixture densities (4). The theoretical densities depict 8% (RNA-seq), 23% (4sU-seq) and 60% (TT-seq) intron coverage. Average coverage relative to the maximum in the first kb (left) is shown around the first TSS (left), the 5'-splice site (SS) of an intron of at least 10 kb (first intron excluded, middle), and the last pA site (right) for the sense (top panel) and antisense (bottom panel) strand or relative to the maximum in the first kb from the 5' SS (5' bias, top panel) **(B)** Metagenome analysis around the first TSS comparing TT-seq to different experimental methods for 4,388 genes classified as Class II paused genes in (10) (left) and for the complementary set of 7,535 non-paused genes (right).

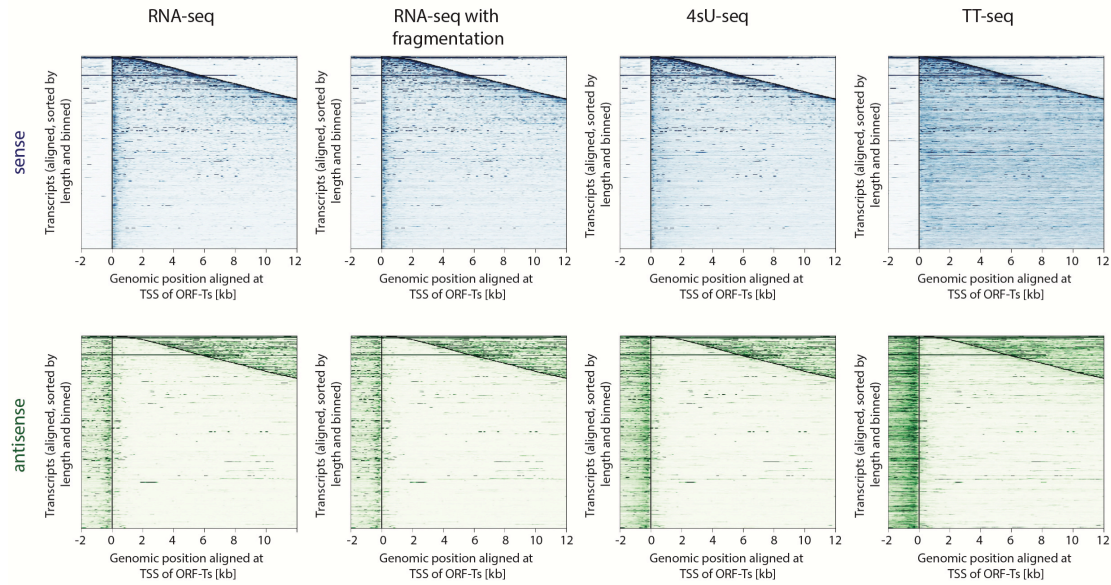


Fig. S2. TT-seq enables uniform sampling of pre-mRNAs and detects ncRNAs with high sensitivity. Coverage profiles of 2,500 most highly expressed mRNA transcripts aligned at the TSS, sorted by length and merged by averaging into horizontal bins of 10 mRNAs each for all measured samples with pooled replicates. The number of position-based read counts is color-coded, ranging from high (dark color) to low (light color). Upper and lower panels represent sense and antisense coverage, respectively.

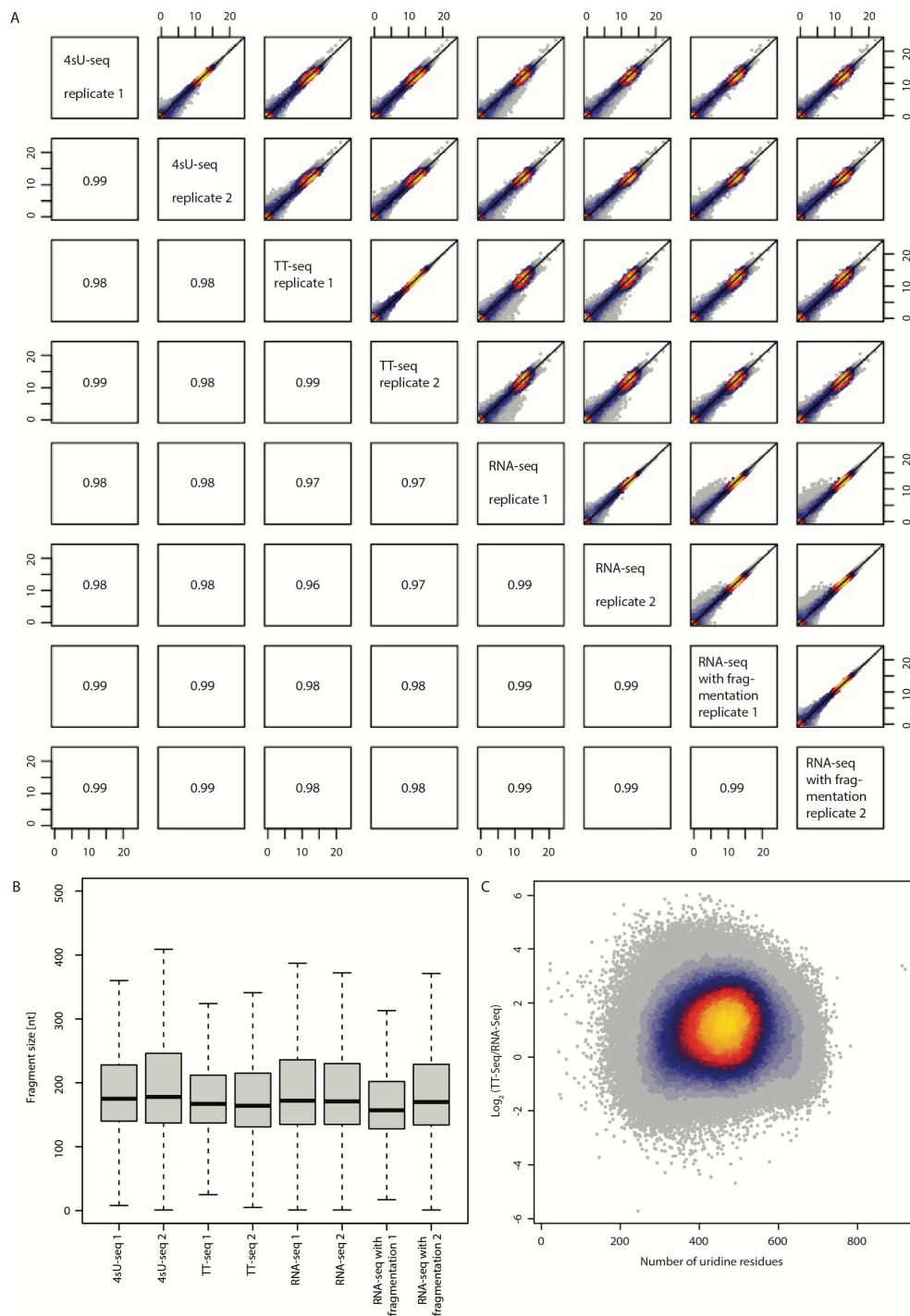


Fig. S3. TT-seq is highly reproducible, and RNA fragmentation does neither alter total RNA levels nor introduce a labeling bias. **(A)** Scatter plots comparing all measured samples in a pairwise fashion (upper triangle) and corresponding Spearman correlation (lower triangle). **(B)** Fragment size distributions for all measured samples. **(C)** Scatter plot comparing the number of uridine residues in bins of 1.5 kb against the respective log-ratio of TT-seq versus RNA-seq with fragmentation.

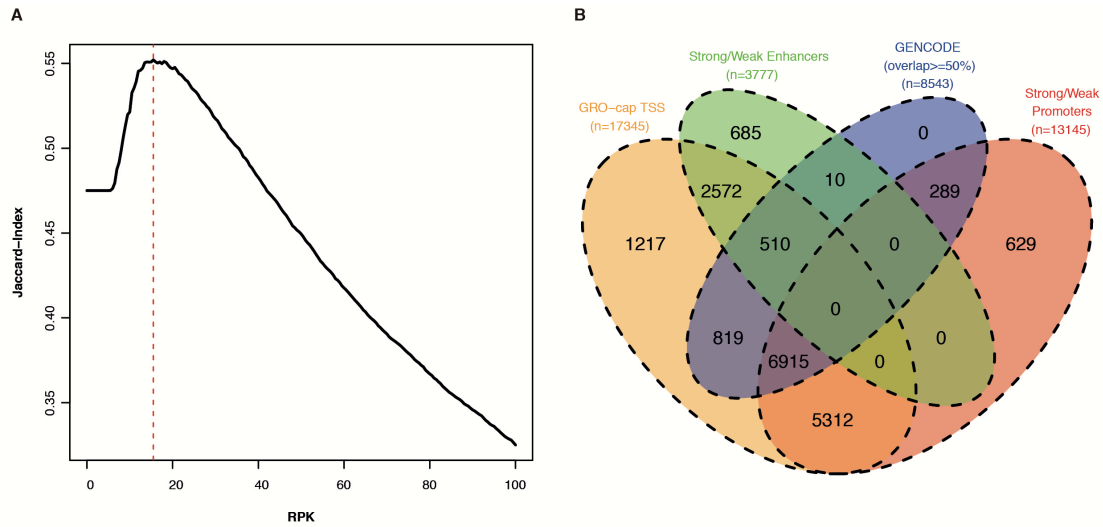


Fig. S4. Accurate annotation of transcripts based on TT-seq data using STAN **(A)** Jaccard index (compared to GENCODE annotation) for different choices of thresholds (RPK, x-axis). **(B)** Venn diagram showing the overlap of the predicted and filtered 21,874 TUs with external data sets.

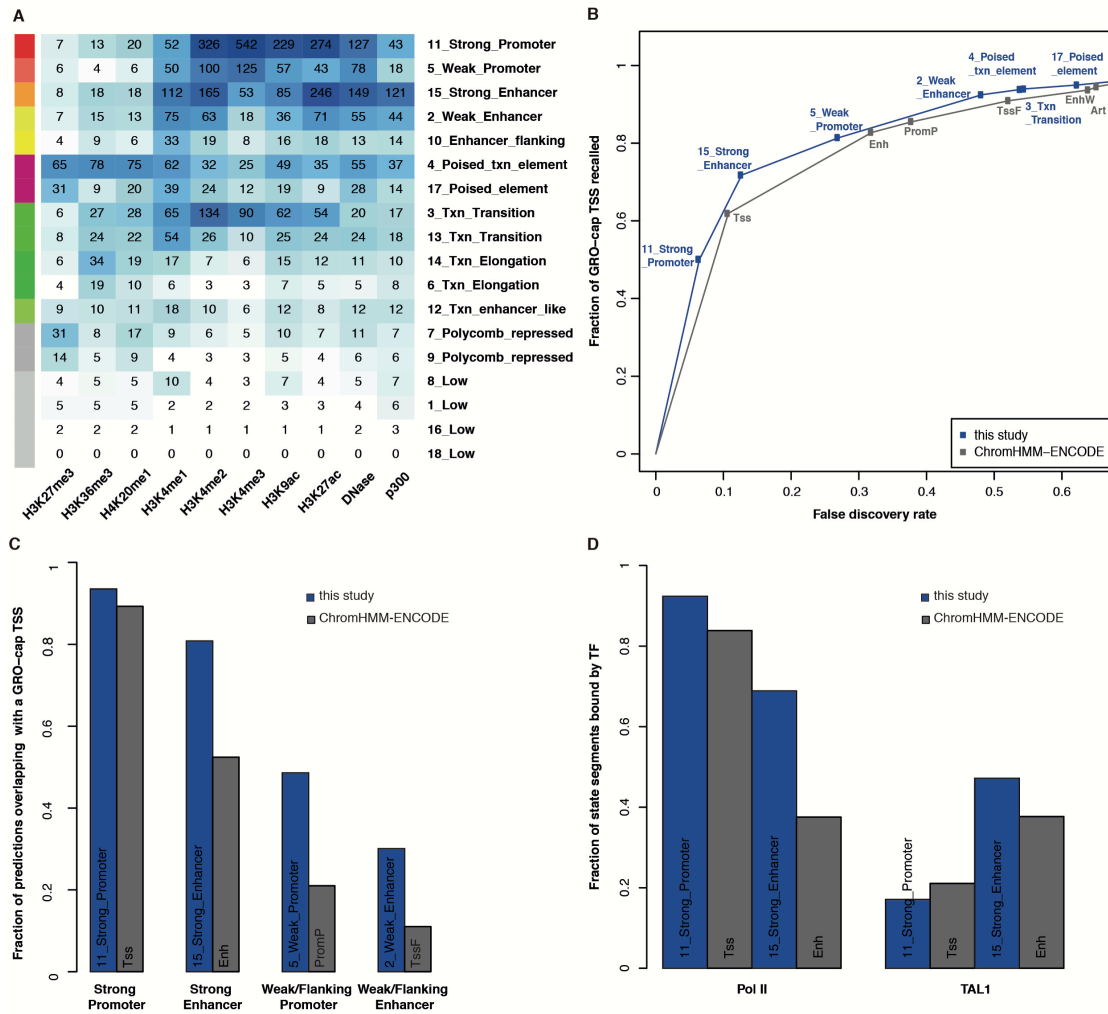


Fig. S5. Chromatin states and transcriptional activity of the corresponding genomic regions. **(A)** Mean read counts for each chromatin mark and state (Txn, Transcription). **(B)** The cumulative FDR is plotted against the recall for this study and ChromHMM-ENCODE. **(C)** Fraction of recovered (overlap) GRO-cap TSSs in 4 different states (strong promoter, strong enhancer, weak/flanking promoter and weak/flanking enhancer) in this study and in ChromHMM-ENCODE. **(D)** Fraction of (promoter and enhancer) chromatin states bound by Pol II and TAL1 are shown for the annotation in this study and the ENCODE ChromHMM annotation.

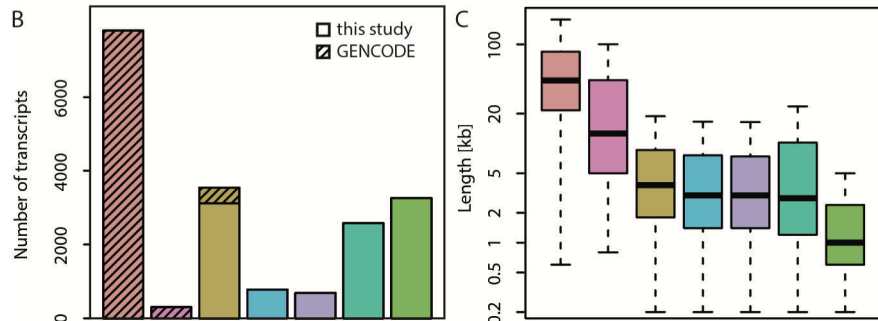
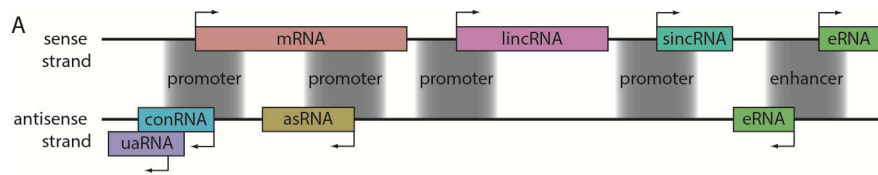


Fig. S6. Annotation and lengths of RNAs mapped by TT-seq. **(A)** Definition and color code of seven transcript classes. **(B)** Number of transcripts in different classes (portions covered by GENCODE hatched). **(C)** Distribution of transcript lengths.

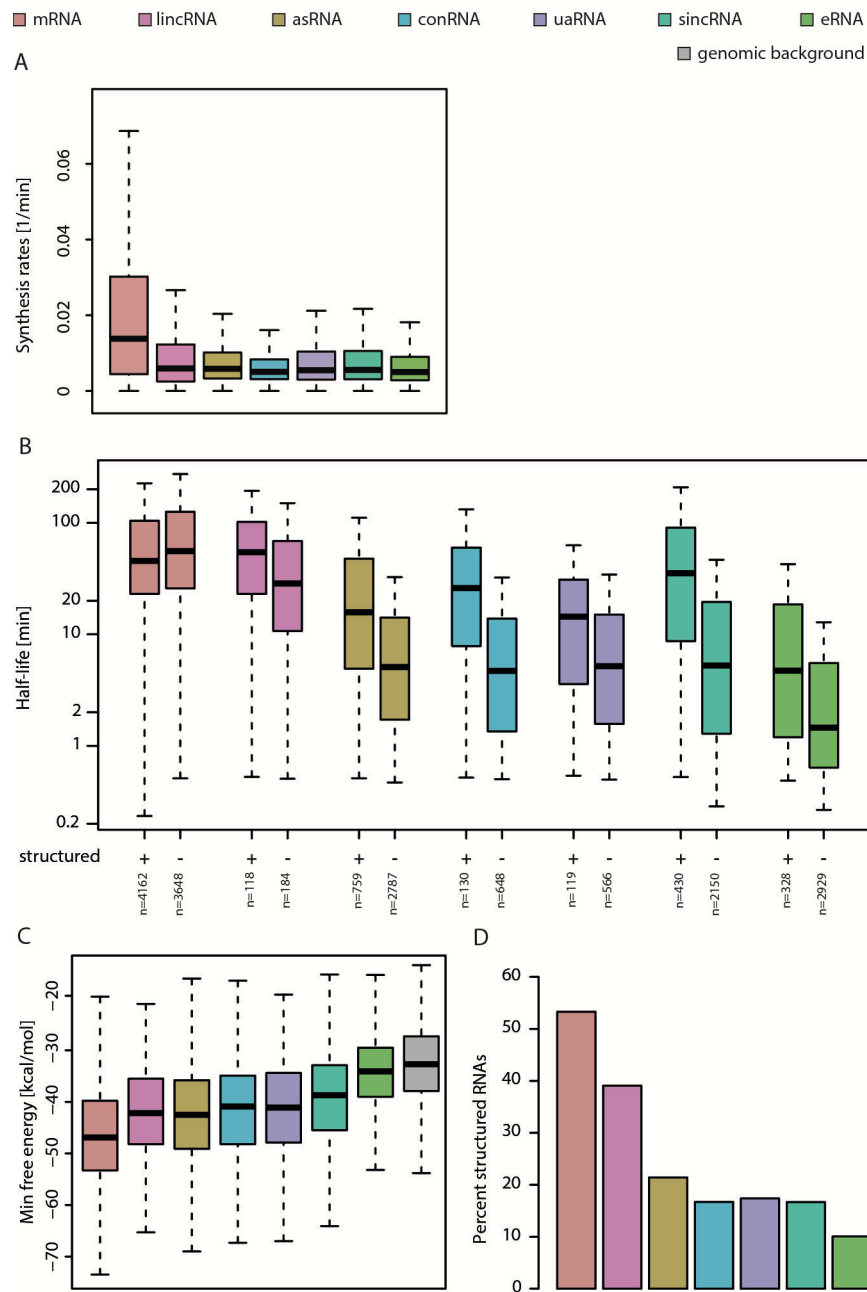


Fig. S7. Estimated transcript synthesis rates, half-lives, and predicted RNA structure. **(A)** Distribution of synthesis rates per transcript class. **(B)** Distribution of half-lives of different transcript classes depending on whether they are predicted to be structured or not (+, -) (37). **(C)** Distribution of the minimum free energy in the first 1000 nucleotides per transcript class. **(D)** Distribution of percentage of structured RNA in different transcript classes.

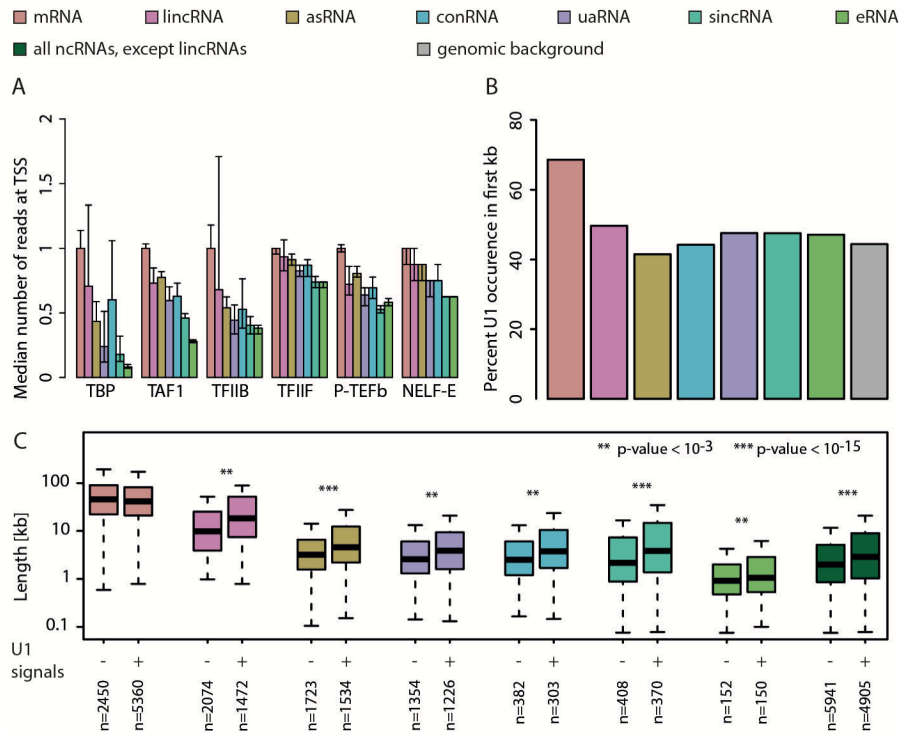


Fig. S8. RNA sequence features. **(A)** Distribution of relative peak occupancies with factors binding promoters (ENCODE, +/- 100 bp from TSS) for transcript classes. **(B)** Occurrence of U1 signal in the first 1000 nt for different transcript classes. **(C)** Distribution of transcript lengths in transcript classes depends on the presence of U1 signals in the first 1000 nt.

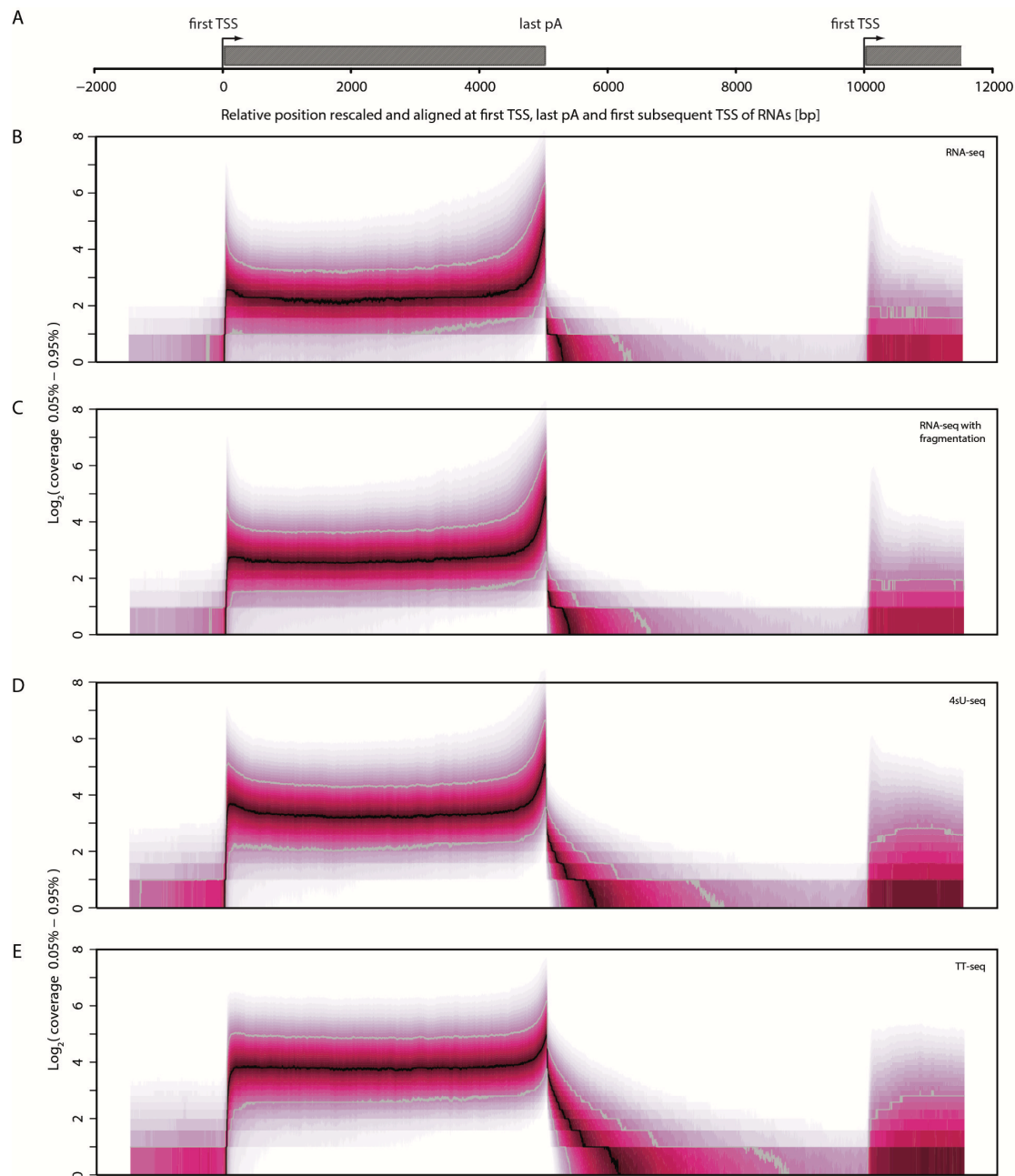


Fig. S9. Comparison of RNA synthesis downstream of the pA site in different RNA-seq experiments. (A) Generic gene architecture. Genomic position of the first TSS is aligned at 0. Last pA site is located at a distance of 5,000 bp from the first TSS for visualization purposes instead of a median of 24,079 bp for 6,977 investigated genes. The subsequent first TSS is depicted at a median distance of additional 5,000 bp from the last pA site. (B) RNA-seq coverage fanned out over 0.05% - 0.95% quantile range, rescaled and aligned according to schematic in (A) in logarithmic scale for 6,977 investigated genes. (C) RNA-seq with fragmentation coverage as in (B). (D) 4sU-seq coverage as in (B). (E) TT-seq coverage as in (B).

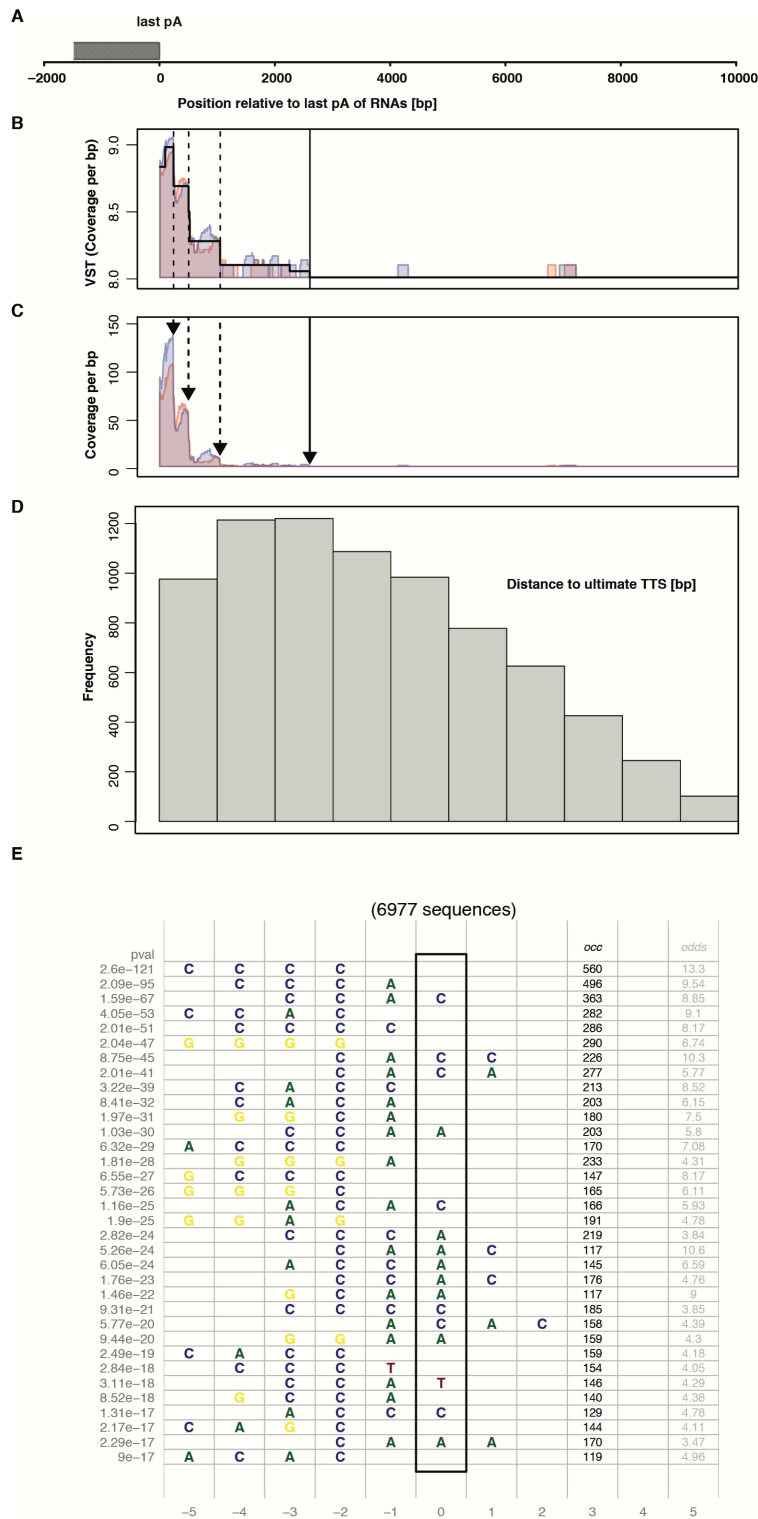


Fig. S10. Estimation of ultimate TTSs via segmentation of potential termination window. (A) Schematic of generic gene architecture showing the last annotated pA site (GENCODE). Genomic position of the last pA site is aligned at 0. The window for site determination is extended to the subsequent first TSS or max. 10 kb. (B) Variance stabilization transformed (VST) TT-seq coverage for two replicates (red, blue) at the ALDH1B1 gene loci. The black line depicts the optimal stepwise linear function estimated via a segmentation algorithm, the breakpoints are indicated by the dashed lines. (C) Plot shows the same loci as in (B) for non-VST normalized data. Dashed and solid arrows indicate putative and ultimate TTSs

resulting from the segmentation depicted in (B). **(D)** Distribution of estimated TTSs relative to last pA site for 6,977 investigated genes. **(E)** Plot showing the top 35 enriched 4-mers found by comparing the frequency of all possible 4-mers in a window of +/- 5 bp around the estimated TTS for fixed positions. Testing was done via Fisher's exact tests against the (background) frequency of the respective 4-mer obtained from a window of the same size shifted 30 bp downstream. The respective p-values and odd-ratios are given in the left and right panel.

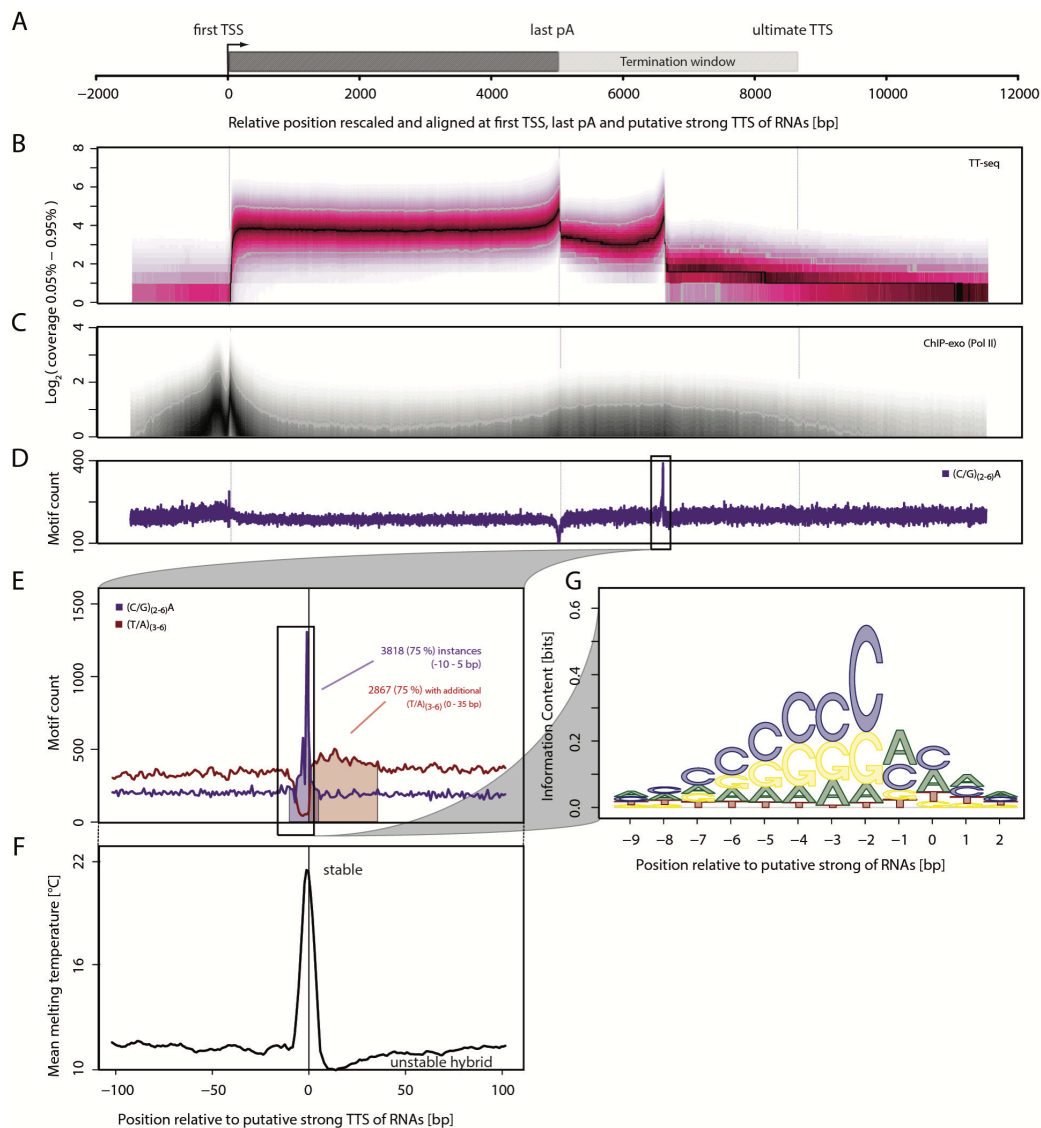


Fig. S11. Global detection and nature of TTSs inside the termination window. **(A)** Generic gene architecture. Genomic position of the first TSS is aligned at 0. The last pA site is located at a distance of 5,000 bp from TSS for visualization purposes for 5,113 investigated genes. The estimated ultimate TTS is depicted at a median distance of 3,359 bp from the last pA site. **(B)** TT-seq coverage fanned out over 0.05% - 0.95% quantiles, rescaled and aligned according to (A) in log-scale for 5,113 investigated genes with the exception that the putative strong TTS is used in a median distance of 1,593 bp from the last pA. **(C)** ChIP-exo Pol II occupancy (23) as in (B). **(D)** (C/G)₍₂₋₆₎A kmer count in the corresponding RNA sequence rescaled and aligned according to schematic in (A). **(E)** (C/G)₍₂₋₆₎A and (T/A)₍₃₋₆₎ kmer count in the corresponding RNA sequence in a window of +/- 100 bp around estimated putative strong TTS. **(F)** The mean melting temperature for a window of +/- 100 bp around the estimated putative strong TTS (4). **(G)** PWM logo representation of -9 to +2 bp of the corresponding RNA sequence around the putative strong TTS (position 0).

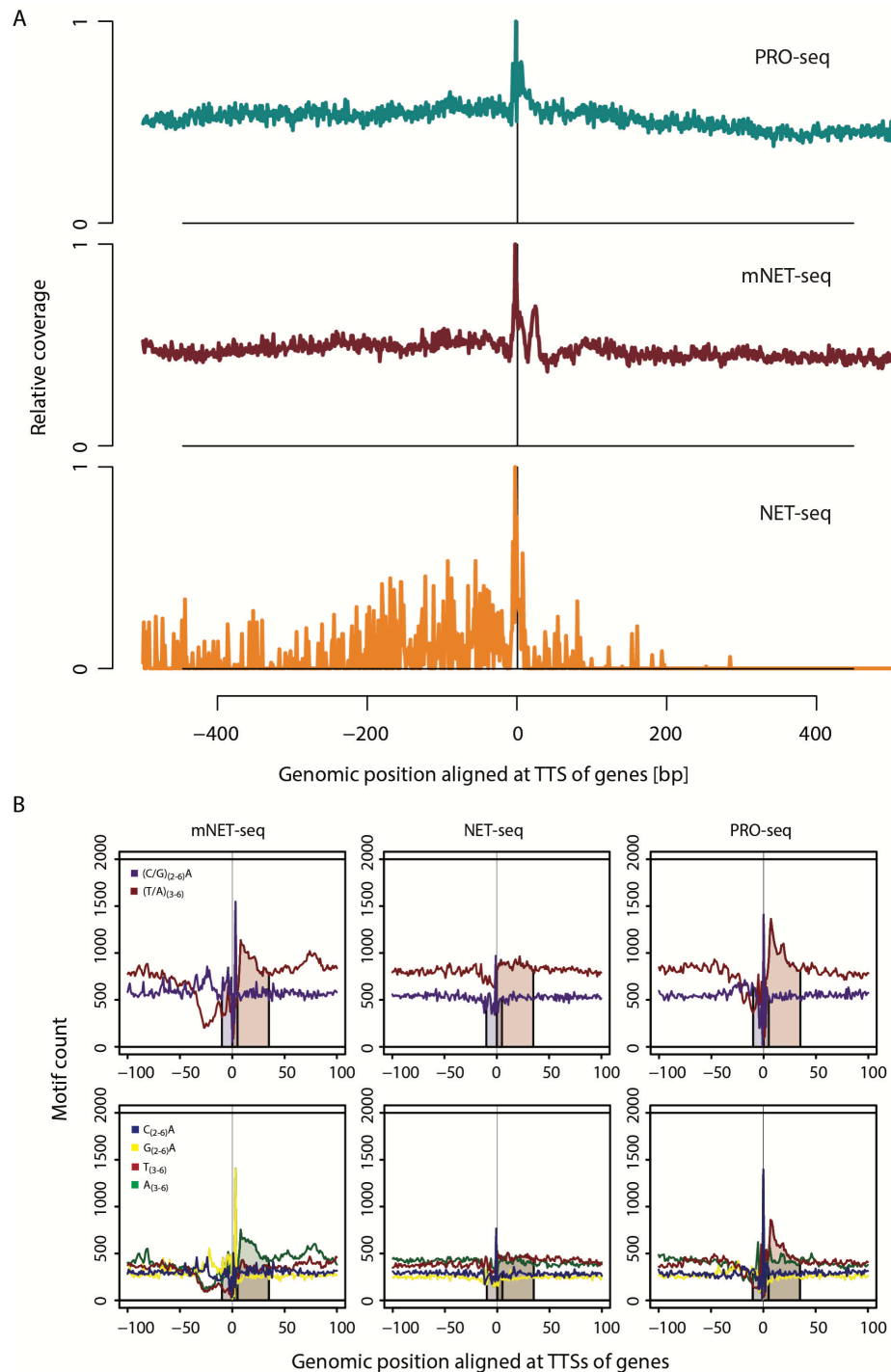


Fig. S12. The derived TTSs coincide with sites of polymerase pausing. **(A)** Average PRO-seq, mNET-seq and NET-seq signal is shown around the putative strong TTSs for 5,113 investigated genes relative to maximum. **(B)** $(C/G)_{(2-6)}A$ and $(T/A)_{(3-6)}$ kmer counts in the corresponding RNA sequence underlying putative strong TTSs derived from PRO-seq, mNET-seq and NET-seq for 14,060 investigated genes in a window of ± 100 bp (the TTS was set to be the position with the maximum number of 3' end read counts for all three methods between the last pA site and the next annotated downstream feature). Lower panel shows $C_{(2-6)}A$, $G_{(2-6)}A$, $T_{(3-6)}$ and $A_{(3-6)}$ kmer counts depicted analogously.

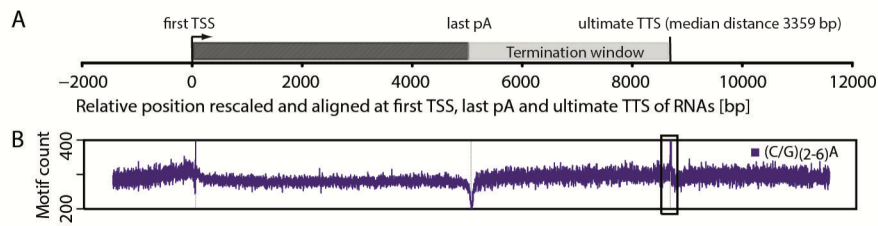


Fig. S13. Distribution of $(C/G)_{(2-6)}A$ kmers. **(A)** Generic gene architecture. The first TSS was aligned at 0, the last pA site was set at a rescaled distance of 5,000 bp from TSS (real median distance is 24,079 bp for 6,977 investigated genes) (4). The ultimate TTS is depicted at a median distance of 3,359 bp from the last pA site (rescaled). **(B)** $(C/G)_{(2-6)}A$ kmer sequence count rescaled and aligned as in (A).

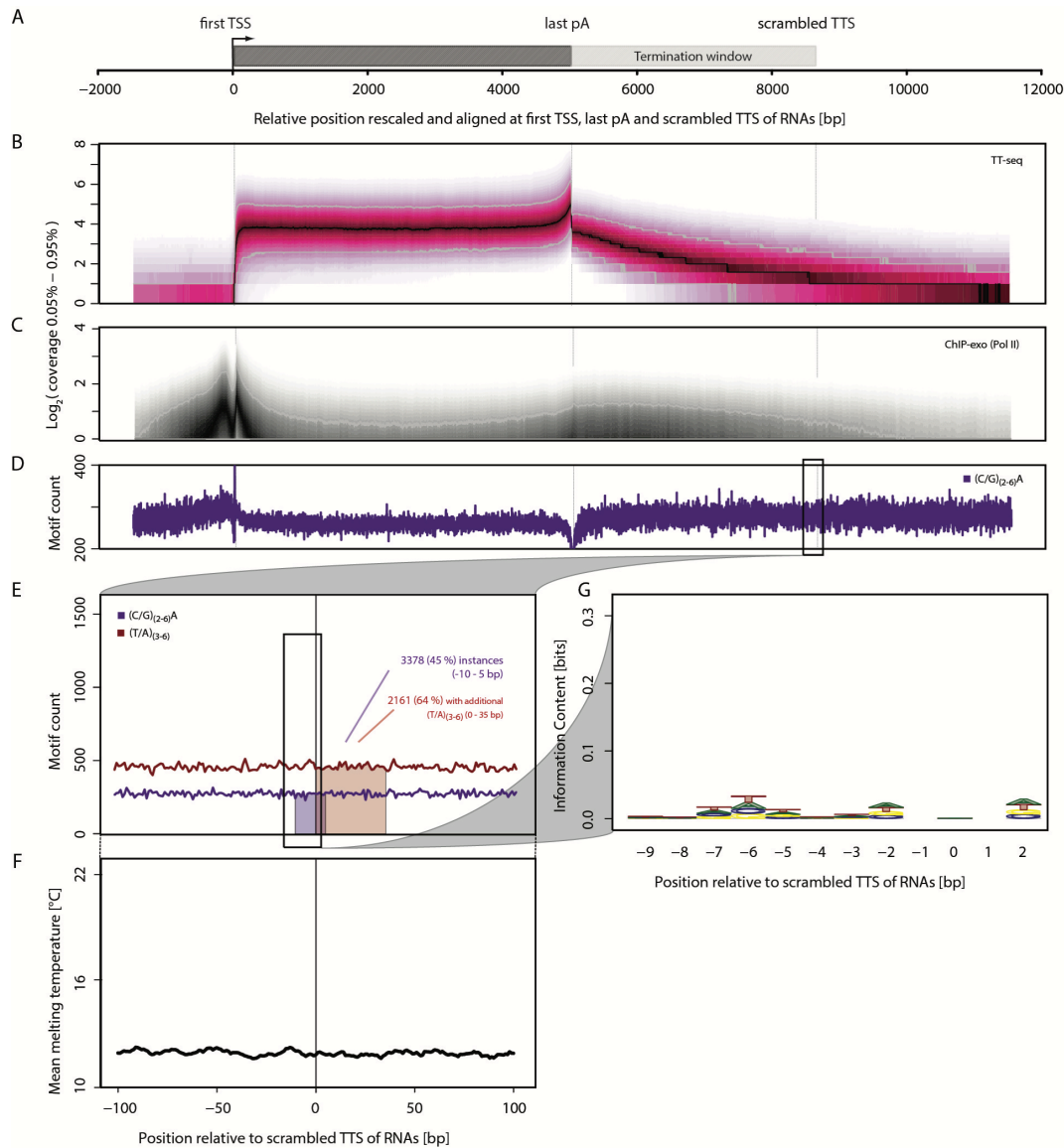


Fig. S14. Supporting evidence for derived TTSs. **(A)** Schematic of generic gene architecture. Genomic position of the first TSS is aligned at 0. The last pA site is located at a distance of 5,000 bp from the first TSS for visualization purposes instead of a median of 24,079 bp for 6,977 investigated genes. The estimated ultimate TTS is depicted at a median distance of 3,359 bp from the last pA site. **(B)** TT-seq coverage fanned out over 0.05% - 0.95% quantile range rescaled and aligned according to schematic in (A) in logarithmic scale for 6,977 investigated genes with the exception that the estimated distances from the last pA to the ultimate TTS are shuffled across all 6,977 genes. **(C)** ChIP-exo (Pol II) coverage (23) as in (B). **(D)** (C/G)₂₋₆A kmer count in the corresponding RNA sequence rescaled and aligned according to schematic in (A). **(E)** (C/G)₂₋₆A and (T/A)₃₋₆ kmer count in the corresponding RNA sequence in a window of +/- 100 bp around estimated shuffled ultimate TTS. **(F)** The mean melting temperature for a window of +/-100 bp around the estimated shuffled ultimate TTS (4). **(G)** PWM logo representation of -9 to +2 bp of the corresponding RNA sequence around the shuffled ultimate TTS (position 0).

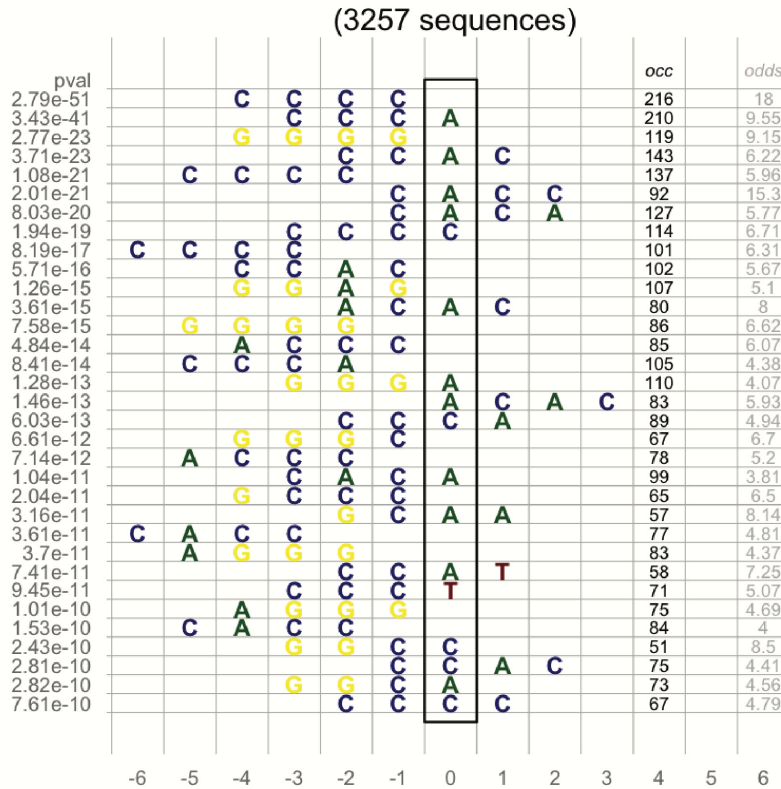
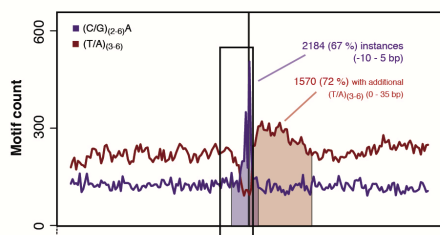
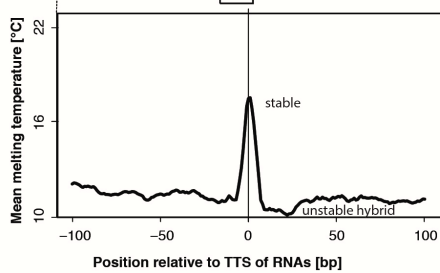
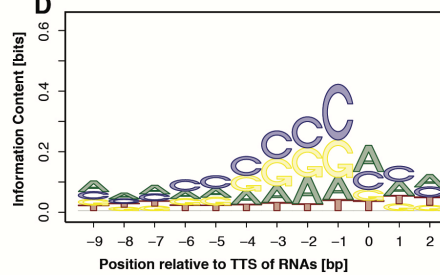
A**B****C****D**

Fig. S15. The 3'-end of eRNAs contains TTS motifs. **(A)** Plot showing the top 35 enriched 4-mers found by comparing the frequency of all possible 4-mers in a window of ± 5 bp around the estimated TTS for fixed positions. Testing was done via Fisher's exact tests against the (background) frequency of the respective 4-mer obtained from a window of the same size shifted 30 bp downstream. The respective p-values and odd-ratios are given in the left and right panel. **(B)** $(C/G)_{(2-6)}A$ and $(T/A)_{(3-6)}$ kmer count in the corresponding RNA sequence in a window of ± 100 bp around estimated TTS. **(C)** The mean melting temperature for a window of ± 100 bp around the estimated TTS. **(D)** PWM logo representation of -9 to 2 bp of the corresponding RNA sequence around the TTS.

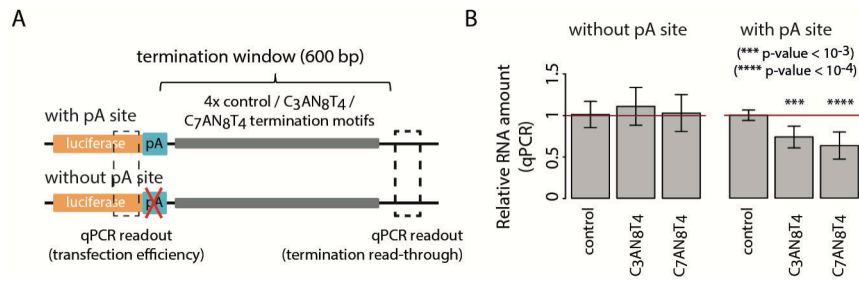


Fig. S16. Experimental support for the functionality of the derived TTS motifs. **(A)** Schematic of *in vivo* transcription assay to test the TTS motif. **(B)** Barplot showing the relative RNA abundance (qPCR) downstream of four TTS motifs relative to a control sequence without and with pA site.

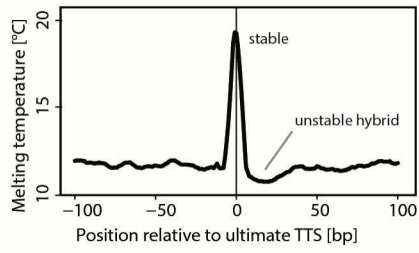


Fig. S17. Distribution of hybrid stability at TTS. The mean melting temperature for each window of +/-100 bp around the estimated TTSs was calculated as the gene-wise position based estimate of the melting temperature of 8-base pair DNA-RNA hybrids

Table S1: primers used for cloning

Name:	Sequence (5'-3'):	5' phosphorylation:
Insert 1 control forward	AGTATCGAAGTCAGCAACTGgcggtgggctctatggctt	yes
Insert 2 control forward	AGTATCGAAGTCAGCAACTGcccgcctcttcgctttctcc	yes
Insert 3 control forward	AGTATCGAAGTCAGCAACTGcttgattagggtgatggtcacgtagtg	yes
Insert 4 control forward	AGTATCGAAGTCAGCAACTGttataaggatttggggatttcggc	yes
Insert 1 C ₃ AN ₈ T ₄ forward	AGTACCCAAGTCAGCATTTTgcggtgggctctatggctt	yes
Insert 2 C ₃ AN ₈ T ₄ forward	AGTACCCAAGTCAGCATTTTcccgcctcttcgctttctcc	yes
Insert 3 C ₃ AN ₈ T ₄ forward	AGTACCCAAGTCAGCATTTTcttgattagggtgatggtcacgtagtg	yes
Insert 4 C ₃ AN ₈ T ₄ forward	AGTACCCAAGTCAGCATTTTttataaggatttggggatttcggc	yes
Insert 1 C ₇ AN ₈ T ₄ forward	CCCCCCCAAGTCAGCATTTTgcggtgggctctatggctt	yes
Insert 2 C ₇ AN ₈ T ₄ forward	CCCCCCCAAGTCAGCATTTTcccgcctcttcgctttctcc	yes
Insert 3 C ₇ AN ₈ T ₄ forward	CCCCCCCAAGTCAGCATTTTcttgattagggtgatggtcacgtagtg	yes
Insert 4 C ₇ AN ₈ T ₄ forward	CCCCCCCAAGTCAGCATTTTttataaggatttggggatttcggc	yes
Insert 1 reverse	tcccagcatgcctgctattg	no
Insert 2 reverse	cgctagggcgctggcaag	no

Insert 3 reverse	tcgaggtgccgtaaagcactaaatc	no
Insert 4 reverse	atcaaaagaatagaccgagatagggttgag	no
Remove pA forward	ttcttactgtcatgcccaagtaagatgctt	yes
Remove pA reverse	gcggccgcttagtcacca	no

Table S2: primers used for qPCR analysis

Name:	Sequence (5'-3'):
Transfection efficiency readout forward	CCTCAAAGGGCTTGCCAACG
Transfection efficiency readout reverse	CCTTGATCTTGTCCACCTGGC
Termination read-through readout forward	GTGGAATGTGTGTCAGTTAGGGTG
Termination read-through readout reverse	GACTTCCACACCTGGTTGCT

Table S3: Ct-values from qPCR experiments

Samples	Transfection efficiency primers – Ct values, technical replicates				Termination read-through primers - Ct values, technical replicates			
	1	2	3	4	1	2	3	4
pA_Control_A	12.86	12.66	13.11		17.07	17.1	17.04	
pA_Control_B	12.8	12.56	12.69		17.27	17.32	17.03	
pA_Control_C	14.08	14.12	14.1		20.66	20.7	20.81	
pA_C ₃ AN ₈ T ₄ _A	13.07	13.21	13.34		18.05	18.21	18.17	
pA_C ₃ AN ₈ T ₄ _B	13.15	13.02	13.09		17.79	17.67	17.93	
pA_C ₃ AN ₈ T ₄ _C	14.33	14.28	14.35		21.37	21.37	21.44	
pA_C ₇ AN ₈ T ₄ _A	13.05	13.04	13.27		17.55	17.85	17.89	
pA_C ₇ AN ₈ T ₄ _B	13.33	13.42	13.36		18.39	18.28	18.33	
pA_C ₇ AN ₈ T ₄ _C	12.36	12.52	12.47		20.2	20.32	20.39	
no_pA_Control1_A	11.73	12.38	12.31	12.45	13.04	13.4	13.27	13.31
no_pA_Control1_B	11.64	12.53	12.32	12.26	12.92	13.23	13.34	13.77
no_pA_Control2_A	13.4	13.79			14.67	14.93		
no_pA_C ₃ AN ₈ T ₄ _1_A	11.69	12.41	12.33	12.27	13.03	13.23	13.8	13.36
no_pA_C ₃ AN ₈ T ₄ _1_B	12.01	12.64	12.6	12.69	13.02	13.27	13.3	13.46
no_pA_C ₃ AN ₈ T ₄ _2_A	13.11	13.11			14.24	14.02		
no_pA_C ₇ AN ₈ T ₄ 1_A	11.94	12.63	12.62	12.76	13.37	13.89	13.8	13.84
no_pA_C ₇ AN ₈ T ₄ 1_B	11.62	12.51	12.46	12.51	12.93	13.41	13.28	13.7
no_pA_C ₇ AN ₈ T ₄ 2_A	13.78	13.59			14.68	14.48		

Table S4: normalized values from qPCR experiment:

Samples	normalized qPCR values - technical replicates			
	1	2	3	4
pA_Control_A	1	0.979	1.021	NA
pA_Control_B	0.957	0.924	1.13	NA
pA_Control_C	1.045	1.016	0.942	NA
pA_C ₃ AN ₈ T ₄ _A	0.637	0.57	0.586	NA
pA_C ₃ AN ₈ T ₄ _B	0.883	0.959	0.801	NA
pA_C ₃ AN ₈ T ₄ _C	0.744	0.744	0.709	NA
pA_C ₇ AN ₈ T ₄ _A	0.849	0.689	0.671	NA
pA_C ₇ AN ₈ T ₄ _B	0.709	0.765	0.739	NA
pA_C ₇ AN ₈ T ₄ _C	0.458	0.421	0.401	NA
no_pA_Control1_A	1.161	0.904	0.99	0.963
no_pA_Control1_B	1.315	1.061	0.983	0.73
no_pA_Control2_A	1.094	0.914	NA	NA
no_pA_C ₃ AN ₈ T ₄ _1_A	1.135	0.988	0.665	0.903
no_pA_C ₃ AN ₈ T ₄ _1_B	1.508	1.268	1.242	1.111
no_pA_C ₃ AN ₈ T ₄ _2_A	1.053	1.227	NA	NA
no_pA_C ₇ AN ₈ T ₄ 1_A	1.113	0.776	0.826	0.804
no_pA_C ₇ AN ₈ T ₄ 1_B	1.388	0.995	1.089	0.814
no_pA_C ₇ AN ₈ T ₄ 2_A	1.157	1.329	NA	NA

Author contributions: MM designed and carried out all experiments. BS carried out all bioinformatics analysis except transcript calling, RNA classification, analysis of U1 sequence motifs, and prediction of RNA secondary structure, which were carried out by BZ. BZ, JG and AT developed the chromatin state annotation. KF designed RNA spike-in probes. CD established the spike-in normalization method. BS, JG and PC designed research. JG and PC supervised research. BS, MM, JG, and PC prepared the manuscript, with input from all authors.

References and Notes

1. T. H. Jensen, A. Jacquier, D. Libri, Dealing with pervasive transcription. *Mol. Cell* **52**, 473–484 (2013). [Medline doi:10.1016/j.molcel.2013.10.032](#)
2. R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. Jørgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Müller, FANTOM Consortium, A. R. R. Forrest, P. Carninci, M. Rehli, A. Sandelin, An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014). [Medline doi:10.1038/nature12787](#)
3. M. D. Cleary, C. D. Meiering, E. Jan, R. Guymon, J. C. Boothroyd, Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nat. Biotechnol.* **23**, 232–237 (2005). [Medline doi:10.1038/nbt1061](#)
4. Materials and methods are available as supplementary materials on *Science* Online.
5. L. J. Core, A. L. Martins, C. G. Danko, C. T. Waters, A. Siepel, J. T. Lis, Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014). [Medline doi:10.1038/ng.3142](#)
6. R. Kodzius, M. Kojima, H. Nishiyori, M. Nakamura, S. Fukuda, M. Tagami, D. Sasaki, K. Imamura, C. Kai, M. Harbers, Y. Hayashizaki, P. Carninci, CAGE: Cap analysis of gene expression. *Nat. Methods* **3**, 211–222 (2006). [Medline doi:10.1038/nmeth0306-211](#)
7. H. Kwak, N. J. Fuda, L. J. Core, J. T. Lis, Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013). [Medline doi:10.1126/science.1229386](#)
8. A. Mayer, J. di Iulio, S. Maleri, U. Eser, J. Vierstra, A. Reynolds, R. Sandstrom, J. A. Stamatoyannopoulos, L. S. Churchman, Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* **161**, 541–554 (2015). [Medline doi:10.1016/j.cell.2015.03.010](#)
9. T. Nojima, T. Gomes, A. R. Grosso, H. Kimura, M. J. Dye, S. Dhir, M. Carmo-Fonseca, N. J. Proudfoot, Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **161**, 526–540 (2015). [Medline doi:10.1016/j.cell.2015.03.027](#)
10. L. J. Core, J. J. Waterfall, J. T. Lis, Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008). [Medline doi:10.1126/science.1162228](#)
11. B. Zacher, M. Michel, B. Schwalb, P. Cramer, A. Tresch, J. Gagneur, <http://biorxiv.org/content/early/2016/03/06/041020> (2016).
12. J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M.

- Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, T. J. Hubbard, GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012). [Medline doi:10.1101/gr.135350.111](#)
13. S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Röder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakraborty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigó, T. R. Gingeras, Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012). [Medline doi:10.1038/nature11233](#)
 14. T. K. Kim, M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, E. Markenscoff-Papadimitriou, D. Kuhl, H. Bito, P. F. Worley, G. Kreiman, M. E. Greenberg, Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010). [Medline doi:10.1038/nature09033](#)
 15. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012). [Medline](#)
 16. R. A. Flynn, A. E. Almada, J. R. Zamudio, P. A. Sharp, Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 10460–10465 (2011). [Medline doi:10.1073/pnas.1106630108](#)
 17. M. Rabani, R. Raychowdhury, M. Jovanovic, M. Rooney, D. J. Stumpo, A. Pauli, N. Hacohen, A. F. Schier, P. J. Blackshear, N. Friedman, I. Amit, A. Regev, High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell* **159**, 1698–1710 (2014). [Medline doi:10.1016/j.cell.2014.11.015](#)
 18. D. Kaida, M. G. Berg, I. Younis, M. Kasim, L. N. Singh, L. Wan, G. Dreyfuss, U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**, 664–668 (2010). [Medline doi:10.1038/nature09479](#)
 19. M. G. Berg, L. N. Singh, I. Younis, Q. Liu, A. M. Pinto, D. Kaida, Z. Zhang, S. Cho, S. Sherrill-Mix, L. Wan, G. Dreyfuss, U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150**, 53–64 (2012). [Medline doi:10.1016/j.cell.2012.05.029](#)
 20. A. E. Almada, X. Wu, A. J. Kriz, C. B. Burge, P. A. Sharp, Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360–363 (2013). [Medline doi:10.1038/nature12349](#)

21. S. West, N. Gromak, N. J. Proudfoot, Human 5' → 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature* **432**, 522–525 (2004). [Medline doi:10.1038/nature03035](#)
22. M. Kim, N. J. Krogan, L. Vasiljeva, O. J. Rando, E. Nedeá, J. F. Greenblatt, S. Buratowski, The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. *Nature* **432**, 517–522 (2004). [Medline doi:10.1038/nature03041](#)
23. B. J. Venters, B. F. Pugh, Genomic organization of human transcription initiation complexes. *Nature* **502**, 53–58 (2013). [Medline](#)
24. R. Ashfield, A. J. Patel, S. A. Bossone, H. Brown, R. D. Campbell, K. B. Marcu, N. J. Proudfoot, MAZ-dependent termination between closely spaced human complement genes. *EMBO J.* **13**, 5656–5667 (1994). [Medline](#)
25. M. L. Kireeva, N. Komissarova, D. S. Waugh, M. Kashlev, The 8-nucleotide-long RNA:DNA hybrid is a primary stability determinant of the RNA polymerase II elongation complex. *J. Biol. Chem.* **275**, 6530–6536 (2000). [Medline doi:10.1074/jbc.275.9.6530](#)
26. L. Dölken, Z. Ruzsics, B. Rädle, C. C. Friedel, R. Zimmer, J. Mages, R. Hoffmann, P. Dickinson, T. Forster, P. Ghazal, U. H. Koszinowski, High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA* **14**, 1959–1972 (2008). [Medline doi:10.1261/rna.1136108](#)
27. A. Dobin, T. R. Gingeras, Mapping RNA-seq reads with STAR. *Curr. Protoc. Bioinformatics* **51**, 11.14.1–11.14.19 (2015). [Medline](#)
28. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012). [Medline doi:10.1038/nmeth.1923](#)
29. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). [Medline doi:10.1093/bioinformatics/btp352](#)
30. S. Anders, W. Huber, Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010). [Medline doi:10.1186/gb-2010-11-10-r106](#)
31. S. Anders, P. T. Pyl, W. Huber, HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015). [Medline doi:10.1093/bioinformatics/btu638](#)
32. W. Huber, J. Toedling, L. M. Steinmetz, Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* **22**, 1963–1970 (2006). [Medline doi:10.1093/bioinformatics/btl289](#)
33. J. H. Bullard, E. Purdom, K. D. Hansen, S. Dudoit, Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010). [Medline doi:10.1186/1471-2105-11-94](#)
34. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014). [Medline doi:10.1186/s13059-014-0550-8](#)

35. C. Miller, B. Schwalb, K. Maier, D. Schulz, S. Dümcke, B. Zacher, A. Mayer, J. Sydow, L. Marcinowski, L. Dölken, D. E. Martin, A. Tresch, P. Cramer, Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol. Syst. Biol.* **7**, 458 (2011). [Medline](#) [doi:10.1038/msb.2010.112](https://doi.org/10.1038/msb.2010.112)
36. R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdisen, H. Tafer, C. Flamm, P. F. Stadler, I. L. Hofacker, ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011). [Medline](#) [doi:10.1186/1748-7188-6-26](https://doi.org/10.1186/1748-7188-6-26)
37. S. Washietl, I. L. Hofacker, M. Lukasser, A. Hüttenhofer, P. F. Stadler, Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.* **23**, 1383–1390 (2005). [Medline](#) [doi:10.1038/nbt1144](https://doi.org/10.1038/nbt1144)
38. W. Huber, A. von Heydebreck, H. Sülmann, A. Poustka, M. Vingron, Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, S96–S104 (2002). [Medline](#) [doi:10.1093/bioinformatics/18.suppl_1.S96](https://doi.org/10.1093/bioinformatics/18.suppl_1.S96)
39. J. SantaLucia Jr., A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 1460–1465 (1998). [Medline](#) [doi:10.1073/pnas.95.4.1460](https://doi.org/10.1073/pnas.95.4.1460)