Mapping Human Transient Transcriptomes Using Single Nucleotide Resolution 4sU Sequencing (SNU-Seq)

Philipp Lorenz, Anna Lamstaes, Harry Fischl, Shidong Xi, Aksel Saukko-Paavola, Struan Murray, Thomas Brown, Charlotte George, Andre Furger, Andrew Angel, Jane Mellor*

Department of Biochemistry, University of Oxford, South Parks Road, Oxford, OX1 3QU, UK

*Corresponding author: jane.mellor@bioch.ox.ac.uk

Key words: transient nascent transcriptome mapping; 4-thiouridine; HEK293; HEP3B; IFNγ-inducible enhancers; alternative polyadenylation; transcription start site mapping; SNU-Seq; size-selected SNU-Seq.

Running Title: Single nucleotide resolution 4sU (SNU) sequencing

Abstract

Genomes are pervasively transcribed leading to stable and unstable transcripts that define functional regions of genomes and contribute to cellular phenotypes. Defining comprehensive nascent transcriptomes is pivotal to understand gene regulation, disease processes, and the impact of extracellular signals on cells. However, currently employed methods are laborious, technically challenging and costly. We developed single-nucleotide resolution 4sU-sequencing (SNU-Seq), involving pulse labelling, biotinylation and direct isolation of nascent transcripts. Artificial poly-(A)-tailing of the 3' most nucleotide of nascent transcripts ensures oligo-d(T) primer-based library preparation and sequencing using commercial 3' RNA-Seq kits. We show that SNU-Seq is a cost-effective new method generating even read profiles across transcription units. We used SNU-Seq to identify transcription elongation parameters, to map usage of polyadenylation (PAS) sites and novel enhancers. Remarkably, 4sU labelled nascent RNA accumulates short ~100nt transcripts that map to the 5' end of genes. We show that isolation of these short nascent RNA and sequencing the 5' and 3' ends using size-selected SNU-Seq (ssSNU-Seq) provides highly sensitive annotations of mapped and novel TSSs, promoter-proximal pause/termination sites. Thus, SNU-seq and ssSNU-seq combined yield comprehensive transcriptomics data at low cost with high spatial and temporal resolution.

Highlights

- SNU-Seq maps nascent transcripts at base-pair resolution, with high sensitivity and low cost
- SNU-Seq detects comprehensive polyadenylation sites.
- SNU-Seq maps the promoter proximal pause 60-80 nt from the TSS.
- Size-selected SNU-Seq yields highly sensitive and novel TSS annotations

Introduction

To understand the process of transcription by RNA polymerase II (Pol2), methodology is required to capture nascent transcripts. Monitoring nascent transcripts has improved the mapping of transcription start sites (TSS) (Tome et al. 2018) and identified novel concepts such as promoter-proximal pausing (Core and Adelman 2019) and enhancer RNAs (eRNA) (Hou and Kraus 2021). Recent advances in nascent RNA technologies demonstrated how these processes are integral to gene regulation in eukaryotes (Laitem et al. 2015; Nojima et al. 2018; Core and Adelman 2019; Larke et al. 2021) in response to a range of different conditions and during development. These techniques (see Supplemental Table 1) were successfully employed to monitor nascent transcription quantitatively e.g. TT-Seq (Schwalb et al. 2016), PRO-Seq (Kwak et al. 2013), GRO-Seq (Core and Lis 2008), NET-Seq (Churchman and Weissman 2011; Nojima et al. 2015; Larke et al. 2021), to assess elongation rates (TT-Seq), to identify and map promoter proximal pause sites (NET-Seq, PRO-Seq), and TSSs e.g. START-Seq, ScaRNA-Seq (Nechaev et al. 2010; Scruggs et al. 2015; Lavender et al. 2016; Larke et al. 2021) and CoPRO-Seq (Tome et al. 2018). Whilst these methods have been extensively used and yielded the data to formulate and provided the basis to understand these new concepts, they are costly and each have limitations. These drawbacks include a lack of resolution e.g. TT-Seq (Gregersen et al. 2020) and GRO-Seq (Core and Lis 2008) and the reliance on laborious and challenging technical steps including the isolation of intact nuclei prior to labelling (GRO-seq and PRO-seq) and non-standard library preparations e.g. PRO-Seq (Mahat et al. 2016), TT-Seq (Schwalb et al. 2016), NET-Seq (Churchman and Weissman 2012). The current approaches are relatively costly as they generally also suffer from low read numbers forcing very deep sequencing of samples in order to generate more comprehensive data e.g. PRO-Seq (Woo et al. 2018), NET-Seq (Schlackow et al. 2017; Nojima et al. 2018). This emphasises the acute need for simpler techniques that provide nucleotide resolution with the high sensitivity and versatility provided by a metabolic label (Cleary et al. 2005; Fuchs et al. 2014; Sidaway-Lee et al. 2014;

Stoeckius et al. 2014; Schwalb et al. 2016; Akbalik et al. 2017) whilst relying on cost effective standard RNA-Seq library preparation and sequencing kits. Such improvements will be pivotal to accelerate further insight into the complexity and integration of transcriptional events with post-transcriptional downstream fates (Babour et al. 2016; Fischl et al. 2017; Brown et al. 2018; Brown 2019; Fischl et al. 2019).

To combine the benefits and address some of the shortfalls of the current methods (Supplemental Table 1), we developed single-nucleotide resolution 4sU-sequencing, or SNU-Seq, and size-selected ssSNU-Seq. Similar to TT-Seq, SNU-Seq relies on 4sU-pulse labelling of cells, which is very efficient, but omits sonication and includes addition of an artificial poly-(A)-tail at the 3' nucleotide on the nascent transcript. This unique step allows the use of standard oligo-d(T) primer-based RNA-Seq library kits and commercial 3' RNA-Sequencing kits whilst retaining single nucleotide resolution.

Supplemental Table 1: Comparison SNU-Seq and ssSNU-Seq with Existing Techniques to Assess Nascent Transcription

Technique	**GRO-Seq**	**PRO-Seq**	**NET-Seq**	**TT-Seq**	**SNU-Seq**	**SNU-Seq (with size-selection)**
In Vivo ?	No	No	Yes	Yes	Yes	Yes
Nascent RNA Selection	BrdU-Antibody	Biotin-NTPs / Streptavidin	Pol II –Antibody	4sU-Specific Biotin / Streptavidin	4sU-Specific Biotin / Streptavidin	4sU-Specific Biotin / Streptavidin
Fragmentation	Alkaline Hydrolysis	Alkaline Hydrolysis	MNase Digestion	Sonication (chemical fragmentation for TTchem-Seq)	None	Size Selection by RNA Gel
Resolution	Lower	Single bp	Single bp	Lower	Single bp	Single bp
Usage	Nascent Transcription	Enhancer Detection; Single basepair resolution; Detects promoter-proximal pause	Possibility of sequencing nascent RNA specific to Pol II modifications; Detects promoter-proximal pause	Determination of Elongation Rates; Sensitive; Straightforward protocol	Accurate detection of poly(A)-site, novel enhancers and promoter proximal pause; High sensitivity; straightforward protocol; Single basepair resolution	Detects promoter-proximal pause; Identification of TSSs; Single basepair resolution

This approach generates an even nascent transcriptional profile at single nucleotide resolution across transcription units with distinct reproducible spikes where reads accumulate at specific genomic locations. We confirm that the majority of these read spikes can be mapped to the 5' and 3' end of transcription units and show that this approach can be used to reliably identify promoter proximal pause sites, 60-80 nt downstream of the transcription start site (TSS), and poly A sites (PAS) usage including alternative polyA sites (Neve et al. 2017) in the nascent transcripts, respectively. A feature of SNU-Seq is the accumulation of short 4sU-labelled nascent RNAs (~100 nt) that predominantly map to the 5' end of transcription units. We exploited this feature to develop size-selected SNU-Seq (ssSNU-Seq) and show that the additional size

selection step prior to sequencing provides a highly accurate tool to map both TSSs and promoter-proximal pause/early transcription termination sites, at single base-pair resolution. To demonstrate the resourcefulness of this novel approach, we successfully employed SNU-Seq to study the dynamic nature of the IFNγ response in HEP3B cells and identify hundreds of novel IFNγ-induced enhancers and extensive alternative polyadenylation site usage in response to IFNγ.

Results

SNU-Seq reports nascent transcription at near-single nucleotide resolution (Fig. 1)

Similar to 4sU-Seq methods, SNU-Seq is based on 4sU-labelling of cells for 5 to 10 minutes followed by purification of nascent RNA using biotin-labelling and streptavidin pull-down (Fig. 1A).

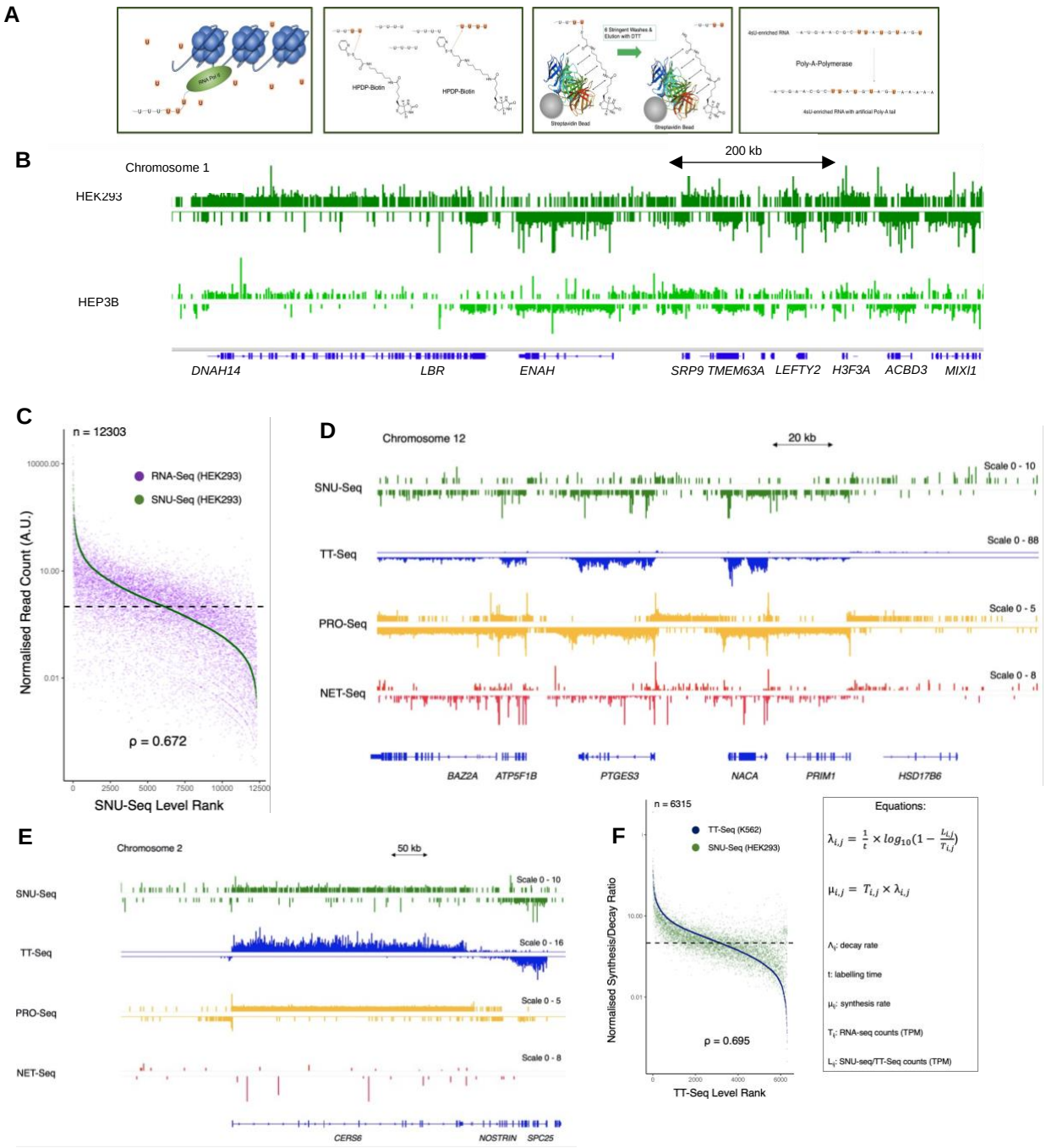
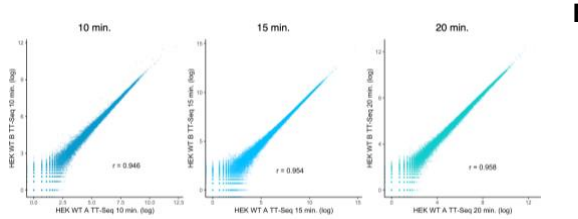


Figure 1. Comparison of SNU-Seq using a 10 min 4sU pulse label in different cell types and with other methods for assessing nascent transcription: **A** The SNU-Seq workflow includes thio-labelling in vivo followed by RNA purification without sonication. The RNA is then biotinylated with HPDP-biotin, which specifically forms a covalent bond to thio-labelled RNA which is then purified using strepdavidin columns. An artificial poly-A tail is added before performing strand-specific library preparation using oligo-dT primers. **B** IGV screenshot displaying a comparison of SNU-Seq in HEK293 cells and HEP3B cells (this study) on a region of chromosome 1. **C**. Correlation between SNU-Seq (this study) and RNA-Seq in HEK293 cells (Banks et al. 2014). The Spearman correlation coefficient (rho) is indicated. **D,E** IGV screenshots displaying a comparison of SNU-Seq (HEK293, this study), NET-Seq (Nojima et al. 2015), HeLa), TT-Seq (this study, HEK293), and PRO-Seq (Woo et al. 2018), HEK293). **F** Ranked correlation plot of the synthesis-to-decay ratios between SNU-Seq and TT-Seq. The equations used are indicated in the box on the right. SNU-Seq data are in HEK293 cells (from this study), and, for comparison, to see if SNU-Seq can be used for this mathematical approach, the original TT-Seq dataset in K562 cells was used (Schwalb et al. 2016). The Spearman correlation coefficient (rho) is indicated.

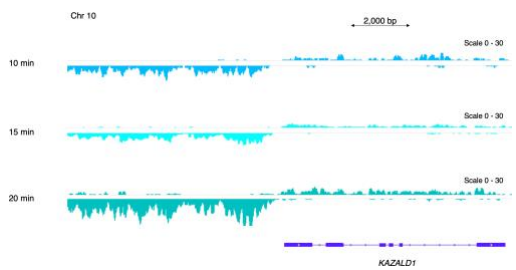
Supplemental Figure 1

A

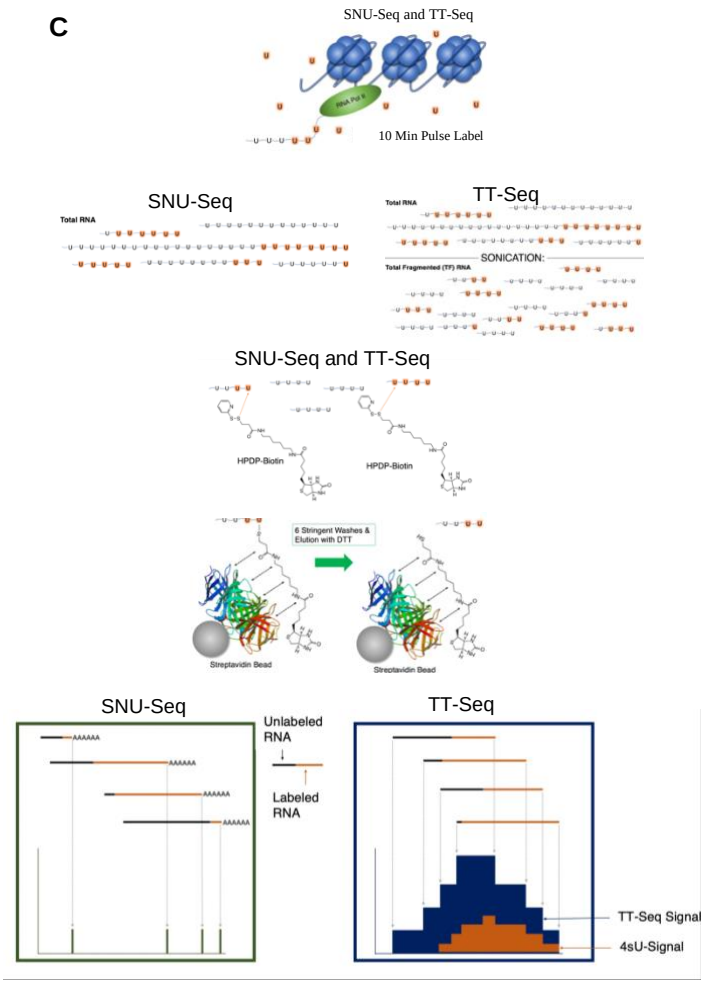


Repeat A	84.3 %	82.5 %	85.0 %
Repeat B	78.0 %	81.4 %	86.6 %

B



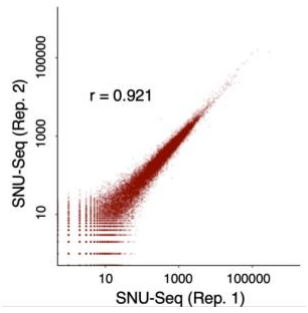
C



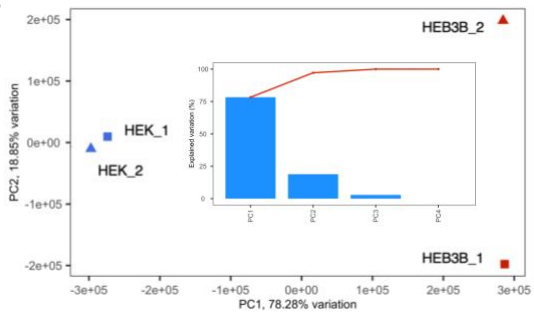
D

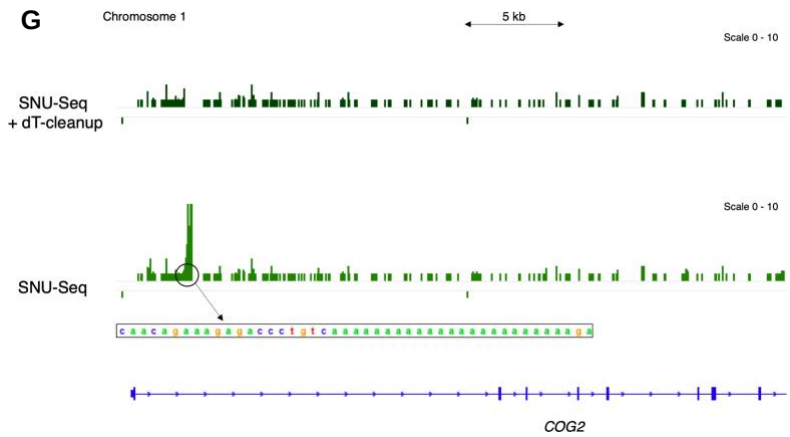
Alignment Scores	Repeat 1	Repeat 2
Aligned Exactly Once (Percentage)	64.86	65.88
Overall Alignment (Percentage)	77.8	78.63

E



F





Supplemental Figure 1. Establishment of TT-Seq and SNU-Seq in HEK293 cells. A Scatter plot of log-transformed read counts for two repeats of 10, 15 and 20 -minute labelled TT-Seq experiment in HEK293 cells and alignment percentages of TT-Seq experiments. **B** TT-Seq IGV Genome Browser screenshot at the KAZALD1 locus from HEK293 cells (10, 15 and 20 min labelling times). **C** Comparison between SNU-Seq (left, green) and TT-Seq (right, blue) with respect to how their signals are generated. **D** Scatter plot of SNU-Seq repeats read counts (log scale) in HEK293 cells. The Pearson correlation coefficient is shown (r = 0.921). **E** Alignment scores for SNU-Seq repeats in HEK293 cells. **F** Principal component analysis of SNU-Seq in HEK293 and HEP3B cells with two repeats each. The first (x-axis) and second (y-axis) principal components are plotted against each other in a biplot. **Insert:** Scree plot of principal component analysis of SNU-Seq in HEK293 and HEP3B cells with two repeats each. **F** IGV screenshots of SNU-Seq with and without dT-cleanup are shown around the *COG2* locus. The DNA sequence of the forward strand surrounding the peak in the SNU-Seq signal is indicated.

This gives sufficient time for cells to process the metabolic label into a triphosphate for incorporation into the nascent transcripts, but is short enough to capture primarily nascent, as opposed to fully processed transcripts. Optimal labelling times were confirmed by performing TT-Seq on HEK293 cells and visualizing loci in IGV and found to be similar to those required for HeLa cells (Fig. S1A,B). To generate single-nucleotide resolution, an artificial poly-A-tail is added to the labelled nascent transcripts after enrichment for labelled RNA but before library preparation using *E. coli* poly-(A)-polymerase (Fig. 1A, S1C). The artificial poly-(A)-tail allows for library preparation with oligo-dT primers. This means that positional information about the last incorporated nucleotide during the pulse labelling window is retained but also leads to priming from rare internal Poly(A) tracks on the nascent transcripts, for example at *COG2* (Fig. S1G), which were removed computationally from the datasets. In contrast to TT-Seq (Schwalb et al. 2016), SNU-Seq does not incorporate a sonication step after total RNA purification (Fig. S1C), does not generate a 5' bias (Fig. S1C), approximately 65% of reads align (Fig. S1D) and biological repeats correlate well (Fig. S1E) with no major batch effects observed in HEK293 or HEP3B cells with a 10 min pulse label (Fig. S1F). The IGV snapshot reveals uniform coverage over the transcribed regions of the coding and non-coding genome and similar read coverage in both cell types (Fig. 1B). For HEK293 cells, the SNU-Seq signal correlates well (Spearman correlation 0.672) with transcript levels at 12,303 genes (Banks et al. 2014) (Fig. 1C). As SNU-

seq is a simple technique, comparable to 3' end RNA-Seq (Hoque et al. 2013; Zheng et al. 2016), it offers real advantages compared to other methods for assessing nascent transcription.

Comparison between SNU-Seq and other techniques for assessing nascent transcription.

We compared SNU-Seq with other techniques for assessing nascent transcription: GRO-seq (HeLa cells, (Laitem et al. 2015)), PRO-Seq (HEK293, (Woo et al. 2018)), NET-Seq (HeLa cells,(Nojima et al. 2015)) and TT-Seq (K562 cells,(Schwalb et al. 2016)) (Supplemental Table 1; Fig. 1D,E). The PRO-seq data (HEK293,(Woo et al. 2018)) is of very high quality but requires a very high read depth to generate comprehensive data and the cost involved, plus the need to isolate nuclei, rules this out as a routine method. SNU-Seq is most similar to TT-Seq as both involve a short pulse label with 4sU (Fig. S1C). SNU-Seq has a more even signal over genes and eliminates the 5' bias observed with TT-Seq (Fig. 1D,E, S1C). At some, but not all loci, TT-Seq shows a higher signal over exons than introns suggesting that some of the signal is derived from transcripts that have already been processed into mRNA (Fig. 1D,E), thus some of the nascent signal in introns may be lost from this dataset. To verify that SNU-Seq can accurately reflect transcription elongation profiles, genome-wide, we compared elongation parameters between that established for the TT-Seq protocol (K562 cells, (Schwalb et al. 2016)) and SNU-Seq (HEK293) by determining synthesis rates, decay rates, and the synthesis-to-decay ratio (Fig. 1F). The transcript synthesis-to-decay ratio indicated the two methods show a good correlation (Spearman correlation of 0.695) between their elongation parameters. Thus, together with an RNA-Seq dataset, SNU-Seq can be used to determine elongation parameters, and these can be used with mathematical simulations to relate changes in elongation rates to transcript fate (Brown et al. 2018; Uzun et al. 2021).

Another general feature of SNU-Seq is the high sensitivity for nascent transcription compared to GRO-Seq and NET-Seq at some loci (Fig. 1D) but not others (Fig. 1F). For SNU-Seq, PRO-Seq and GRO-Seq the signal is derived predominantly from the act of nucleotide incorporation, while NET-seq captures nascent transcripts associated with elongating, paused and backtracked polymerase and generates a distinct peak at the 5' region of transcription units. Consistently, SNU-Seq generally shows a very low variance of the signal height and provides a very flat profile, with the density of reads reflecting the amount of label incorporated during the labelling window (Fig. 1F). The profile appears to be more variable for GRO-seq and TT-seq, which are not single nucleotide resolution techniques. SNU-Seq is most similar to the very deep data set for the PRO-Seq (Woo et al. 2018).

SNU-seq detects hundreds of novel eRNAs at putative enhancers in IFNγ treated HEP3B cells (Fig. 2)

There are many unstable transcripts in the mammalian genome. To assess the ability of SNU-Seq to detect these, we used our IFNγ treated HEP3B model to detect nascent eRNAs, which at an enhancer positively correlate with target promoter transcription, and thus define enhancer activity (Andersson et al. 2014; Catarino and Stark 2018). First, we focused on enhancer annotations from the FANTOM5 atlas of promoters, enhancers, lncRNA and miRNAs (Kawaji et al. 2017) (http://fantom.gsc.riken.jp/5/datafiles/) using SNU-Seq and DESeq2 to detect differential transcription within 5kb of an annotated enhancer at each of the three time points. We detected a total of 112 changes (Fig. 2A). Examples of IFNγ-inducible transcription (30 mins) at annotated FANTOM5 enhancers include e49007 (within IRF1-AS1) and e15662 (Fig. 2B,C). Both show the characteristic features of divergent nascent transcription, inducible peaks of H3K27 acetylation (H3K27ac), ATAC-seq chromatin accessibility but no H3K4me3, which distinguishes them as enhancers not promoters (see *IRF1-AS1* promoter; Fig. 2B). We noted that close to e15662 there are two unannotated regions with divergent SNU-seq transcription, characteristic of enhancers (Fig. 2B).

To examine other unannotated putative enhancers, we counted the differentially expressed SNU-seq reads with 5kb of ATAC-seq peaks (Fig. 2C; S2A,B). ATAC-Seq peaks are present at both

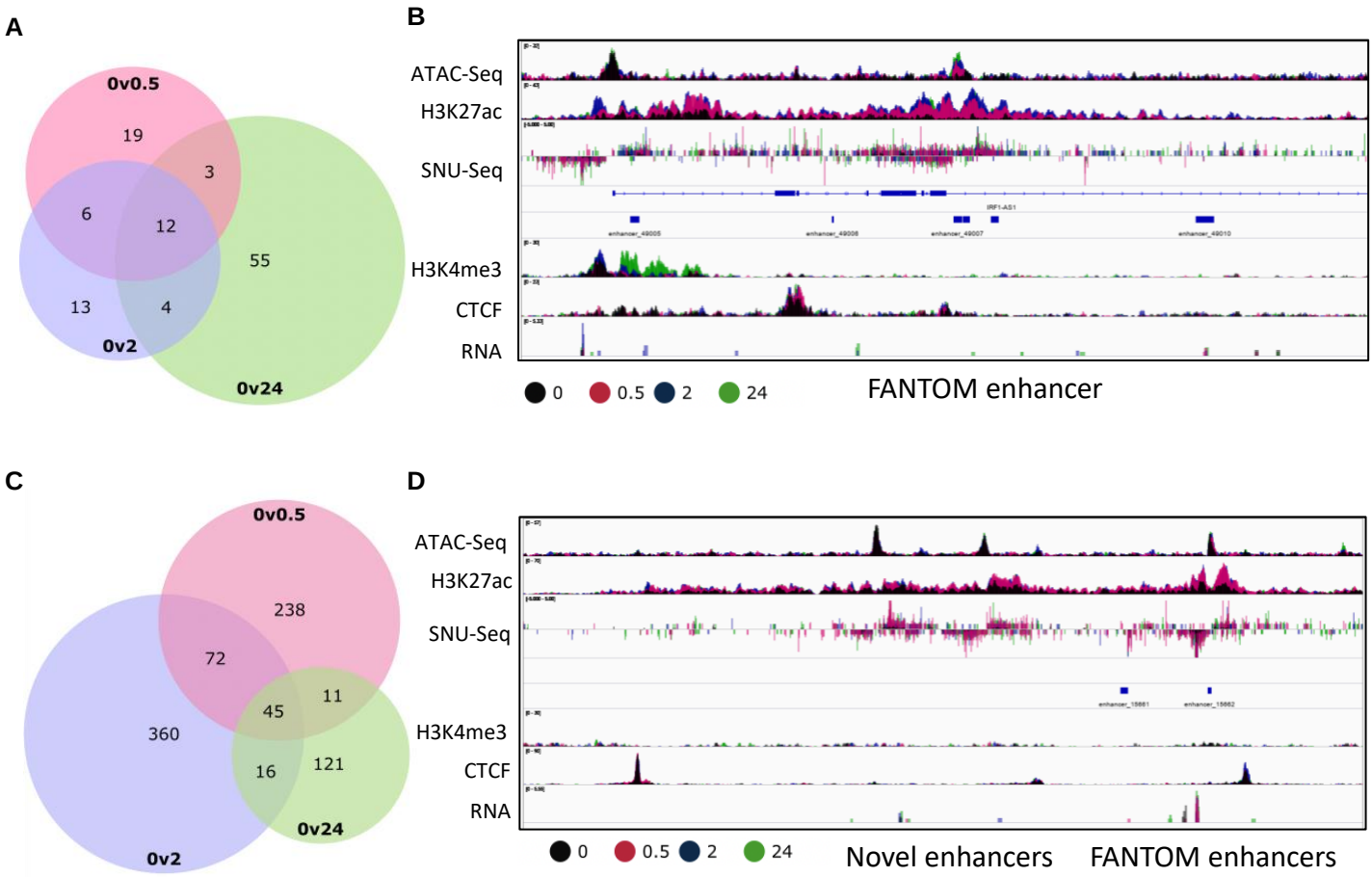
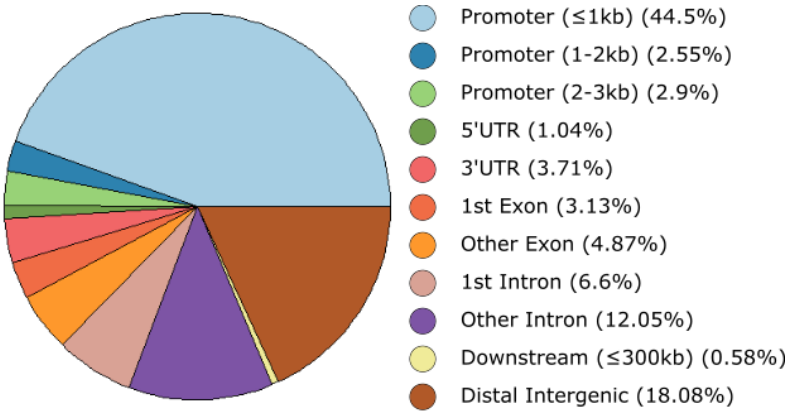


Figure 2 SNU-seq identifies IFNγ-induced changes in the enhancer landscape. A,C Venn diagrams showing IFNγ-induced transcriptional changes within a 5kb window surrounding FANTOM5 enhancers (**A**) or ATAC-seq peaks (**C**) for each condition; untreated vs 0.5 hrs (pink), 2 hrs (blue) or 24 hrs (green) of IFNγ treatment. **B,D** IGV images of the regions surround the *IRF1-AS1* locus (**B**) or a 12kb stretch of chr3 with no annotated loci (**D**) showing strand-specific SNU-Seq, RNA-Seq, DNA accessibility (ATAC-Seq) and ChIP-Seq profiles of CTCF, H3K27ac and H3K4me3 data for each treatment sample (n=2-3), with the untreated sample in black, and 0.5, 2 and 24 hrs of IFNγ treatment in pink, blue and green, respectively. All tracks are spike-in normalised except ATAC-Seq which is normalised based on sequencing depth. Annotated FANTOM5 enhancers are represented by the blue boxes.

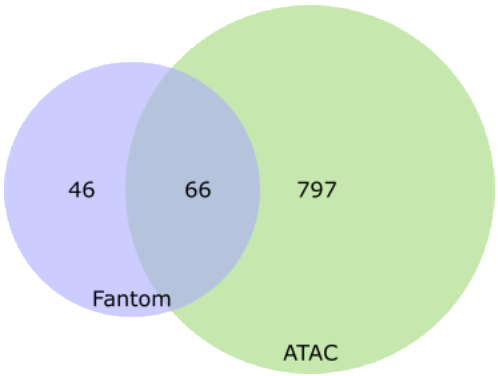
A



B

	# of regions	
	↑	↓
0v0.5	200	166
0v2	262	231
0v24	173	20

C



Supplemental Figure 2. A Summary of the numbers of differential nascent transcription events within a 5kb window of an ATAC-Seq peak per condition; untreated vs 0.5 hrs, 2 hrs or 24 hrs of IFNγ treatment. **B** Genome-wide distribution of regions of differential transcription at ATAC peaks determined using CHIPseeker. **C** Number of ATAC-Seq peaks compared to annotated FANTOM5 enhancers with differential transcripts within 5kb.

promoters and enhancers (see Fig. 2B). Analysing distributions of the 863 ATAC-Seq accessible regions with differential SNU-Seq reads, revealed that approximately 50% were present within 3kb of a promoter (Fig. S2A), suggesting that half of the differentially transcribed peaks were changes in gene expression or transcription of intragenic enhancers, while the remaining peaks may represent the differential activity of gene distal enhancers, of which over 300 are not annotated in the FANTOM5 database (Fig. S2B,C). Thus SNU-seq combined with ATAC-seq identifies putative novel enhancers specific to IFNγ treated HEP3B cells, not established in public enhancer databases. It has been argued that eRNA production provides a better indication of enhancer activity than H3K27ac levels (Fishilevich et al. 2017; Tyssowski et al. 2018).

Mapping polyadenylation sites (PAS) and alternative PAS in nascent transcripts using SNU-seq (Fig. 3)

Spikes are evident in the SNU-Seq profiles, particularly at the 5' and 3' regions of transcription units. We focused first on the 3' end of transcription units. SNU-Seq detects poly(A) tails on nascent transcripts and reports these as a spike, for example at *SSR3* which has well defined proximal and distal PAS (Fischl et al. 2019) (Fig. 3A; Fig. S3A). After normalising the SNU-Seq read counts for each gene to make every gene (N=9336) contribute equally to the average profile, a clear spike is evident in the metagenomic analysis around the poly(A) site (Fig. 3B).

The ability to detect usage of annotated PAS is a feature unique to SNU-Seq, as no spikes are evident in the profiles using TT-Seq, PRO-Seq or NET-Seq (Fig. 3A;S3A). To validate that the SNU-Seq spikes correspond to authentic PAS signals (Fischl et al. 2019; Wang and Tian 2020), we identified a set of genes which show differential APA between nuclear and cytoplasmic fractions using 3' end RNA-Seq in HEK293 cells. The relative levels of transcript isoforms within the nuclear fraction will be more representative of the nascent RNA population than the cytoplasmic fraction. This is because post-transcriptional regulation at the level of nuclear export and RNA degradation will have less of an impact on the nuclear fraction. At these genes showing APA, the level of transcript isoforms using particular PAS detected using SNU-Seq is more similar to the nuclear fraction, demonstrating that SNU-Seq is detecting the level at which particular PAS are being used by nascent transcripts. These events are illustrated for *SCO*1 (Fig. 3C), *COX20* and *HNRNPU* (Fig. S3B).

We were keen to validate the ability of SNU-Seq to detect APA using a new data set. We examined SNU-Seq profiles in HEP3B cells treated with IFNγ for 30 min, 2 h and 24 h after removal of the internal poly A tracks. An example of APA in this data set is seen at *STAT1* which encodes a transcription factor necessary for the IFNγ response (Fig. 3D, S3C). *STAT1* is an early gene, transcriptionally induced after 30 mins treatment with IFNγ and maintained for 24 h. At each time point we observe an increasingly complex pattern of spikes at the 3' end of the STAT1, indicative of APA, and the corresponding appearance of transcripts with similar 3' ends. APA is evident at other early induced genes over time, including *EDN1* (Fig. S3C).

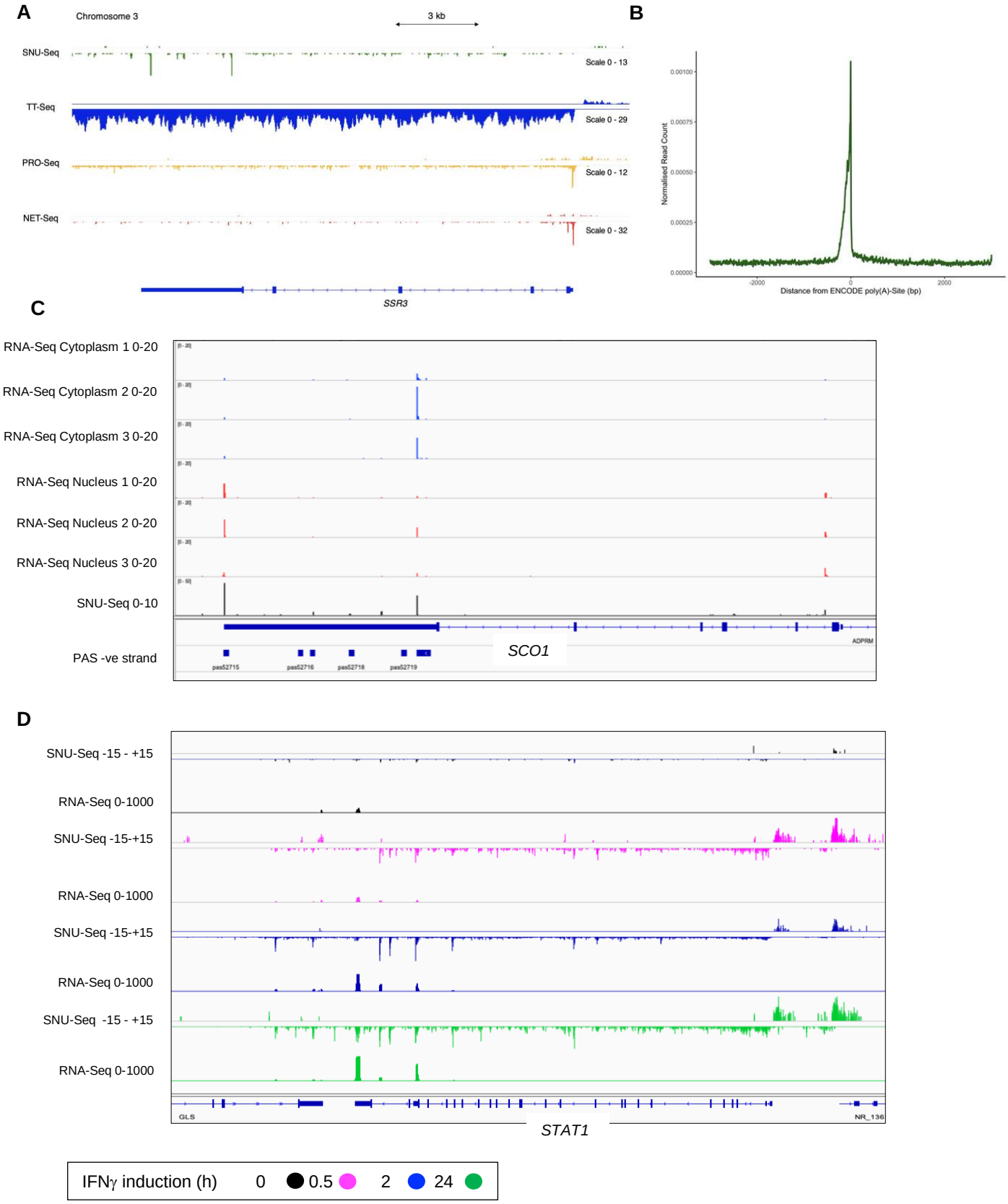
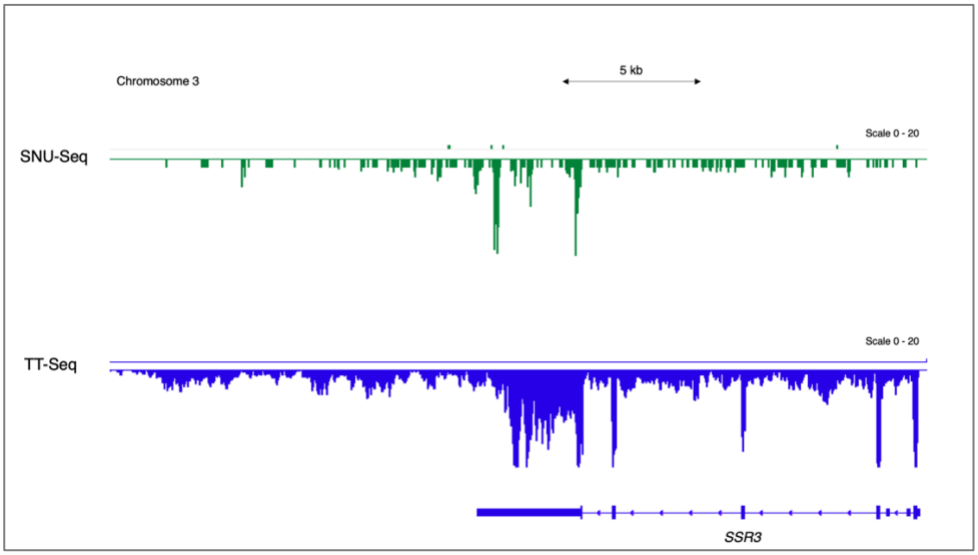
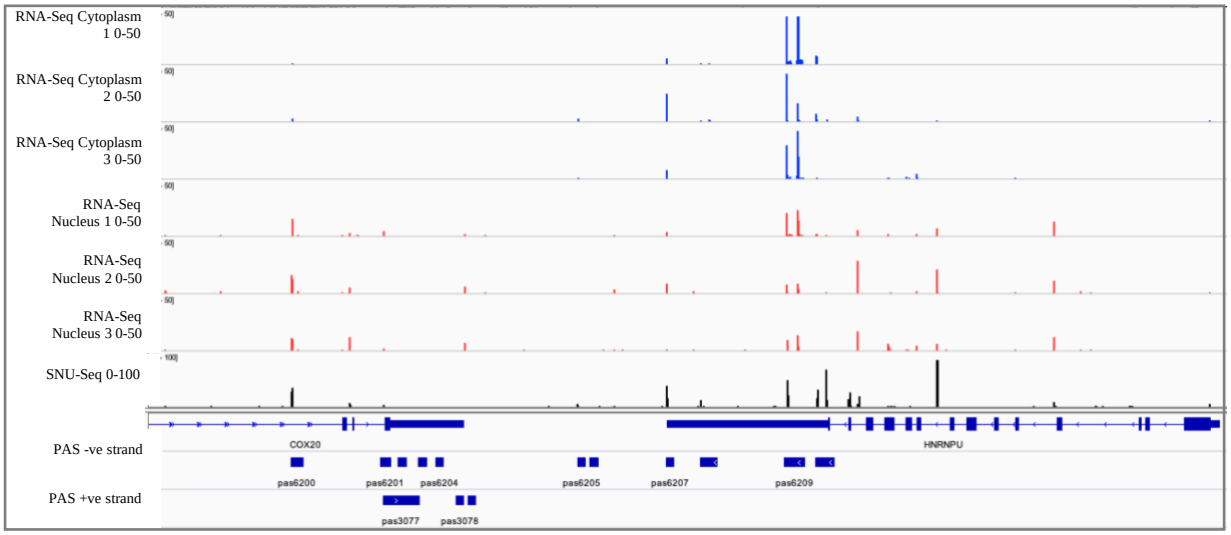


Figure 3 Detection of polyadenylation sites and alternative polyadenylation sites using SNU-Seq. A The TT-Seq (HEK293, navy), SNU-Seq (HEK293, green), PRO-Seq (yellow, HEK293, (Woo et al. 2018), and NET-Seq (red, HeLa (Nojima et al. 2015)) profiles at the *SSR3* locus are shown. **B** SNU-Seq metagene around ENCODE poly(A)-sites (n = 9336). If multiple poly(A)-sites are known, the most proximal option was chosen. **C** Signal at *SCO1* PASs for SNU-Seq (black), and nuclear (red) and cytoplasmic (blue) 3' end RNA-Seq in HEK293 cells. Windows containing mapped PAS are indicated by the blue boxes. **D** IFNγ induction of HEP3B cells results in relative and differential APA over induction time at *STAT1* evident in SNU-Seq spikes and total RNA 3' end reads. Scales are indicated for comparison.

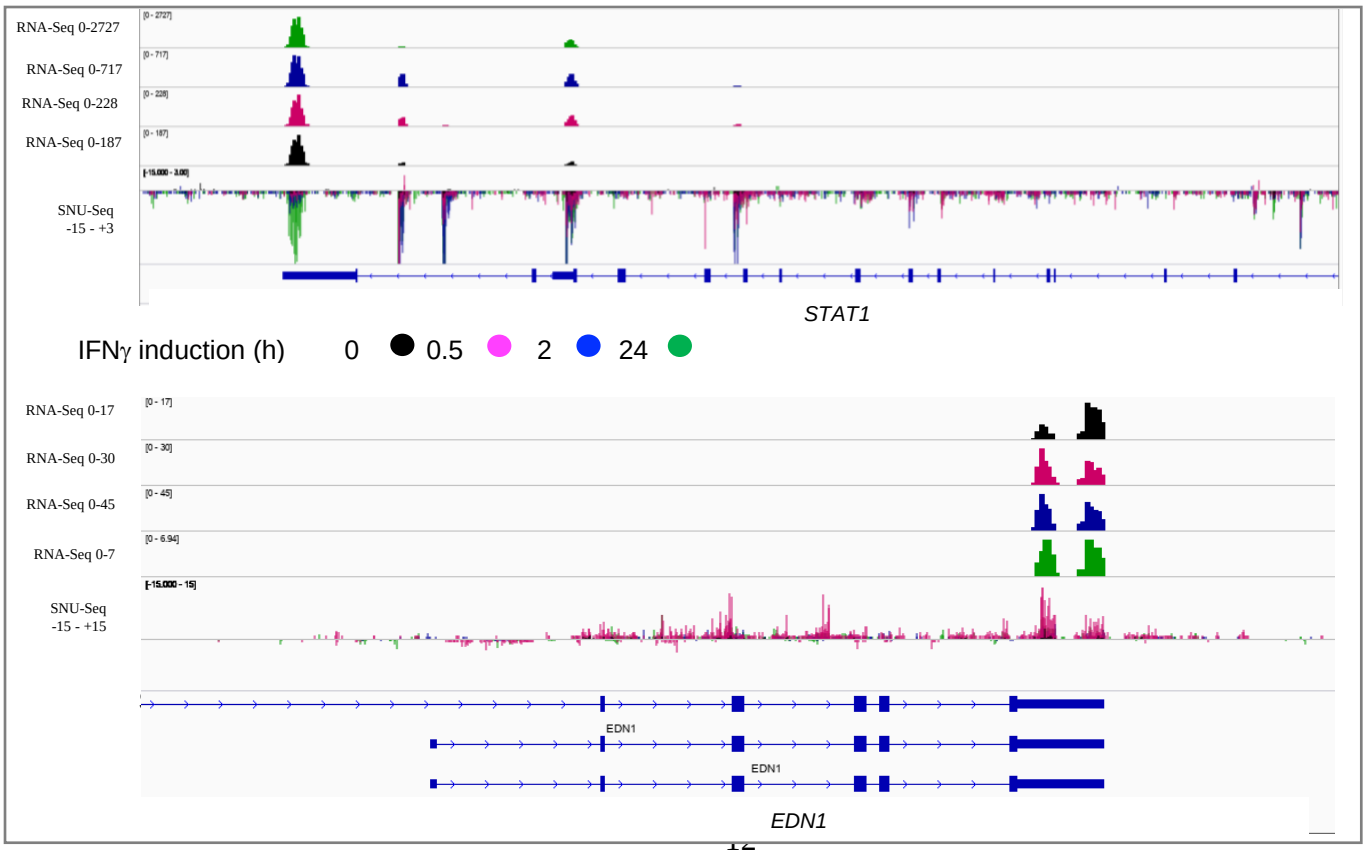
A


Chromosome 3
5 kb
SNU-Seq
Scale 0 - 20
TT-Seq
Scale 0 - 20
SSR3

B


RNA-Seq Cytoplasm 1 0-50
RNA-Seq Cytoplasm 2 0-50
RNA-Seq Cytoplasm 3 0-50
RNA-Seq Nucleus 1 0-50
RNA-Seq Nucleus 2 0-50
RNA-Seq Nucleus 3 0-50
SNU-Seq 0-100
PAS -ve strand
PAS +ve strand
COX20
HNRNPU
pas6200
pas6201
pas6204
pas6205
pas6207
pas6209
pas3077
pas3078

C


RNA-Seq 0-2727
RNA-Seq 0-717
RNA-Seq 0-228
RNA-Seq 0-187
SNU-Seq -15 - +3
STAT1
IFNγ induction (h)
0
0.5
2
24
RNA-Seq 0-17
RNA-Seq 0-30
RNA-Seq 0-45
RNA-Seq 0-7
SNU-Seq -15 - +15
EDN1
EDN1
EDN1

Supplemental Figure 3 Detection of polyadenylation sites and alternative polyadenylation sites using SNU-Seq. A Detailed view of the TT-Seq (HEK293, navy) and SNU-Seq (HEK293, green) profiles at the *SSR3* locus. Note the reads from exonic regions in the TT-Seq data. Scales are indicated for comparison. **B** Signals at *COX20* and *HNRNPU* PASs for SNU-Seq (black), and nuclear (red) and cytoplasmic (blue) 3' end RNA-Seq in HEK293 cells. Windows containing mapped PAS are indicated by the blue boxes. **C** IFNγ induction of HEP3B cells results in relative and differential APA evident in nascent transcripts (SNU-Seq) and steady-state transcripts (3' end RNA-Seq) over induction time at *STAT1* (different view) and *EDN1* evident in SNU-Seq spikes and total RNA 3' end reads. Scales are indicated for comparison.

SNU-Seq captures the 5' promoter proximal pause/site of early termination (Fig. 4)

At some genes, a spike close to the transcription start site (TSS) is evident in the IGV plots, for example at *CHEK2* (Fig. 4A). To produce an average profile for a metagenomic analysis at the 5' region, we made the SNU-Seq read counts for each of the 9336 genes contribute equally, by normalising. This resulted in a complex profile around the TSS (Fig. 4B). The metagene show clear signal increase after the TSS followed by a small peak. However, the average SNU-Seq coverage over the gene body is much higher than in the promoter-proximal region suggesting a technical issue with SNU-Seq at the extreme 5' region of some genes (Fig. 4B). We plotted separate metagenes for genes with a high 5' end signal (2548) and those with a low 5' end signal (6788) nt TSS to +150 nt) (Fig. 4C). This revealed two distinct profiles. Interestingly, the average SNU-Seq signal within the gene body is comparable between the two sets of genes. For the genes with the high end signal, the profile is consistent with the detection of the promoter proximal pause/site of early termination, such as at *CHEK2* (Fig. 4A). This suggests a bias in generating reads at the 5' ends of genes for the low 5' end signal class of genes, which may arise due to use of random primers at the 5' end during library preparation, meaning that some very short nascent RNAs may not be fully reflected in the library.

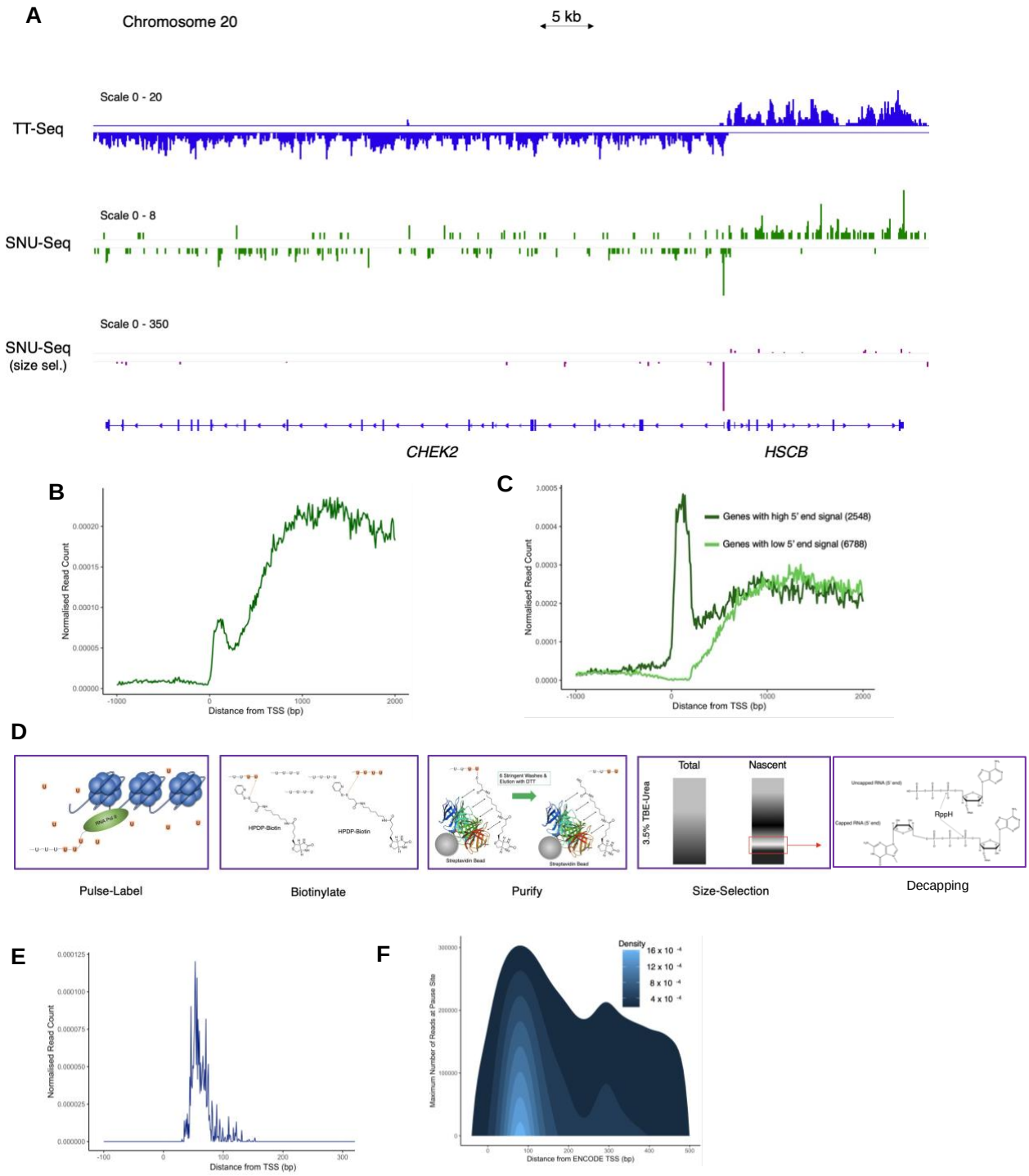
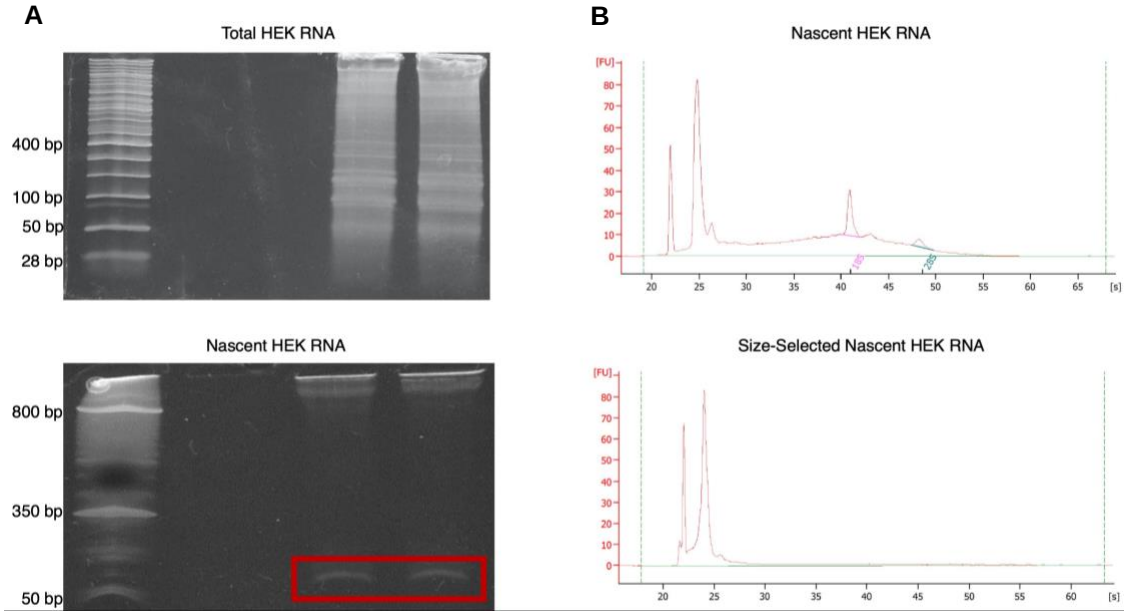
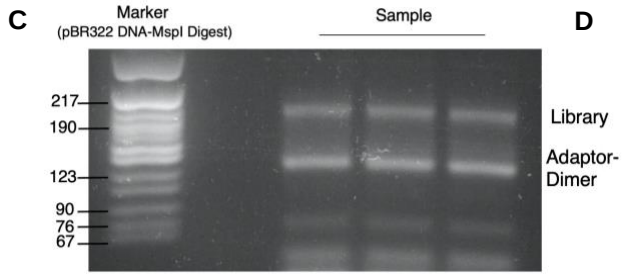


Figure 4 SNU-Seq and ssSNU-Seq capture the promoter-proximal pause/site of early termination. A IGV screenshot displaying TT-Seq (HEK293, this study), SNU-Seq (HEK293, this study), and size- selected SNU-Seq (HEK293, this study) profile around the *CHEK2* locus on chromosome 20. **B** A metagene plot of SNU-Seq reads around transcription start sites (TSS annotations from Brown et al. (2018)). **C** The genes were split up based on whether there was a considerable number of reads in the first 150 bp (> 5 % of reads in the window) or not. **D** Workflow for SNU-Seq with size-selection including thio-labelling, biotinylation, and streptavidin purification followed by a gel-based size selection step. After purifying RNAs of a size range between 50 and 100 nucleotides and decapping, adapters were added immediately for library preparation, retaining single nucleotide resolution at both the 5' and the 3' ends. **E**. Normalised metagene of size-selected SNU-Seq reads around human TSSs. TSS annotations were derived from Brown et al. (2019) (Brown 2019) (n = 9336). **F** 2-dimensional density plot showing the distribution of the maximum read number of each gene against its location relative to the gene's ENCODE transcription start site. Density refers to the number of occurrences per pixel (n = 19874).

SNU-Seq Combined with Size-Selection Allows for Characterization of the Promoter-Proximal Region of Genes with High Precision (Fig. 4)

While establishing controls for the SNU-Seq protocol, we observed that the nascent thio-labelled RNA reproducibility yields a clear band at approximately 50-100nt and a smear from 600nt on gels (Fig. 4D, S4A,B). The short RNA was isolated and subject to linker ligation at both the 3' end (75nt) and 5' end (54nt) after decapping with RppH. The coverage showed more the 86% of the library aligned to the human genome (Fig. S4C,D) predominantly at the 5' region of transcription units, and sequencing allows the first and last nucleotides of the gel-purified nascent transcripts to be reported.





D

Sequencing Stats	Number	Percentage
Number of Total Reads	333,002,721	
Reads that align 0 times	44,549,628	13.3 %
Aligned Reads	288,453,093	86.7 %

Supplemental Figure 4. SNU-Seq and ssSNU-Seq capture the promoter-proximal pause. **A** RNA-Gels run with total and nascent HEK293 RNA. **B** RNA pico-chip Bioanalyzer traces of nascent HEK293 RNA before and after gel-based size selection. **C** Size-selected SNU-Seq library preparation. DNA gel of size-selected SNU-Seq library products after adaptor ligation and PCR amplification. The top bands represent the desired library product whereas the second highest bands are a result of adaptor dimerisation. **D** Size-selected SNU-Seq sequencing depth and alignment percentage.

Using NET-Seq profiles to validate the promoter proximal pause/site of early termination at the 3' end of ssSNU-Seq reads (Fig. 4)

We examined the 3' end reads in the size-selected (ss) SNU-Seq library. Reassuringly the library from the size selected nascent RNA contained reads that align precisely at *CHEK2* with the spike obtained from the SNU-Seq library (Fig. 4A). Next, a calibrated metagene and heatmap were created to examine the genome wide-distribution of reads in the ssSNU-Seq library (Fig. 4E,F). Overall, the highest density of the maximum 3'-end read signal is located around 60-80 bp downstream of the TSS, and a minor density peak at around 300 bp downstream of the TSS. We speculate that these two peaks represent the promoter-proximal, and -distal pausing sites, respectively, or alternatively, sites of early transcription termination, and set out to validate this using (i) NET-Seq data which captures all form of Pol2 including the promoter proximal pause (Nojima et al. 2015) and a non-supervised machine learning approach, and (ii) ChIP-Seq data for NELF, a pausing factor, INTS3, a subunit of integrator, linked to the pausing signal by premature termination and MED26, a subunit of mediator involved in transcription initiation (Liu et al. 2014; Stadelmayer et al. 2014; Nojima et al. 2018; Elrod et al. 2019; Takahashi et al. 2020).

First, we performed k-means clustering (k = 6) on the shape of human NET-Seq profiles for 2748 genes in order to identify clusters of genes with different promoter-proximal pausing profiles (Fig. 5A, S5A). This yielded one cluster with a flat metagene (cluster 1), three clusters with a sharp peak directly downstream of the TSS (clusters 2, 3, and 4), one cluster with a slightly broader peak (cluster 6), and one cluster with a broad peak much further into the gene body (cluster 5). We then plotted ssSNU-Seq data for the genes in each cluster and, remarkably, the patterns match those in the NET-Seq data (Fig. 5A). Therefore SNU-Seq in combination with size-selection can be used as a highly sensitive and accurate way of measuring the promoter-proximal pause. By contrast, TT-Seq does not have a high enough resolution to detect the promoter-proximal pause (Fig. 5A).

As a control to show that the signal in our experiments is transcription dependent, we performed calibrated TT-Seq on HEK293 cells with and without 6 hours of alpha-amanitin treatment to inhibit Pol2. Unexpectedly, we found that α-amanitin treatment limits the entire TT-Seq signal to the 5' region of genes with no signal over gene bodies (Fig. S5B). Furthermore, the profiles for the NET-Seq, size-selected SNU-Seq, and alpha-amanitin treated TT-Seq profiles are very similar, indicating that in the presence of α-amanitin, Pol2 can transcribe in the window up to the promoter proximal pause site, or that the pauses

are long lived and exist before the α-amanitin treatment, or that these regions are transcribed by another α-amanitin polymerase when Pol2 is inhibited (Fig. 5A).

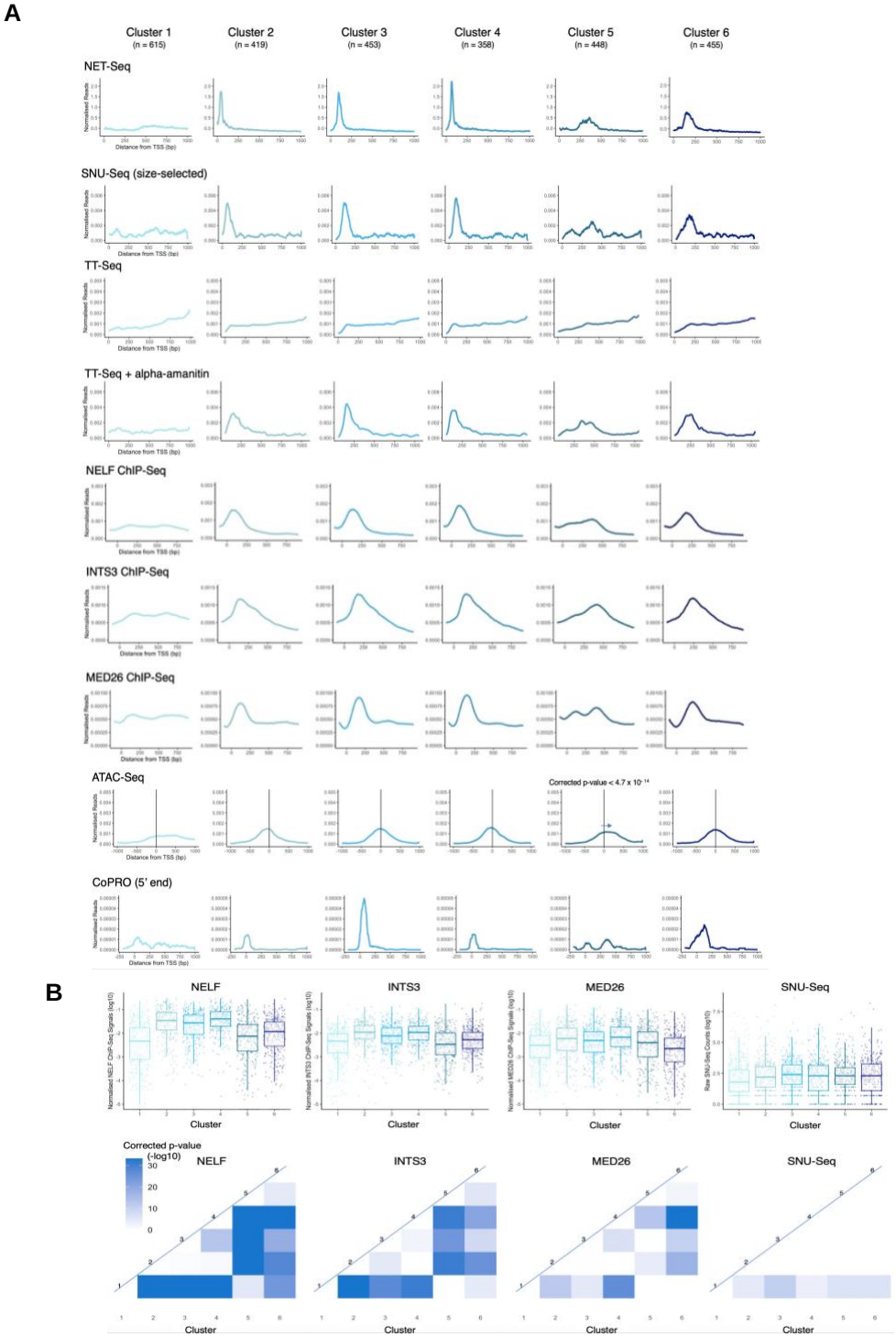
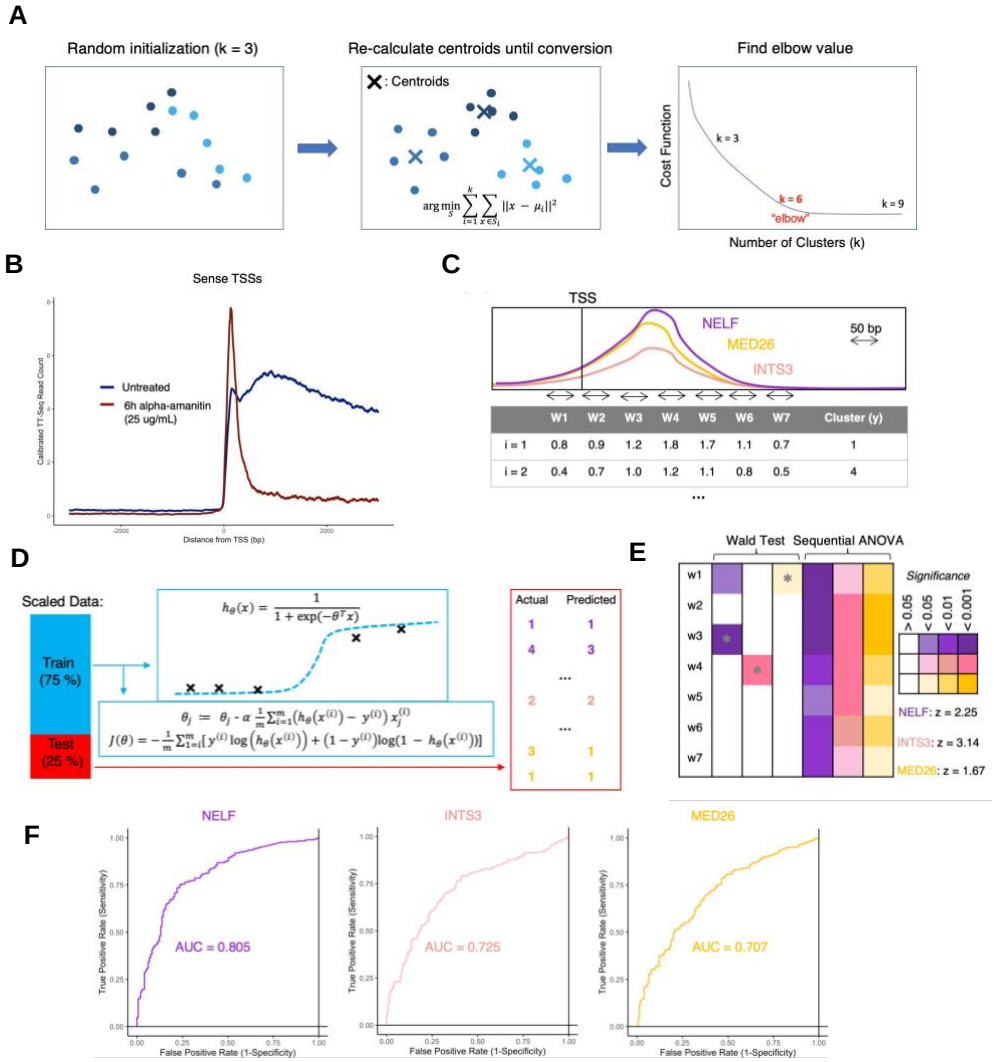


Figure 5. ssSNU-Seq defines the promoter proximal pause/site of early termination. **A** Generation of k-means clusters (k = 1 to k = 6) using human (HeLa) NET-Seq (Nojima et al. 2015). The same cluster indices for each gene were applied to sort (i) size-selected SNU-Seq data (HEK293 cells), (ii) TT-Seq (HEK293 cells), (iii) TT-Seq + a-amanitin (HEK293 cells), (iv) NELF ChIP-Seq, (v) INTS3 ChIP-Seq, (vi) MED26 ChIP-Seq (ChIP-Seq data in HeLa cells; (Liu et al. 2014); (Nojima et al. 2018); (Takahashi et al. 2020), (vii) ATAC-Seq (HEK293 cells, (Weltner et al. 2018), (viii) CoPRO (K562 cells,(Tome et al. 2018)), into the respective clusters. Metagenes are normalised to make every gene contribute equally. **B** Statistical analysis of NELF, integrator, and mediator occupancy within gene clusters. Boxplot diagrams for NELF, INTS3, MED26, and SNU-Seq levels at the 5' end of genes for each cluster (300 bp window for ChIP-Seq and 1000 bp window for SNU-Seq). ChIP-Seq data was normalised to make each gene contribute equally. For SNU-Seq, the raw counts were used. All data was log-transformed. Triangular heatmaps of between-cluster significance for NELF, INTS3, MED26, and SNU-Seq levels. P-values were calculated using the Wilcox-rank sum test and corrected using the Bonferroni method.



Supplemental Figure 5: ssSNU-Seq defines the promoter proximal pause/site of early termination. A The k-means clustering algorithm, in a simplified version, is shown. The data to be used is initialised into clusters before calculating centroids. The optimisation problem is to reduce the distance between the data points and the centroids of their corresponding cluster by re-calculating the centroids and re-assigning the clusters until convergence. This is performed by calculating Euclidian distance. In this example, only two dimensions are shown, but for sequencing data, each basepair position is considered one dimension. The number of clusters is determined heuristically by finding the elbow value, that corresponds to the position after which increasing k does not further reduce the cost function. **B** TT-Seq of HeLa cells with or without alpha-amanitin treatment around sense TSSs as annotated by Brown et al. (2019) (Brown 2019). **C** Classification approach of NELF, integrator, and mediator levels into NET-Seq based clusters. ChIP-Seq data preparation for machine learning approach. 50-bp long summed windows are generated with one window upstream and 6 windows downstream of the TSS of each gene. **D** Logistic regression approach depicting the data partitioning into training and testing sets and the sigmoid hypothesis function for classification. The gradient descent and cost function equations are presented in the lower blue box. m denotes the total number of training examples (total number of genes, i.e. 1845). J denotes the cost function, i denotes the training example (i.e. the gene), and j the feature (i.e. the window). The vector containing the 7 parameters is termed θ. $x^{(i)}$ is a vector containing the summed ChIP-Seq level for the i-th gene in the 7 windows. $y^{(i)}$ denotes the value to be predicted, i.e. the cluster. h is the hypothesis function. **E** P-values associated with the probability of parameters being significantly different from zero are displayed for NELF (purple), INTS3 (pink), and MED26 (gold), as determined by the Wald test, which tests the hypothesis whether the variable assigned to the window can be removed without affecting the prediction model. P-values in the left half of the table are determined by ANOVA which sequentially compares the model containing only the previous variables to the model containing the respective variable. Z-values indicate the magnitude of how this parameter influences cluster classification towards clusters 2-4. A negative number would indicate that higher levels in this window would make it more likely to classify the gene as belonging to cluster 1. The window to which the maximum z-values belong are indicated with a grey star. **F** Sensitivity vs FPR curves and the area under the curve for NELF, INTS3 and MED26.

Next, we used to ChIP-Seq data for NELF, INTS3 or MED26, to validate these peaks as pauses or sites of early termination (Fig. 5A). For all three factors, the most pronounced and most 5' peaks are found in clusters 2, 3, and 4 whereas cluster 1 shows very flat profiles. Cluster 6 has again a slightly further downstream and broader peak in its profile for all three factors compared to cluster 2, 3, and 4. Cluster 5 has a later peak, although only the mediator Chip-Seq profile (MED26) shows the dual peaks that were also observed in the ssSNU-Seq and the TT-Seq with alpha-amanitin treatment, suggesting that these peaks might represent alternative start sites. This was confirmed using CoPRO, a nuclear-run on variation of PRO-Seq that enriches for 5'end states of nascent RNAs (Tome et al. 2018) and ATAC-Seq data from HEK293 cells (Weltner et al. 2018) which revealed that the peak in cluster five is shifted significantly further downstream compared to the other clusters (Fig. 5A, adjusted p-value $< 3.5 \times 10^{-5}$). Boxplots show the difference between NELF levels when comparing cluster 1 to cluster 2 to 4 in particular (Fig. 5B). This difference is slightly less pronounced for INTS3 and MED26 levels, but particularly small when looking at the levels of nascent transcription as measured by SNU-Seq (without size-selection). This suggests that levels of NELF predict genes with a pause, but levels of nascent transcription do not. Finally, we used a simple machine learning algorithm (logistic regression) to assess whether NELF, integrator, or mediator levels alone are predictive of which cluster a gene may belong (pausy; cluster 2-4 or non-pausy; cluster 1) (Fig. S5C). NELF shows the strongest predictive power, compared to integrator or mediator subunits, in terms of being able to assign NET-Seq -based clusters to pausy or non-pausy. This supports the 5' spikes in the SNU-Seq data as primarily reflecting pausing around 60-80 nt downstream of the TSS which may be accompanied by early termination of transcription, reflecting the INT3 signal.

Genes that exhibit promoter-proximal pausing are enriched for transcription factors regulating developmental processes (Fig. 6)

To characterize the regulatory context of promoter-proximal pausing, we calculated a calibrated pausing signal by dividing the summed size-selected SNU-Seq signals for each gene (0 – 150 bp downstream of the TSS) by the sum of TT-Seq reads for each gene (0 – 500 bp downstream of the TSS, Fig. 6A). Gene ontology analysis on the top and bottom quintile of genes ranked by their relative pausing signal revealed that genes exhibiting pausing are involved in primary metabolic processes, cell-cycle regulation and DNA-damage

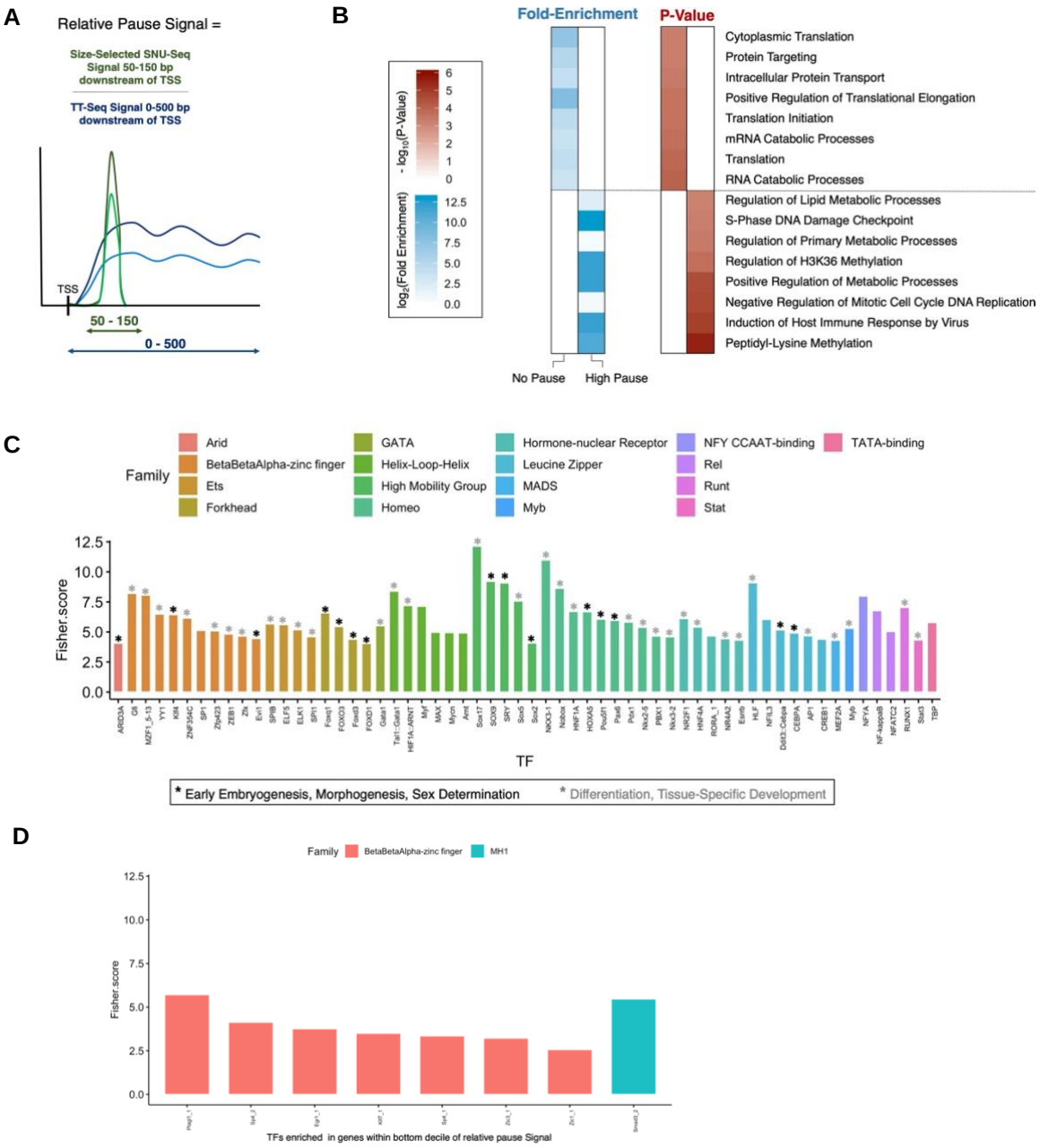


Figure 6 Gene Ontology Analysis of Genes with and without Promoter-Proximal Pausing/Early termination **A** Calculation of the relative level of the pause signal. **B** Fold-enrichment and p-values of gene ontologies as determined by GORILLA are shown for genes that exhibit high levels of calibrated pausing (see **A**) (top decile) or no pausing (bottom decile). log2-transformed values for the fold-enrichment and negative log10-transformed values for p-values are shown. The top and bottom decile each contained 934 genes (total number of genes was 9336). Genes with zero TT-Seq or zero SNU-Seq reads were ignored. **C** Transcription Factor Enrichment in Genes with High Levels of Pausing. **C** Transcription factor enriched in genes within the top decile (n = 934) of the relative pausing score are shown. A Fisher score threshold of 4 was applied. The transcription factor family is indicated by the bar colour. A black or grey star indicates whether the transcription factor is involved in early embryogenesis or tissue-specific development, respectively. **D** PLAGL1, a zinc finger protein known to suppress cell proliferation (Kas et al. 1998) is enriched at those genes with the lowest levels of relative pausing (bottom decile).

checkpoints (Fig. 6B). This agrees with previous analyses on promoter-proximal pausing based on Pol II ChIP-Seq (Day et al. 2016). We conclude that because of the single-nucleotide resolution, as well as the high sensitivity of our approach, we managed to generate highly statistically significant results on the gene ontology of promoter-proximal pause compared to the ChIP-Seq approach which generates low-resolution Pol II positioning profiles rather than precise information about the act of transcription.

To further characterise the biological context of promoter-proximal pausing, we used the same relative pausing score to identify groups of transcription factors that are enriched at genes that exhibit high levels of pausing (Fig. 6C). We find that that transcription factors regulating development are particularly enriched, which supports a range of previous studies in this field (Levine 2011; Gaertner et al. 2012; Smith and Shilatifard 2013; Gaertner and Zeitlinger 2014; Yang et al. 2016). MYC, which has recently been shown to regulate the promoter-proximal pausing stage of the transcription cycle by recruiting SPT5 (Baluapuri et al. 2020), also appears in our analysis. In the reverse analysis, we found that fewer transcription factors overall are enriched amongst those genes that show a lower relative pausing score (Fig. 6D). However, amongst them are EGR1 (Baron et al. 2006; Inoue and Fry 2018) and PLAGL1 (Huang et al. 2001; Abdollahi et al. 2003), both of which can act as tumour-suppressor genes via regulating p53.

Using ssSNU-Seq to annotate transcription start sites (TSS) and discover new TSS (Fig. 7)

Based on the 5' signal in the ssSNU-Seq data, we developed a pipeline that allows us to annotate transcription start sites (Fig. 7A). This is, in effect, using the same principle that has been used for TSS-annotations based on START-Seq (Nechaev et al. 2010; Scruggs et al. 2015; Lavender et al. 2016). However, with the 4sU-labelling and enrichment steps before size-selection, we utilized truly nascent RNA instead of chromatin-associated/nuclear RNAs used in the traditional START-Seq protocols. First, we carried forward only those 5'-end signals that overlap with ATAC-Seq peaks to ensure chromatin accessibility, for which we used previously published ATAC-Seq data in HEK293 cells (Weltner et al. 2018). We then merged TSSs together that were close enough together based on a previously published TSS-clustering algorithm used in CAGE-Seq approaches (Kodzius et al. 2006; Frith et al. 2008; Takahashi et al. 2012). The ssSNU-seq based clusters are predominantly shorter than 10 bp (Fig. S7D) and the cluster-centres map to ENCODE TSS annotations as expected (Fig. S7E). To verify transcriptional activity emanating from these TSS candidates we filtered out candidates for which HEK293 TT-Seq signal in the first 1000 bp downstream of the TSS candidate was lower than 5 times the signal of the 1000 bp preceding the TSS candidate. We identified a total of 4672 active TSSs (Fig 7E) and the confirmed TSSs are shown in a TT-Seq heatmap (Fig. 7F). We identify TSSs as either previously annotated (obsTSSs) or unknown (uTSSs) using the open-source code TSScall (Lavender et al. 2016). We checked that the previously annotated TSSs do in fact map around TSSs and found that a large fraction of our novel or unknown TSSs are upstream antisense TSSs or PROMPTs (Fig 7G). The newly identified antisense TSSs are predominantly divergent and are located between 100 and

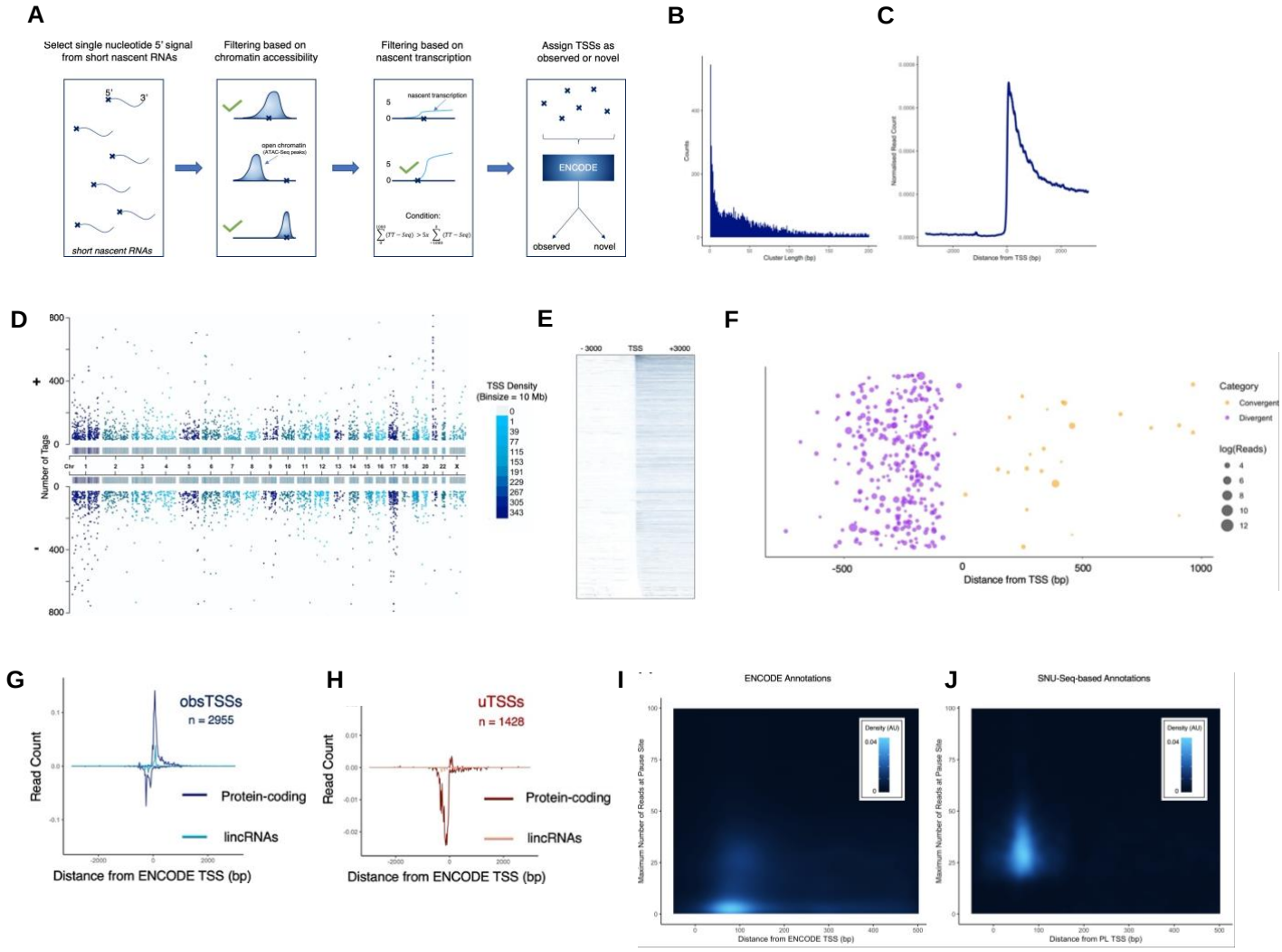


Figure 7 ssSNU-Seq based TSS characterisation. A simplified overview of the TSS annotation workflow. 5' end signals of size-selected SNU-Seq was filtered through ATAC-Seq peaks and, after clustering into TSS-clusters (not shown), filtered based on a minimum 5-fold increase in the TT-Seq signal immediately downstream of the TSS candidate. The verified TSSs were then classified as either previously annotated (obsTSSs) or novel/ un-annotated TSSs (uTSSs). **B C** Clustering of SNU-Seq based TSS candidates. **B** A histogram of the cluster lengths. **C** A metagene of the full-length clusters around ENCODE transcription start sites. The signal is normalised so that each locus contributes equally to the metagene, by making the sum of reads for each cluster equal to 1. **D** Manhattan plot of SNU-Seq-based TSS annotations. The plot shows the location of each verified TSS annotation on the respective strand (+ or -) and chromosome. The y-axis corresponds to the number of reads that were associated with the original 5' end signal. The bars show the density of TSSs within each chromosome on each strand. **E** A heatmap of TT-Seq around ssSNU-Seq based TSS annotations (signal range 0 - 500 reads), ordered by number of reads from 1 basepair downstream of the TSS in a descending order and displayed in a 6000-basepair window around the TSS. **F** Categories of SNU-Seq-based TSSs that have not been annotated previously are classified as either convergent or divergent if they are in the vicinity of already annotated TSSs and on the opposite strand. The size of the data points corresponds to the log-transformed read count. **G** Metagene of already annotated TSSs (obsTSSs, n = 2955) around protein-coding or lincRNA ENCODE TSSs. **H** Metagene of novel/un-annotated TSSs (uTSSs, n = 1428) around protein-coding or lincRNA ENCODE TSSs. **I,J** Density Plots of 3'ends of Short Nascent RNAs around Different TSS Annotations **I** A 2-dimensional density plot depicting the density of 3'ends of short nascent RNAs (from size- selected SNU-Seq). The maximum read value and its corresponding position was chosen, with the density indicated by the colour. The distance from the ENCODE TSS and maximum read value are indicated on the x-axis and y-axis, respectively. The density (arbitrary units) corresponds to the number of TSSs per pixel (n = 19874) **J** A 2-dimensional density plot depicting the density of 3'ends of short nascent RNAs (from size-selected SNU-Seq). The maximum read value and its corresponding position was chosen, with the density indicated by the colour. The distance from the SNU-Seq based TSSs and maximum read value are indicated on the x-axis and y-axis, respectively. The density (arbitrary units) corresponds to the number of TSSs per pixel (n = 2955).

500 bp upstream of the sense TSSs. A further piece of analysis was done in this context to see where the 3' ends of the short nascent RNAs are located with respect to either the ENCODE TSSs or the SNU-Seq based TSS annotations. This comparison is shown in Figure 6.5 for ENCODE annotations (A) and the SNU-Seq based annotations (B). Strikingly, the pause signals, when mapped around the ENCODE TSSs, are almost distributed horizontally, centred between 50 and 150 bp downstream of the TSS, but with lower densities positioned at a much wider distance range. In contrast, when mapping the 3'end of the short nascent RNAs around the SNU-Seq based annotations, they are distributed almost vertically, with a sharp positional distribution between 50 and 100 bp downstream of the TSSs. Only negligible numbers of 3'ends are located outside this range. This would also point to potential issues with the ENCODE database of TSS annotations, since some of these TSSs may not have been verified through nascent transcription filters, for example. They might therefore not reflect the true starting point of nascent transcription, but only reflect the most 5' end signal of (alternative) transcripts that were picked as the TSS, even though the true TSS from which some of the nascent transcription emanates is actually located further downstream.

In conclusion, size-selected SNU-Seq can be a viable basis for transcription start site annotation, especially because, in contrast to other methods such as CAGE, the transcripts used for the annotations are nascent. The way in which TSSs were annotated for this study relied on a set of very strict filters, such as ATAC-Seq peaks, and a 5-fold increase in TT-Seq signal, which lead to a relatively small set of TSSs. However, because of these filters, one can assume that these TSSs belong to transcription units that are transcriptionally active and found in open and accessible chromatin.

As noticed with TT-Seq for measuring antisense transcription, a 4sU-based approach seems particularly suitable when interested in detecting non-coding and antisense transcription. This holds true for 4sU-based TSS annotations obtained through size- selected SNU-Seq, too, given the high proportion of unknown uTSSs that are antisense to the TSSs of already annotated protein coding genes.

Discussion

The aim of this work was to develop and validate a method to interrogate nascent transcriptomes which combine the sensitivity of the 4sU pulse label used in TT-Seq with the resolution offered by nucleotide resolution techniques such as PRO-Seq and NET-Seq, but with simple library preparation to reduce cost and improve general accessibility. **S**ingle **N**ucleotide resolution 4s**U**-**Seq** (SNU-Seq) meets these criteria

and allows the generation of profiles of nascent transcripts in human cells with high resolution, high sensitivity and low cost. We are able to detect an even, uniform signal over gene bodies representing productive elongation and show the mathematical parameters describing transcription, based on synthesis-to-decay ratios, correlate well between SNU-Seq and TT-Seq. In this context, it is key to point out that, whilst sonication or chemical fragmentation during TT-Seq reduces the 5' bias, it does not completely eliminate it. SNU-Seq, because of the retention of full-length fragments followed by mapping the last incorporated nucleotide, does not have any 5' bias. In addition to the higher resolution, our protocol is very straightforward and can be carried out in less than two days. This is in part due to the library preparation; use of 3'end kits using oligo-d(T) primers is well established and optimised for convenience and a speedy workflow. Performing library preparation this way also greatly reduces costs compared to many of the kits or protocols used for NET-Seq, PRO-Seq or TT-Seq. This is also supported by the fact that approximately 65% of SNU-Seq reads uniquely align. We further find that our technique is robust and reproducible, producing similar profiles between repeats and different human cell lines (HEK293 and HEP3B). As SNU-Seq is much more straightforward that TT-Seq, it is suitable to derive relative elongation rates from mutants, drug treated cells, or during development and, as shown here, to characterise new enhancers and alternative sites of polyadenylation in IFN-γ treated HEP3B cells over time. In addition, SNU-Seq data can be used to parametrise simulations to enable other metrics of transcription to be derived. We show the utility of SNU-Seq for discovering new transcription units, particularly non-coding transcription units for example those transcribed divergent from annotated genes. Using IFNγ induced HEP3B cells we characterise hundreds of novel enhancers using ATAC-Seq peaks coupled with nascent transcription as the criterion for active enhancers.

SNU-Seq is not quite perfect; because of the simple library preparation method, for some genes, the reads at the extreme 5' end of nascent transcripts are not fully represented in the library. For routine assessment of the nascent transcriptome, elongation parameters or APA sites this is not problematic; it affects novel TSS mapping. ssSNU-Seq, with immediate adaptor ligation on both the 3' and 5' ends of the fragments, overcomes this limitation and detects novel TSSs, particularly those associated with unstable transcripts which are less likely to be represented in the encode database. In the HEK293 cells, many of the novel TSSs were from upstream antisense transcription units. Our initial analysis, using the encode TSS annotations and our *k*-means clustering, identified different pausing behaviours that can be measured by

both ssSNU-Seq and NET-Seq whereas TT-Seq lacks the resolution to do so. Crucially, our clustering approach raised the question of why short, nascent RNAs could lead to a promoter-proximal pause signal much further into the gene body (at around 300-400 bp downstream of the TSS). Rather than this being attributed to a 'late' pausing site, we discovered that this is actually occurring because of alternative TSS usage, by integrating ATAC-Seq and CoPRO sequencing data. This points to a key observation: When analysing nascent transcription and regulatory events during the transcription cycle, a comprehensive and sensible approach would be to integrate a range of different methods due to their various benefits and strengths.

When considering the biological purpose of the promoter-proximal pause, our data confirm, on a genome-wide level, what has previously been observed on a gene-by-gene basis or with lower resolution approaches: We find that genes that undergo promoter-proximal pausing are frequently involved in cell cycle regulation and DNA-damage response (Day et al. 2016). Crucially, our enrichment analysis shows transcription factors regulating development are enriched at genes exhibiting high levels of pausing. This has been a heavily studied topic, particularly in *Drosophila*, with the argument being made that the purpose of pausing is not limited to keeping RNA Polymerase II in a poised state, ready for rapid induction. Instead, pausing is suggested to occur in order to promote the synchronisation of transcriptional induction through the avoidance of stochastic or 'bursty' transcription (Gaertner and Zeitlinger 2014), which would be particularly relevant during development.

At the 3' ends of genes, many SNU-Seq spikes are coincident with polyadenylation sites. Interestingly, we show that at *STAT1* both the position and number of these sites changes over time of IFNγ induction. Again, the position of many (but not all) of the nascent spikes overlap with those in the mRNA transcript population. The fate of these transcripts with different 3' UTRs or 3' ends remain to be determined. This may be particularly important as, in HEK293 cells, many SNU-Seq peaks colocalise with PAS in nuclear polyadenylated RNA preparations, but not in the cytoplasmic pool, suggesting the transcripts are targeted for degradation in the cytoplasm and are unlikely therefore to function in gene expression. We note that there are many spikes in the SNU-Seq data, in addition to those at annotated PASs, perhaps reflecting the potential for differential 3' end formation and gene expression.

Methods

Resources

Sequencing Data

Dataset (+ Citation)	Cell Line	Accession / Source
SNU-Seq	HEK293	This Study
SNU-Seq	HEP3B	This Study
SNU-Seq with size-selection	HEK293	This Study
TT-Seq	HEK293	This Study
TT-Seq (Schwalb *et al*., 2016)	K562	GSE75792
NET-Seq (Nojima *et al*., 2015)	HeLa	GSE60358
GRO-Seq (Laitem *et al*., 2015)	HeLa	E-MTAB-3360
CoPRO (Tome *et al*., 2018)	K562	GSE116472
ATAC-Seq (Weltner *et al*., 2018)	HEK293	E-MTAB-6195
RNA-Seq (Banks *et al*., 2014)	HEK293	E-GEOD-57027

Kits and Reagents

Kit / Reagent Name	Provider	Code
NEBNext rRNA Depletion Kit (Human/Mouse/Rat)	NEB	E6350S
NEBNext Multiplex Oligos (Set 1)	NEB	E7335L
NEBNext Ultra II Directional RNA Library Prep	NEB	E7765S
NEBNext Small RNA Library Prep Kit	NEB	E7300S
QuantSeq 3’end Kit for Ion Torrent	Lexogen	012.24A
uMACS Streptavidin Kit	Miltenyi Biotech	130-074-101
NextSeq 500/550 High Output v2.5 kit (150 cycles)	Illumina	20024907
4-thiouridine	Carbosynth	NT06186
EZ-link HPDP-Biotin	Thermo Fisher	21341
SnakeSkin Dialysis Tubing, 16 mm	Thermo Fisher	88242
E.coli Poly(A) Polymerase Kit	NEB	M0276S

RNA 5' Pyrophosphohydrolase (RppH)	NEB	M0356S
ThermoPol® Reaction Buffer Pack	NEB	B9004S
miRNeasy Micro Kit	Qiagen	217084
5-PRIME Phase Lock Gel Heavy	VWR	733-2478

Software / Packages

Software / Package / Tool Name	Citation / Source
Trim-Galore	Babraham Bioinformatics
Trimmomatric	(Bolger et al. 2014)
fastqc	Babraham Bioinformatics
Hisat2	(Kim et al. 2019)
Bowtie2	(Langmead and Salzberg 2012)
samtools	(Li et al. 2009)
bedtools	(Quinlan and Hall 2010)
Macs2	(Zhang et al. 2008)
wigToBigWig	UCSC
featureCount	(Liao et al. 2014)
Paraclu	(Frith et al. 2008)
TSScall	(Lavender et al. 2016)

Cell Culture

HEK293 and HEP3B cells were cultured in DMEM supplemented with 10 % (v/v) FBS and 1 % (v/v) Penicillin-Streptomycin. The incubator was set to 37 ˚C at 5 % CO_2. HEK293 cells were passaged 24 hours before harvesting for next-generation sequencing experiments at which point they were grown to approximately 80% confluency. Cells were counted on a Nexcelom Biosciences Auto 2000 Cell Counter. Hep3B cells were passaged 48hrs before harvesting at 80% confluency. After passage they were left for 24hrs to adapt, then left untreated or treated with 10 ng/ml of IFN for 0.5, 2 or 24 hrs unless stated otherwise.

TT-Seq

TT-Seq was performed as described in the original publication (Schwalb et al. 2016). Sequencing libraries were prepared using the NEBNext rRNA Depletion kit for Human Cells followed by the NEBNext II Ultra Directional RNA Library prep kit, following both kit instructions. Library quality was assessed using the

Agilent Bioanalyzer with a DNA-High Sensitivity chip. Pooled libraries were sequenced on the Illumina NextSeq 500 platform using the High-Output Illumina Nextseq 500 kit.

For calibrated TT-Seq with spike-ins, we followed the instructions for thio-labelled spike-in usage described in (Schwalb et al. 2016) Schwalb *et al.* (2016). Alpha-amanitin treatment of was carried out by incubating HEK293 cells for 6 hours with or without addition of 25 ug/mL alpha-amanitin before following the TT-seq protocol with spike-ins.

TT-Seq - Sequencing Analysis

Paired-end sequencing reads were quality-checked with fastqc and then trimmed using Trim-Galore to remove sequences with a quality below a Q score of 20. Trimmed reads were then aligned to the human genome (hg38 assembly) with hisat2 (Kim *et al*., 2015) using the –no-mixed and the -no-discordant flags. Aligned files in the sam format were then filtered by using samtools (Li *et al*., 2009) with the flags -q 40, -f 99, -F3852, -bS. Calibration of samples (if necessary) was achieved by calculating scaling factors of spike-in counts between samples based on counts tables generated by the bioconductor package featureCounts (Liao *et al*., 2013). Bedgraph and bigwig files were generated from bam files using bedtools (Quinlan and Hall, 2010), and wigTobigWig (UCSC Genome Browser), respectively.

SNU-Seq without size-selection

SNU-Seq libraries were generated by following the TT-Seq protocol (Schwalb *et al*., 2016), but omitting the sonication step. Before library preparation, nascent RNAs were polyadenylated using the NEB Poly(A)-Polymerase kit following the kit instructions. The reaction with 150 ng thiolabeled RNA was left for 45 minutes at 37 ˚C before isopropanol precipitation. The pellet was resuspended in 11 μL RNase-free water. Qualities and amounts were checked on the Qbit fluorometer and the Agilent Bioanalyzer (RNA Pico Chip).

$6x10^6$ Hep3B cells were treated with 10 ng/ml IFN for 0, 0.5, 2 or 24 hrs. During the final 12 min of IFN treatment, 500 μM 4sU was additionally added. Following treatments, cells were washed with cold PBS before lysing in Qiazol on ice. Samples were spiked with *Saccharomyces cerevisiae* RNA (0.01%). 300 μg RNA was biotinylated.

Libraries were prepared following the Quant-Seq Lexogen 3' mRNA kit instructions with 13 PCR cycles. Library qualities were checked using the Agilent Bioanalyzer (DNA High Sensitivity Chip). Following the manufacturer's instructions, the chips prepared by the Ion Chef were then sequenced on an Ion Proton Sequencing platform.

SNU-Seq without size-selection – Sequencing Analysis

After quality control with fastqc, fastq files were trimmed with trimmomatric (Bolger *et al*., 2014) to remove reads with a quality score below 20 in a sliding window of 5 bp. The poly(A) reads were removed with the clip option. Sequences were then aligned with hisat2 with the same settings as for TT-Seq. Sorted bam files were generated using samtools. Calibration of samples (if necessary) was achieved by calculating scaling factors of spike-in counts between samples based on counts tables generated by the R package featureCounts. Bedgraph and bigwig files were generated from bam files using bedtools, and wigTobigWig, respectively. 3'end single basepair coverage was achieved by using the bedtools -bg -3 option.

SNU-Seq with size-selection

Nascent, thio-labelled RNA was generated as described for SNU-Seq without size selection. The nascent RNA was then run on a 3.5 % TBE-Urea gel (8M urea). For this, 250 ng of nascent RNA was mixed with 2x loading dye and boiled for 2 minutes at 80 ˚C. The gel was kept on ice and pre-run for 10 minutes at 80 V, using ice-cold ultra-pure TBE (1x) as the running buffer. Wells of the gel were washed with TBE before loading samples. Samples were then run at 80 V for 90 minutes. The gel was transferred onto a square petri dish and incubated in 1x ultra-pure TBE with SYBR Gold (1/10,000 v/v) for 5 minutes. The gel was rinsed twice in 1x ultra-pure TBE before imaging.

For purification of the small nascent RNA fraction, the gel was incubated at -80 ˚C for 10 minutes. The small band of 70-100 bp size was cut using a razor blade. We extracted the RNA from the gel slice using dialysis: The gel slice, along with 0.8 mL 10 mM Tris-HCl (pH 7), was placed into SnakeSkin dialysis membrane sealed by two Eppendorf caps. The dialysis was run for 35 minutes at 80 V. The RNA was then purified with isopropanol precipitation and eluted in 11 uL RNase-free water. Quality and size distribution of the size-selected RNA was checked on the Agilent Bioanalyzer using an RNA pico chip.

The size selected nascent RNA was treated with 5' Pyrophosphohydrolase (NEB) in ThermoPol® reaction buffer following the manufacturer's instructions. The reaction was stopped with 1 uL 500 mM EDTA and heat-inactivated by incubation at 65 ˚C for 5 minutes. RNA was again purified with isopropanol

precipitation and resuspended in 6 uL RNase-free water. Libraries were generated using the NEBNext Small RNA Library kit for Illumina following the kit instructions. Library quality was assessed with the Agilent Bioanalyzer using a DNA High Sensitivity chip. Sequencing was performed on an Illumina NextSeq 500 platform.

SNU-Seq with size-selection – Sequencing Analysis

Fastq files were quality-trimmed with Trim-Galore and aligned with bowtie2 (Langmead and Salzberg, 2012). Sorted bam files were generated from aligned sam files using samtools with a filtering step added to only retain reads that have a mapping quality score above 30 (-q 30). As above, 3'end single basepair resolution was achieved by using the bedtools -bg -3 option when generating bedgraph files.

TSS annotations

To generate TSS annotations from the size-selected SNU-Seq samples, 5' end coverage was generated using the -bg -5 option (instead of -3) when generating bedgraphs from bam files in bedtools. To generate 5'end reads that only occur in nucleosome-depleted regions, ATAC-Seq data in HEK293 cells (Weltner *et al.,* 2018, E-MTAB-6195) were used. The ATAC-Seq fastq files were trimmed with Trim-Galore, aligned with bowtie2, and converted to bam and bed files using samtools and bedtools, respectively. Subsequently, MACS2 (Zhang *et al*., 2008) was used to call peaks (no model assumed). The intersect option in bedtools was then used to retain only those 5'end SNU-Seq (size-selected) signals that are located within ATAC-Seq peaks. The resulting bedgraph files were then used as an input to identify TSS cluster centres using Paraclu (Frith *et al*., 2008) which identifies clusters in a sliding window, and a cluster value threshold of 30 was applied. Cluster-centres were calculated using the mean position of the cluster, and TSS candidates were verified in Matlab as follows: Only those TSS candidates were assigned as true active TSSs where the summed TT-seq signal (using HEK WT 10 minute-labelled TT-Seq) in the 1000 bp downstream of the candidate TSS was at least 5 times greater than in the 1000 bp upstream of the candidate TSS. TSScall, a python code (Lavender *et al*., 2016), was then used to identify annotated and un-annotated (novel) TSSs using the comprehensive GENCODE (Frankish et al. 2019) (v29) annotation file which yielded 2955 previously annotated TSSs and 1428 novel transcription start sites. The read threshold for this was set to 5, which corresponded to an FDR of 0.001.

Mathematical Modelling and Determination of Transcription Constants

Synthesis and decay rates were determined based on a previously published modelling approached (Villamil *et al*., 2019). HEK293 RNA-seq counts were used from Banks et al. (2014). The following equations were used:

$\lambda_{i,j} = \frac{1}{t} \times log_{10}(1 - \frac{L_{i,j}}{T_{i,j}})$ and

$\mu_{i,j} = T_{i,j} \times \lambda_{i,j}$

where λ_{ij} is the decay rate, t is the labelling time, μ_{ij} is the synthesis rate, T_{ij} is the RNA-seq counts (length-normalised tags per million, TPM) and L_{ij} is the SNU-seq counts (TPM). Scaling factors will become cancelled out in the calculation of the Synthesis Rate / Decay Rate ratio.

Bioinformatics and Data Analysis

For metagenes, we used TSS annotations that we generated previously (Brown *et al*., 2018), or, if indicated otherwise, ENCODE TSS and poly(A)-site annotations (v29). In MATLAB, we normalised the read counts for each gene in the specified window to a total sum of 1, in order to make each gene contribute equally to the metagene plots. At each position, we removed the top and bottom 0.5 % of entries, and the final signals were then averaged over 10bp-moving windows. p-values were determined using the non-parametric Wilcoxon rank sum test, and the Bonferroni correction for multiple testing was applied when required. To determine the correlation between repeats, we calculated the Pearson correlation coefficient based on the counts tables. In all other cases, the Spearman ranked correlation (rho) was calculated. Principle Component Analysis (PCA) and data visualization were performed in R using the PCAtools and ggplot2 packages, respectively.

Cluster Analysis

Clustering was performed on standardised profiles of mNETseq reads. All genes greater than 1,000bp in length and with total mNETseq reads across the whole genes expressing above background levels for the entire genome (t-test, p>0.01) were included. K-means clustering was then performed on the standardised profiles in MATLAB, with a value of K=6 determined as the highest value of K which explained the variance across the entire set of genes without redundancy (elbow-method). Heuristically, this meant that we chose enough groups so that each group has a distinct average profile and choosing any more groups would result in splitting a group into two almost-identical subgroups. We then plotted normalised size-selected SNU-Seq, TT-Seq (with or without alpha-amanitin treatment), ATAC-Seq, and CoPRO (5') signals around TSSs

within the same cluster annotations. Metagenes for these sequencing methods were generated as above, however the removal of the top and bottom 0.5% of signals was omitted apart from the CoPRO metagenes.

Nuclear and cytoplasmic RNA extraction

Extraction of RNA from nuclear and cytoplasmic subcellular fractions was carried out as described in (Fischl et al., 2020). This was performed for 3 biological replicates on HEK293 cells cultured in DMEM containing 10 % (v/v) FBS, 2.5 mM L-Glutamine and 1 % Penicillin/Streptomycin (Sigma cat#P0781) in a 37°C incubator with 5 % CO2 to 80-90 % confluency before harvesting.

QuantSeq

Barcoded libraries for multiplexed, strand-specific sequencing of the 3' end of polyadenylated RNAs were generated using the QuantSeq 3' mRNA-Seq Library Prep Kit for Ion Torrent (Lexogen) as per the manufacturer's instructions, using 500 ng and 1700 ng input RNA for nuclear and cytoplasmic RNA samples, respectively, and using 13 PCR cycles. Libraries were loaded onto the Ion Chef System (ThermoFisher) for template preparation and chip loading and the resulting chips were sequenced on the Ion Proton Sequencing System (ThermoFisher) as per the manufacturer's instructions.

Reads were aligned to the hg19 genome build using the Ion Torrent Server TMAP aligner with default alignment settings (-tmap mapall stage1 map4). Human polyA site (PAS) annotations were obtained from PolyA_DB3 (Wang et al. 2018). Each PAS was extended 20 nt 3' and 200 nt 5' from the site of cleavage and those that overlapped on the same strand after extension were combined into a single PAS annotation. Mapped reads were narrowed to their 3' most nucleotide and those which overlapped with the extended PAS annotations were counted. PASs associated with non-coding RNAs and genes not in the RefSeq (O'Leary et al. 2016) gene database were excluded. Genes with only one PAS were also excluded.

Alternative polyadenylation analysis

Genes showing alternative polyadenylation (APA) between nuclear and cytoplasmic fractions were determined using the DEXseq algorithm (Anders et al. 2012). While designed for analysis of differential exon usage, this algorithm was adapted to analyse differential PAS usage by using PAS counts in place of exon counts. For each PAS, this then gives a base mean value, which is the mean of normalized counts for all samples after normalizing for sequencing depth, the log2 fold change in relative usage between the nuclear and cytoplasmic samples being compared and a p-value, adjusted for multiple testing using the Benjamini-Hochberg method (padj), showing the significance of any difference.

For each gene, the results were filtered to select the PAS with the lowest padj, excluding genes without a PAS showing a padj less than 0.01 and base mean greater than 30. For each of these PASs, the PAS within the same gene was selected showing the most significant opposing log2 fold change in relative usage and also a base mean greater than 30. This then gives a pair of PASs for each gene showing a significant APA event, with one PAS showing greater relative usage in the cytoplasm and the other in the nucleus. To increase reliability, both PASs then had to have at least 5 reads per million according to the original PolyA_DB3 database. Genes were then further filtered to limit events to those where both PASs are in the 3'UTR to give a final list of 22 stringent 3'UTR APA events.

Chromatin analysis in HEP3B cells

ATAC-seq was performed based on methods previously described (Buenrostro et al. 2015). $5x10^4$ Hep3B cells were treated with 10 ng/ml IFN for 0, 0.5, 2 or 24 hrs. Cells were washed in cold PBS (containing IFN for treated groups) and resuspended in lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL). The nuclei were pelleted and resuspended in 1x TD with 2.5 l TDE1 (Nextera DNA library prep kit, Illumina). These were incubated for 30 min at 37°C. The tagmented DNA was purified using the MinElute PCR purification kit (Qiagen, 28004). Each sample was amplified using the Nextera DNA library prep kit and Nextera index kit (Illumina) with the following thermocycler programme: hold at 72°C (5 min), 98°C (30 s), 9 cycles of 98°C (10 s), 63°C (30 s), 72°C (30 s) with a final 72°C extension for 1 min. The libraries were purified by adding 1.8 x volume of RT AxyPrep beads (Appleton Woods AxyPrep Mag PCR clean up) to each reaction. They were then size selected with a ratio of 0.6:1 beads to remove fragments larger than 600bp. Libraries were run on the Agilent Bioanalyzer with a DNA-High Sensitivity chip to assess quality. Paired-end sequencing was performed on a NextSeq 500 with the 75 cycles NextSeq 500/550 High Output v2 kit (Illumina). FastQCs were performed on each dataset to ensure good quality of the run. The adapter sequences were trimmed and paired using Trimmomatic, removing reads with a quality score below 20 in a 5bp sliding window and reads below a minimum length of 30bp (Bolger et al. 2014). Trimmed reads were aligned to the hg38 genome if un-spiked, or to a combined hg38-dm6 genome if spiked, using Bowtie2 (Langmead and Salzberg 2012). SAMtools were used to remove duplicates and filter out reads with a MAPQ quality score <30 as well as mitochondrial reads (Li et al. 2009). MACS2 was used to call peaks with a minimum FDR (q-value) of 0.01 (Zhang et al. 2008). Following this, ENCODE blacklisted regions were removed.

For visualisation of tracts, reads were normalised based on sequencing depth. Filtered BAM files were converted to bedgraphs using BEDtools (Quinlan and Hall 2010).

The ChIPmentation protocol used is largely based on the protocol published by Schmidl et al. (Schmidl et al. 2015). 10^6 cells were treated with 10 ng/ml IFN for 0, 0.5, 2 or 24 hrs. Cells were washed and collected in cold PBS (containing IFN for treated groups). The cell pellet was resuspended in RT PBS and fixed for 5 min with 1% formaldehyde before quenching with 0.125 M Glycine. 5000 fixed *Drosophila* sg4 cells (0.05%) were added as a spike-in control. The cell pellet was obtained, washed in cold PBS and incubated on ice in cold swelling buffer (10 mM Tris-HCl pH 8, 10 mM NaCl, 0.2% NP-40, 1 mM AEBSF and 1x complete mini EDTA-free proteasome inhibitor, Roche) for 10 min. The subsequent nuclear pellet was resuspended in cold lysis buffer (10 mM Tris-HCl pH 8, 1% NP-40, 0.5% Na- deoxycholate, 0.1% SDS, 1 mM AEBSF and 1x complete mini EDTA-free proteasome inhibitor, Roche) and sonicated for 90 min, 30 s on/30 s off at 4°C (Bioruptor, Diagenode). One tenth of the sample was taken as input control. The remaining sample was incubated overnight, rotating at 4°C with antibody; anti-H3K4me3 (MERCK, 05-745R), anti-H3K27ac (Millipore, 07-360), anti-CTCF (CST, 3418S), anti-H2AV *Drosophila* Ab (Active Motif, 39715) and PE Mouse IgG2b, Isotype Ctrl Antibody (Biolegend, 400313).

Scaling factors were calculated based on spike-in counts between samples, or based on sequencing depth for CTCF ChIP, and used when converting the filtered BAM files to bedgraphs using BEDtools (Quinlan and Hall 2010). Bedgraphs were further converted to BigWig files using the UCSCtools (UCSC genome browser).

Data Access

The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus (Edgar et al. 2002) and are accessible through GEO Series accession numbers shown below.

Data in the HEP3B cells is available at GSE172053.
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE172053. Reviewer token is: mxwzyqqozhuxnyp

Data in HEK293 cells is available at GSE179306.
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE179306. Reviewer token is: wbkzgoeejbulror

Data on polyadenylation in HEK293 cells is available at GSE165251.
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE165251. Reviewer token is: mloxkqoqprcjvqb

Competing interest statement

JM acts as an advisor to and/or holds stock in Oxford Biodynamics plc., Chronos Therapeutics Ltd., and Sibelius Natural Products Ltd. None of these companies has any interest in the data presented in this manuscript.

Acknowledgements

This work was supported by: The Wellcome Trust (WT089156MA to J.M.); the BBSRC (BB/P00296X/1 to J.M.); the Leverhulme Trust (RPG-2016-405 to J.M.); a Wellcome Trust Strategic Award (091911); BBSRC and EPSRC studentships (BB/M011224/1 to P.L. and EP/F500394/1 to T.B.); a Wellcome Trust studentship (209897/Z/17/Z to A.L.) and a Royal Society University Research Fellowship (UF120327 to A.A.). We thank Shona Murphy for critical insights during this development of this technique.

Author Contributions

SNU-Seq and SNU-Seq with size-selection was developed and performed by PL and SX. Bioinformatics analysis was done by PL, SM and CG. SNU-Seq in HEP3B cells together with chromatin analysis and bioinformatics was carried out by AL. Mathematical modelling and determination of transcription rate constants was carried out by ASK and AA. HF conducted the polyadenylation analysis. TB and PL performed the k-means clustering and downstream analysis. Data was visualized by PL. JM and PL conceived the experiments. The manuscript was written by PL and JM with additional input from AA, HF, AL and other authors.

References

Abdollahi A, Pisarcik D, Roberts D, Weinstein J, Cairns P, Hamilton TC. 2003. LOT1 (PLAGL1/ZAC1), the candidate tumor suppressor gene at chromosome 6q24-25, is epigenetically regulated in cancer. *The Journal of biological chemistry* **278**: 6041-6049.

Akbalik G, Langebeck-Jensen K, Tushev G, Sambandan S, Rinne J, Epstein I, Cajigas I, Vlatkovic I, Schuman EM. 2017. Visualization of newly synthesized neuronal RNA in vitro and in vivo using click-chemistry. *RNA Biol* **14**: 20-28.

Anders S, Reyes A, Huber W. 2012. Detecting differential usage of exons from RNA-seq data. *Genome Res* **22**: 2008-2017.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455-461.

Babour A, Shen Q, Dos-Santos J, Murray S, Gay A, Challal D, Fasken M, Palancade B, Corbett A, Libri D et al. 2016. The Chromatin Remodeler ISW1 Is a Quality Control Factor that Surveys Nuclear mRNP Biogenesis. *Cell* **167**: 1201-1214 e1215.

Baluapuri A, Wolf E, Eilers M. 2020. Target gene-independent functions of MYC oncoproteins. *Nature reviews* **21**: 255-267.

Banks CA, Lee ZT, Boanca G, Lakshminarasimhan M, Groppe BD, Wen Z, Hattem GL, Seidel CW, Florens L, Washburn MP. 2014. Controlling for gene expression changes in transcription factor protein networks. *Mol Cell Proteomics* **13**: 1510-1522.

Baron V, Adamson ED, Calogero A, Ragona G, Mercola D. 2006. The transcription factor Egr1 is a direct regulator of multiple tumor suppressors including TGFbeta1, PTEN, p53, and fibronectin. *Cancer Gene Ther* **13**: 115-124.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* **30**: 2114-2120.

Brown T. 2019. Study of the dynamics of gene expression by mathematical modelling and systems approaches. In *Biochemistry*, Vol D.Phil. University of Oxford.

Brown T, Howe FS, Murray SC, Wouters M, Lorenz P, Seward E, Rata S, Angel A, Mellor J. 2018. Antisense transcription-dependent chromatin signature modulates sense transcript dynamics. *Molecular systems biology* **14**: e8007.

Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. 2015. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current protocols in molecular biology / edited by Frederick M Ausubel [et al* **109**: 21 29 21-21 29 29.

Catarino RR, Stark A. 2018. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes & development* **32**: 202-223.

Churchman LS, Weissman JS. 2011. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**: 368-373.

Churchman LS, Weissman JS. 2012. Native elongating transcript sequencing (NET-seq). *Current protocols in molecular biology / edited by Frederick M Ausubel [et al* **Chapter 4**: Unit 4 14 11-17.

Cleary MD, Meiering CD, Jan E, Guymon R, Boothroyd JC. 2005. Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nature biotechnology* **23**: 232-237.

Core L, Adelman K. 2019. Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. *Genes & development* **33**: 960-982.

Core LJ, Lis JT. 2008. Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science (New York, NY* **319**: 1791-1792.

Day DS, Zhang B, Stevens SM, Ferrari F, Larschan EN, Park PJ, Pu WT. 2016. Comprehensive analysis of promoter-proximal RNA polymerase II pausing across mammalian cell types. *Genome Biol* **17**: 120.

Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* **30**: 207-210.

Elrod ND, Henriques T, Huang KL, Tatomer DC, Wilusz JE, Wagner EJ, Adelman K. 2019. The Integrator Complex Attenuates Promoter-Proximal Transcription at Protein-Coding Genes. *Molecular cell* **76**: 738-752 e737.

Fischl H, Howe FS, Furger A, Mellor J. 2017. Paf1 Has Distinct Roles in Transcription Elongation and Differential Transcript Fate. *Molecular cell* **65**: 685-698.e688.

Fischl H, Neve J, Wang Z, Patel R, Louey A, Tian B, Furger A. 2019. hnRNPC regulates cancer-specific alternative cleavage and polyadenylation profiles. *Nucleic acids research* **47**: 7580-7591.

Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M et al. 2017. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* **2017**.

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

Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, Sandelin A. 2008. A code for transcription initiation in mammalian genomes. *Genome Res* **18**: 1-12.

Fuchs G, Voichek Y, Benjamin S, Gilad S, Amit I, Oren M. 2014. 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome Biol* **15**: R69.

Gaertner B, Johnston J, Chen K, Wallaschek N, Paulson A, Garruss AS, Gaudenz K, De Kumar B, Krumlauf R, Zeitlinger J. 2012. Poised RNA polymerase II changes over developmental time and prepares genes for future expression. *Cell Rep* **2**: 1670-1683.

Gaertner B, Zeitlinger J. 2014. RNA polymerase II pausing during development. *Development* **141**: 1179-1183.

Gregersen LH, Mitter R, Svejstrup JQ. 2020. Using TTchem-seq for profiling nascent transcription and measuring transcript elongation. *Nature protocols* **15**: 604-627.

Hoque M, Ji Z, Zheng D, Luo W, Li W, You B, Park JY, Yehia G, Tian B. 2013. Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat Methods* **10**: 133-139.

Hou TY, Kraus WL. 2021. Spirits in the Material World: Enhancer RNAs in Transcriptional Regulation. *Trends in biochemical sciences* **46**: 138-153.

Huang SM, Schonthal AH, Stallcup MR. 2001. Enhancement of p53-dependent gene activation by the transcriptional coactivator Zac1. *Oncogene* **20**: 2134-2143.

Inoue K, Fry EA. 2018. Tumor suppression by the EGR1, DMP1, ARF, p53, and PTEN Network. *Cancer Invest* **36**: 520-536.

Kas K, Voz ML, Hensen K, Meyen E, Van de Ven WJ. 1998. Transcriptional activation capacity of the novel PLAG family of zinc finger proteins. *The Journal of biological chemistry* **273**: 23026-23032.

Kawaji H, Kasukawa T, Forrest A, Carninci P, Hayashizaki Y. 2017. The FANTOM5 collection, a data series underpinning mammalian transcriptome atlases in diverse cell types. *Sci Data* **4**: 170113.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37**: 907-915.

Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M et al. 2006. CAGE: cap analysis of gene expression. *Nat Methods* **3**: 211-222.

Kwak H, Fuda NJ, Core LJ, Lis JT. 2013. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science (New York, NY* **339**: 950-953.

Laitem C, Zaborowska J, Isa NF, Kufs J, Dienstbier M, Murphy S. 2015. CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II-transcribed genes. *Nature structural & molecular biology* **22**: 396-403.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.

Larke MSC, Schwessinger R, Nojima T, Telenius J, Beagrie RA, Downes DJ, Oudelaar AM, Truch J, Graham B, Bender MA et al. 2021. Enhancers predominantly regulate gene expression during differentiation via transcription initiation. *Molecular cell* **81**: 983-997 e987.

Lavender CA, Cannady KR, Hoffman JA, Trotter KW, Gilchrist DA, Bennett BD, Burkholder AB, Burd CJ, Fargo DC, Archer TK. 2016. Downstream Antisense Transcription Predicts Genomic Features That Define the Specific Chromatin Environment at Mammalian Promoters. *PLoS Genet* **12**: e1006224.

Levine M. 2011. Paused RNA polymerase II as a developmental checkpoint. *Cell* **145**: 502-511.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**: 2078-2079.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)* **30**: 923-930.

Liu P, Xiang Y, Fujinaga K, Bartholomeeusen K, Nilson KA, Price DH, Peterlin BM. 2014. Release of positive transcription elongation factor b (P-TEFb) from 7SK small nuclear ribonucleoprotein (snRNP) activates hexamethylene bisacetamide-inducible protein (HEXIM1) transcription. *The Journal of biological chemistry* **289**: 9918-9925.

Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, Waters CT, Munson K, Core LJ, Lis JT. 2016. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nature protocols* **11**: 1455-1476.

Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y, Adelman K. 2010. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science (New York, NY* **327**: 335-338.

Neve J, Patel R, Wang Z, Louey A, Furger AM. 2017. Cleavage and polyadenylation: Ending the message expands gene regulation. *RNA Biol* **14**: 865-890.

Nojima T, Gomes T, Grosso AR, Kimura H, Dye MJ, Dhir S, Carmo-Fonseca M, Proudfoot NJ. 2015. Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161**: 526-540.

Nojima T, Rebelo K, Gomes T, Grosso AR, Proudfoot NJ, Carmo-Fonseca M. 2018. RNA Polymerase II Phosphorylated on CTD Serine 5 Interacts with the Spliceosome during Co-transcriptional Splicing. *Molecular cell* **72**: 369-379 e364.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **44**: D733-745.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**: 841-842.

Schlackow M, Nojima T, Gomes T, Dhir A, Carmo-Fonseca M, Proudfoot NJ. 2017. Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs. *Molecular cell* **65**: 25-38.

Schmidl C, Rendeiro AF, Sheffield NC, Bock C. 2015. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat Methods* **12**: 963-965.

Schwalb B, Michel M, Zacher B, Frühauf K, Demel C, Tresch A, Gagneur J, Cramer P. 2016. TT-seq maps the human transient transcriptome. *Science (New York, NY* **352**: 1225-1228.

Scruggs BS, Gilchrist DA, Nechaev S, Muse GW, Burkholder A, Fargo DC, Adelman K. 2015. Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Molecular cell* **58**: 1101-1112.

Sidaway-Lee K, Costa MJ, Rand DA, Finkenstadt B, Penfield S. 2014. Direct measurement of transcription rates reveals multiple mechanisms for configuration of the Arabidopsis ambient temperature response. *Genome Biol* **15**: R45.

Smith E, Shilatifard A. 2013. Transcriptional elongation checkpoint control in development and disease. *Genes & development* **27**: 1079-1088.

Stadelmayer B, Micas G, Gamot A, Martin P, Malirat N, Koval S, Raffel R, Sobhian B, Severac D, Rialle S et al. 2014. Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes. *Nat Commun* **5**: 5531.

Stoeckius M, Grun D, Kirchner M, Ayoub S, Torti F, Piano F, Herzog M, Selbach M, Rajewsky N. 2014. Global characterization of the oocyte-to-embryo transition in Caenorhabditis elegans uncovers a novel mRNA clearance mechanism. *The EMBO journal* **33**: 1751-1766.

Takahashi H, Kato S, Murata M, Carninci P. 2012. CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. *Methods in molecular biology (Clifton, NJ* **786**: 181-200.

Takahashi H, Ranjan A, Chen S, Suzuki H, Shibata M, Hirose T, Hirose H, Sasaki K, Abe R, Chen K et al. 2020. The role of Mediator and Little Elongation Complex in transcription termination. *Nat Commun* **11**: 1063.

Tome JM, Tippens ND, Lis JT. 2018. Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nature genetics* **50**: 1533-1541.

Tyssowski KM, DeStefino NR, Cho JH, Dunn CJ, Poston RG, Carty CE, Jones RD, Chang SM, Romeo P, Wurzelmann MK et al. 2018. Different Neuronal Activity Patterns Induce Different Gene Expression Programs. *Neuron* **98**: 530-546 e511.

Uzun U, Brown T, Fischl H, Angel A, Mellor J. 2021. Spt4 Facilitates the Movement of RNA Polymerase II through the +2 Nucleosomal Barrier. *BioRxiv* *https://doiorg/101101/20210303433772*.

Wang R, Nambiar R, Zheng D, Tian B. 2018. PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic acids research* **46**: D315-D319.

Wang R, Tian B. 2020. APAlyzer: a bioinformatics package for analysis of alternative polyadenylation isoforms. *Bioinformatics (Oxford, England)* **36**: 3907-3909.

Weltner J, Balboa D, Katayama S, Bespalov M, Krjutskov K, Jouhilahti EM, Trokovic R, Kere J, Otonkoski T. 2018. Human pluripotent reprogramming with CRISPR activators. *Nat Commun* **9**: 2643.

Woo YM, Kwak Y, Namkoong S, Kristjansdottir K, Lee SH, Lee JH, Kwak H. 2018. TED-Seq Identifies the Dynamics of Poly(A) Length during ER Stress. *Cell Rep* **24**: 3630-3641 e3637.

Yang Q, Liu X, Zhou T, Cook J, Nguyen K, Bai X. 2016. RNA polymerase II pausing modulates hematopoietic stem cell emergence in zebrafish. *Blood* **128**: 1701-1710.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zheng D, Liu X, Tian B. 2016. 3'READS+, a sensitive and accurate method for 3' end sequencing of polyadenylated RNA. *RNA* **22**: 1631-1639.

Supplemental Table 2.

ERCC-Spike-in Mix Preparation for TT-seq

I. Primer Design for PCR-amplifying ERCC-Spike-ins 2, 43, 92, 136, 145, 170 Spike in #	**Forward Primer (with T7 promoter)** 5′ → 3′	**Reverse Primer** 3′ → 5′	**Reverse Primer** 5′ → 3′ reverse compl.
ERCC-00043 (Mix 2)	TAATACGACTCACTATAGGGAATACC TTTACAAATGCTTTAAC	ACAAGATGGGTTAAAAAAAAAA AAAAAAAAAAAAAA	TTTTTTTTTTTTTTTTTTTTTTTTA ACCCATCTTGT
ERCC-00170 (Mix 1)	TAATACGACTCACTATAGGGTATTGG TGGAGGGGCACAAG	ATGTCTTAGGTTAAAAAAAAAAA AAAAAAAAAAAAA	TTTTTTTTTTTTTTTTTTTTTTTTA ACCTAAGACAT
ERCC-00136 (Mix 1)	TAATACGACTCACTATAGGGTTTCGA CGTTTTGAAGGAG	GATTTTCCCGGGTACAAAAAAA AAAAAAAAAAAAAAA	TTTTTTTTTTTTTTTTTTTTTTTGTA CCCGGGAAAATC
ERCC-00145 (Mix 2)	TAATACGACTCACTATAGGG ACTGTCCTTTCATCCATAAG	CGGCGTGCGAATTAAAAAAAAA AAAAAAAAAAAAAAAAA	TTTTTTTTTTTTTTTTTTTTTTTTT AATTCGCACGCCG
ERCC-00092 (Mix 1)	TAATACGACTCACTATAGGGAGATG TATATATGATGTC	CTTTAAGCCGTGGAAAAAAAAA AAAAAAAAAAAAAAA	TTTTTTTTTTTTTTTTTTTTTTTCC ACGGCTTAAAG
ERCC-00002 (Mix 2)	TAATACGACTCACTATAGGGTCCAG ATTACTTCCATTTC	GCGTTTTACCCTTAAAAAAAAAA AAAAAAAAAAAAAA	TTTTTTTTTTTTTTTTTTTTTTTTA AGGGTAAAACGC

The red part of the sequence corresponds to the T7 promoter sequence. The green parts correspond to the 5’ or 3’ end of the Spike-in sequence. The blue parts correspond to the Poly-A-Tail at the 3’ end of the RNA-sequence.