Nucleic Acids Research, 2025, **53**, gkaf240
https://doi.org/10.1093/nar/gkaf240
Methods

True length of diverse capped RNA sequencing (TLDR-seq): 5′–3′-end sequencing of capped RNAs regardless of 3′-end status

Jamie Auxillos[1,2,*,†], Arnaud Stigliani[1,2,†], Christian Skov Vaagensø[1,2], William Garland[3], Adnan Muhammed Niazi[4], Eivind Valen[4,5], Torben Heick Jensen[3], Albin Sandelin[1,2,*]

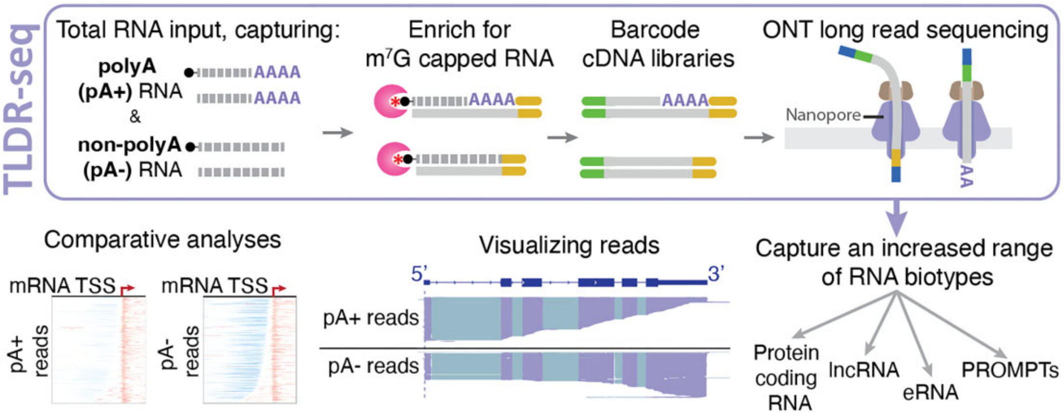
[1]Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, DK2200 Copenhagen, Denmark
[2]Biotech Research and Innovation Centre, University of Copenhagen, DK2200 Copenhagen, Denmark
[3]Department of Molecular Biology and Genetics, Aarhus University, DK8000 Aarhus, Denmark
[4]Computational Biology Unit, Department of Informatics, University of Bergen, N-5008 Bergen, Norway
[5]Department of Biosciences, University of Oslo, N-0371 Oslo, Norway

[*]To whom correspondence should be addressed. Email: jamie.auxillos@bio.ku.dk
Correspondence may also be addressed to Albin Sandelin. Email: albin@bio.ku.dk
[†]The first two authors should be regarded as Joint First Authors.

Abstract

Analysis of transcript function is greatly aided by knowledge of the full-length RNA sequence. New long-read sequencing enabled by Oxford Nanopore and PacBio devices have the potential to provide full-length transcript information; however, standard methods still lack the ability to capture true RNA 5′ ends and select for polyadenylated (pA+) transcripts only. Here, we present a method that, by utilizing cap trapping and 3′-end adapter ligation, sequences transcripts between their exact 5′ and 3′ ends regardless of polyadenylation status and without the need for ribosomal RNA depletion, with the ability to characterize polyadenylation length of RNAs, if any. The method shows high reproducibility, can faithfully detect 5′ ends, 3′ ends and splice junctions, and produces gene-expression estimates that are highly correlated to those of short-read sequencing techniques. We also demonstrate that the method can detect and sequence full-length nonadenylated (pA−) RNAs, including long noncoding RNAs, promoter upstream transcripts, and enhancer RNAs, and present cases where pA+ and pA− RNAs show preferences for different but closely located transcription start sites. Our method is therefore useful for the characterization of diverse capped RNA species and analysis of relationships between transcription initiation, termination, and RNA processing.

Graphical abstract



Introduction

Cells of higher eukaryotes produce a large and diverse population of RNA polymerase II- (RNAPII)-transcribed RNAs, ranging from protein-coding messenger RNAs (mRNAs) to long noncoding RNAs (lncRNAs) and short noncoding RNAs of many different classes and functions. Moreover, at the mRNA level, a high diversity of isoforms commonly arises from the same genomic locus [1]. Such isoform diversity is

Received: October 22, 2024. **Revised:** February 20, 2025. **Editorial Decision:** March 11, 2025. **Accepted:** March 14, 2025

driven by at least three processes: (i) alternative transcription initiation (alternative promoters), (ii) alternative splicing, and (iii) alternative transcription termination, all of which may change mRNA coding properties due to the inclusion or exclusion of functional untranslated regions (UTRs) or translated regions [2–4]. Aside from mRNAs, large proportions of the genome are transcribed into lncRNAs of diverse classes [5].

RNAPII-transcribed RNAs are chemically modified: 5′ ends are m^7G capped, which protects against 5′–3′ exonucleases, while 3′ ends may be polyadenylated (pA+), which serves several functions, including protection from 3′–5′ exonucleases, nuclear export, and cytoplasmic translation (reviewed in [6]). Mature mRNAs are typically pA+, while many lncRNAs are not polyadenylated at their 3′ end (pA−). Moreover, the length of pA tails may differ between otherwise identical transcripts, and can be related to RNA stability, export competence, and translation efficacy (reviewed in [7]). Thus, to examine the function and fate of specific transcripts, it is highly advantageous to know their full sequences, including 5′ ends, splicing patterns, 3′ ends, and the number of nontemplate adenosines. While computational methods can predict the existence and expression of specific transcript isoforms from short-read data, through assembly-based methods (reviewed in [8]) and genome or transcript reference mapping (eg. [9–11]), it is challenging to infer the exact nature of a whole transcript. Moreover, short-read-based methods have method-specific drawbacks: the arguably most-used short-read expression method, RNA sequencing (RNA-seq), has difficulties detecting precise RNA 5′ and 3′ ends, while related methods specialized in sequencing 5′ or 3′ termini (e.g. Cap Analysis of Gene Expression (CAGE) [12] and QuantSeq [13], respectively) cannot sequence the entire body of the transcript.

A new generation of sequencing methods allows for the sequencing of long complementary DNAs (cDNAs) or RNAs (reviewed in [14]), e.g. Oxford Nanopore Technologies (ONT), theoretically enabling end-to-end sequencing of transcripts. However, standard ONT cDNA-sequencing protocols have limitations in terms of 5′- and 3′-end detection. First, 5′ ends of transcripts are often truncated due to the limited processivity of the involved reverse transcriptase (RT) enzymes and the nontemplated addition of bases deriving from template switching [15]. Hence, in any standard cDNA approach, the true transcript 5′ end may not be detected. Second, because the standard ONT cDNA library preparation protocol relies on poly-deoxythymidine (polydT) adapters, largely used to avoid the abundant pA− transfer RNAs and ribosomal RNAs, only pA+ transcripts are captured. Thus, pA− transcripts, including most lncRNAs, will not be sequenced.

To circumvent these issues, a number of ONT-based protocols have been proposed. These can broadly be divided into (i) protocols that aim to locate accurate 5′ ends (through cap trapping, e.g. CapTrap-seq [16] or template switching, e.g. long read CAGE [17]) but rely on pA+ selection, and (ii) protocols that are not reliant on pA+ RNA 3′ ends but do not aim to accurately capture 5′ ends (e.g Nano3P-seq [18]). A small subset of such methods is based on direct RNA-seq, e.g. TERA-Seq [19] and ReCappable-seq [20], but overall produces substantially fewer sequence reads than the cDNA-based methods. To our knowledge, there are no methods that aim at accurate 5′-end detection of capped, RNAPII-transcribed RNAs regardless of 3′-end polyadenylation status.

To this end, we present true length of diverse capped RNA sequencing (TLDR-seq), a method that uses cap trapping to select for capped RNAs, which then are full-length sequenced 5′–3′ regardless of pA status. We show that TLDR-seq faithfully captures true 5′ and 3′ ends, that it can sequence pA+ and pA− transcripts, including diverse lncRNAs and that expression estimates of the data at the gene level are highly correlated with that of short-read methods. TLDR-seq represents a useful complement to other ONT cDNA-based methods as it can be used to characterize full-length pA+ and pA− transcripts in the same experiment.

Materials and methods

Cell lines, depletions, and input RNA sample preparation

HeLa S3 cells (gift from Heick Laboratory, Aarhus University) were cultured in Dulbecco's Modified Eagle's Medium (DMEM) and high glucose and pyruvate (Thermo Fisher, #41966029), supplemented with 10% (v/v) fetal bovine serum (Thermo Fisher, #A3840401) and 1% (v/v) penicillin–streptomycin (Thermo Fisher, #15140122), at 37°C in a humidified incubator with 5% CO_2, and passaged when ∼80%–90% confluent.

HeLa cells were lysed and the total RNA was extracted using the TRIzol Plus RNA Purification Kit and Phasemaker Tubes Complete System (Thermo Fisher, #A33254), DNaseI treated (Thermo Fisher, #EN0523), and repurified using the Purelink RNA Mini Kit (Thermo Fisher, #12183025). RNA quality was assessed using the RNA Nano Bioanalyzer Kit (Agilent, #5067-1511).

All mouse embryonic stem (mES) cell lines used or generated in this study were descendants of wild-type (WT) ES-E14TG2a cells (male genotype, XY). mES cells were cultured on 0.2% gelatin-coated plates in 2i/LIF-containing media [1:1 mix of DMEM/F12 (Thermo Fisher, #31331-028) and Neurobasal (Thermo Fisher, #12348-017) supplemented with 1% (v/v) penicillin–streptomycin (Sigma, #P4333), 2 μM GlutaMAX (Thermo Fisher, #35050-038), 0.1 mM non-essential amino acids (Thermo Fisher, #11140035), 1 mM sodium pyruvate (Thermo Fisher, #11360-039), 0.5× N-2 Supplement (Thermo Fisher, #17502-048), 0.5× B-27 Supplement (Thermo Fisher, #17504-044), 50 μM 2-mercaptoethanol (Thermo Fisher, #31350-038), 3 μM GSK3-inhibitor (CHIR99021, Sigma, #SML1046), 1 μM MEK-inhibitor (PD0325901, Sigma, #PZ0162), and leukemia inhibitory factor (LIF; produced in house)] at 37°C and 5% CO_2. Cells were passaged every 48–72 h by dissociating cells with 0.05% trypsin (Sigma, #P4333) before neutralizing with an equal volume of 1× trypsin inhibitor (Thermo Fisher, #17075-029). Cells were pelleted by centrifugation to remove trypsin before resuspending in 2i/LIF media and plating at $\sim 8 \times 10^4$ cells/ml.

The generation of CRISPR/Cas9-mediated genomic knock-ins of N-terminal 2×HA-FKBP-V (HA-dTAG)-RBM7 mES cells was described in [21] and HA-dTAG-ZCCHC8 cells were generated in a similar way. Depletions of HA-dTAG tagged proteins was performed by the addition of $dTAG^V$-1 (Tocris, #6914) to the cell culture medium for 4 h at a concentration of 100 nM.

Total RNA from WT, HA-dTAG-RBM7, and HA-dTAG-ZCCHC8 mES cells was isolated using TRIzol (Invitrogen, #15596026) according to the manufacturer's instructions and treated with TURBO DNase (Thermo, #AM2239). RNA quality was assessed using the RNA Nano Bioanalyzer Kit (Agilent, #5067-1511).

TLDR-seq and nanopore library preparation

All oligonucleotides used in this study were purchased from Integrated DNA Technologies. We used 8 μg total RNA per sample as starting material, and dephosphorylated using shrimp alkaline phosphatase (rSAP, NEB, #M0371S), and based on the optimal 3′-dephosphorylation enzyme for 3′-end adapter ligation from [22] and following the manufacturer's recommendations. Dephosphorylated RNA was purified using the Purelink RNA Mini Kit and quantified using the NanoDrop One Spectrophotometer (Thermo Fisher) to calculate the volume for 5 μg of dephosphorylated RNA to use for subsequent steps. Briefly, 95.5 pmol of adenylated A1 single-stranded DNA adapter (Supplementary Table S1) was ligated to the 3′ end of dephosphorylated RNA (5 μg) with 10× RNA ligase buffer, 20% polyethylene glycol-8000 (final concentration), 800 units of T4 RNA ligase 2-truncated KQ (NEB, #M0373L), and 16 units of RNaseOUT recombinant ribonuclease inhibitor (Thermo Fisher, 10777019) for 4 h at 25°C. Unligated A1 adapters were deadenylated and degraded using 25 units of 5′ deadenylase (NEB, #M0331S) and 15 units of RecJ$_f$ (NEB, #M0264S) at 30°C for 30 min then purified with RNAClean XP Beads (Beckman Coulter, #A63987). A total of 2 μl of 100 μM A1_rc RT primer (Supplementary Table S1) was added to the ligated RNA then denatured at 65°C for 5 min then placed on ice. cDNA was generated by adding 5× Induro buffer, 2 μl 10 mM dNTP mixture (GenScript, #D0056), 20 units of RNaseOUT recombinant ribonuclease inhibitor, and 800 units of Induro Reverse Transcriptase (NEB, #M0681L) to the RNA and A1_rc mixture and incubated for 2 h at 55°C then purified using RNAClean XP beads. The standard cap-trapping protocol was carried out as described in [23], with the following changes: (i) following RNase ONE (Promega) and Ribonuclease H (Takara, #2150B) digestion steps, the barcoded 5′-linker ligation was replaced with our barcoded A2 adapter ligation (Supplementary Table S1), with each library barcoded with a unique A2 double stranded adapter. The A2 adapter annealing and ligation to the cDNA library were following the same reaction conditions as in [23], (ii) the 3′-linker ligation and USER treatment were completely removed due to the A1 adapter ligation carried out prior to cap trapping, (iii) second-strand cDNA synthesis was carried out using LongAmp Hot-Start Taq 2× Master Mix (NEB, #M0533L) with 2 μl of 20 μM A2 second-strand primer (Supplementary Table S1) and incubated on a thermocycler with the following program: 95°C for 5 min, 60°C for 5 min, 65°C for 30 min, and then hold at 4°C. Excess primers were degraded with 20 units of exonuclease I (NEB, #M0293L) at 37°C for 30 min and then purified by AMPure XP Beads (1.8× bead amount) (Beckman Coulter, #A63881). Of note, as in the nAnT-iCAGE method [23], we incorporated an RNAse I (Promega, #M4261) treatment after reverse transcription but prior to cap trapping to cleave single-stranded RNAs which have undergone incomplete reverse transcription (i.e. not copied till the 5′ end of the RNAs) resulting in truncated cDNAs.

The reaction was purified a second time with a lower AMPure XP Bead amount (1.4×) followed by a vacuum centrifugation using the Savant SpeedVac DNA130 (Thermo, #DNA130-230) to concentrate each library to 10 μl. Preparations of nonamplified libraries are completed at this point and are assessed for fragment-length distribution using a High Sensitivity DNA Bioanalyzer Chip (Agilent, #5067-4626) and concentration using the Qubit dsDNA High Sensitivity Kit (Thermo Fisher, #Q32851). For the nonamplified HeLa library, each barcoded sample was pooled then 180 fmol was used as an input for the LSK110 nanopore library preparation. Endprep/dA-tailing and nanopore adapter ligation (Oxford Nanopore Technologies, #SQK-LSK110) was carried out according to the manufacturer's protocol, loaded onto an R9 MinION flow cell (FLO-MIN106 - R9.4.1), and ran for 72 h.

For amplified libraries, following the exonuclease I digestion and first AMPure XP Bead purification, each library was split into four tubes, each with 10 μl, mixed with 2 μl of the appropriate 10 μM A2 amp primer (Supplementary Table S1), 2 μl of 10 μM A1 amp primer, LongAmp HotStart Taq 2× Master Mix, and nuclease free water. Libraries were amplified using the following program on a thermocycler: 95°C for 30 s, 9–10 cycles of 95°C for 15 s, 60°C for 15 s, and then 65°C for 8 min, followed by 65°C for 8 min and hold at 4°C. Each amplified library was exonuclease treated, purified with AMPure XP Beads, and concentrated as previously described for nonamplified libraries. The four amplified reactions for each library were recombined into one tube prior to fragment length and concentration assessment using the Bioanalyzer and Qubit, respectively. Each barcoded HeLa-amplified sample was pooled at an equimolar ratio then 180 fmol was used as an input for the SQK-LSK110 endprep/dA-tailing and nanopore adapter ligation, carried out according to the manufacturer's protocol, loaded onto a R9 MinION flow cell (FLO-MIN106D - R9.4.1), and ran for 72 h. In parallel, these three HeLa-amplified libraries were also pooled at an equimolar ratio then 180 fmol was used as an input for the LSK114 end-prep/dA-tailing and nanopore adapter ligation (Oxford Nanopore Technologies, #SQK-LSK114) according to the manufacturer's protocol, for subsequent sequencing on an R10 MinION flow cell (FLO-MIN114 - R10.4.1) for 72 h. mES-amplified libraries contained the same A2 barcodes and were therefore not pooled and instead, 180 fmol of each library was individually processed for endprep/dA-tailing and ligated with nanopore adapters using the SQK-LSK110 kit. These mES-amplified libraries were run sequentially on two R9 MinION flow cells, (FLO-MIN106D - R9.4.1), each library was run for 24 h, with flow cells washed using the Flow Cell Wash Kit (Oxford Nanopore Technologies, EXP-WSH004) after each library.

Sequencing and base calling of the TLDR-seq libraries sequenced on the R9 flow cells

The HeLa-amplified and -unamplified libraries that were sequenced for 72 h (HeLa libraries) or 24 h (mES libraries) with the R9.4.1 MinION flow cells (FLO-MIN106D), using the software MinKNOW v22.10.7. Reads were basecalled with Guppy v6.4.2 using the parameters "–trim_strategy none", "–compress_fastq", and "–fast5_out". The last parameter allows for obtaining fast5 files with the move table, needed to determine polyadenylic acid tail (polyA tail) lengths.

Sequencing and base calling of the TLDR-seq libraries sequenced using the R10 flow cells

The same HeLa-amplified TLDR-seq libraries were sequenced on the R9 flow cell (described above) were sequenced for

72 h with the R10.4.1 MinION flow cell (FLO-MIN114), using the software MinKNOW v22.10.7. Reads were basecalled with dorado 0.7.1 using "–no-trim", "–estimate-poly-a", and "sup". For more information about the polyA tail length estimation, see the "polyA tail length estimation" section below.

QC and trimming of the TLDR-seq libraries

Following basecalling, we kept reads with a "mean_qscore_template" >9 for further analysis. As reads have 5′ and 3′ adapters ($A2_i$, i ∈ {1, 2, …, 12} and A1rc, respectively), $A2_i$ was used to demultiplex libraries. Cutadapt v6.3.7 was run successively with "-g $A2_i$.A1rc" and "-g A1.$A2_i$rc" where A1rc and $A2_i$rc are, respectively, A1 and $A2_i$ reverse complement sequences. With this setup, only reads with both A1rc and $A2_i$ adapters (or their reverse complement) are not filtered out. Besides these parameters, cutadapt was used with folowing options:

"-O 8" for a minimum overlap of eight nucleotides between the read and our custom barcodes. If that said overlap was not met, the read was discarded.

"-m 1" to discard any 0-nt long read after trimming.

Genome mapping of TLDR-seq and standard ONT HeLa libraries

TLDR-seq reads with 5′ and 3′ adapters were mapped on the GRCh38 genome (GENCODE release 21) using minimap2 v2.24-r1122 with options "-ax splice", "-un", and "—junc-bed anno.bed". The file "anno.bed" was generated with the program "paftools gff2bed" and the GENCODE v39 gtf file. For multimapped reads, only the best alignment was kept. As the nanopore sequencing method outputs reads on both strands, we wrote an additional script using cutadapt output (see the "QC and trimming" section in the "Materials and methods" section) to orient reads in the direction of their corresponding transcript.

Standard ONT nanopore reads were downloaded from the ENA dataset PRJEB44747 [24] and mapped with minimap2 using the same parameters. However, the same orientation method could not be used as adapters were often too short or not captured by the sequencing. Instead, we mapped the reads to a transcriptome (see next paragraph) and deduced their orientation using their best alignment. If several best alignments were found, we selected one of these at random.

Genome mapping of TLDR-seq mES cell libraries

Reads were mapped with minimap2 v2.24-r1122 on the mm39 genome (GENCODE release 35) using the same methodology and parameters used for mapping the HeLa cell reads. The "–junc-bed anno.bed" file was generated from the GENCODE v35 gtf file.

Transcript and gene quantification of TLDR-seq and standard ONT nanopore libraries

The bambu function from the Bambu package (v3.5.1) [25] was used with the option "discovery = FALSE" to quantify the transcript expression levels from the mapped reads and according to the GENCODE v39 human annotation. When comparing different libraries, we used Bambu Tags per Million (TPM) output (as opposed to the TPM calculation in RNA-seq, these are not normalized by transcript length). For comparisons of expression level on gene level transcripts were grouped by genes and their TPM were summarized as gene expression (see the "Gene expression level comparisons" section).

Mapping of QuantSeq libraries

Fastq files were downloaded from the GEO dataset GSE137612. Reads were mapped to the GRCh38 genome with STAR v2.7.9a [26] (parameters –twopassMode Basic –outSAMmapqUnique 60) and their direction was reversed as the protocol produces stranded antisense transcript reads.

Mapping of SLIC-CAGE libraries

HeLa control [treated with a small interfering RNA (siRNA)-targeting green fluorescent protein (GFP)] SLIC-CAGE fastq files were downloaded from the GEO dataset GSE147655, mapped on the GRCh38 genome, and processed as in [27].

Mapping and quantification of RNA-seq libraries

HeLa control (treated with an siRNA-targeting GFP) RNA-seq paired-end fastq files were downloaded from the GEO dataset GSE84172 and mapped on the GRCh38 with STAR v2.7.9a [26] (parameters –twopassMode Basic –outSAMmapqUnique 60). Additionally, we quantified the transcript expression levels using the GENCODE v39 human transcriptome and Salmon v1.10.2 [9] in quantification mode with the flags -l lU –allowDovetail and –validateMappings. Transcript expression levels were summarized to gene expression levels by summing TPM values for all the isoforms of the same gene.

Processing of ChIP-seq and DNase-seq from mES cells

mES cell Deoxyribonuclease sequencing (DNase-seq) data were downloaded from the ENCODE portal (datasets ENCFF672DJH and ENCFF962TCT) [28]. H3K4me3, H3K4me1, and H3K27 ChIP data were downloaded from the GEO GSE137491 dataset [29]. All the downloaded data were already in the bigWig format. Files were normalized with the RSeQC [30] (v5.0.1) normalize_bigwig.py tool and coordinates were converted from mm10 to mm39 assembly using the lifOver tool [31]. After coordinate conversion, the signal was partitioned with the bedops –partition [32, 33] and averaged over overlapping coordinates with the bedmap –mean function. The two DNase-seq replicates were concatenated, partitioned, and averaged in a similar fashion. These bedgraph files were then converted to the bigWig format with the bedGraphToBigWig program [32]. DNaseI hypersensitive sites peaks were obtained from the ENCODE portal (dataset ENCFF048DWN) [28] and coordinates were converted from mm10 to mm39 using the liftOver tool.

5′ end, 3′ end, and splice sites comparisons

Replicates of each sequencing method were pooled together. For each method, we defined a transcription start site (TSS), if there were at least n reads with the same orientation starting at a given genomic location. This threshold n was determined by (number of reads in the pooled library)/10^6 ensuring that the number of TSS per method was independent of sequencing depth. Additionally, because SLIC-CAGE read processing includes 1-nt trimming on the 5′ end, we shifted the standard ONT nanopore and TLDR-seq read start by 1 nt toward the read direction. When comparing TSS between two methods, we defined a foreground and a background: for each TSS in the background, we assigned the closest TSS in the foreground with the same orientation. If a given background TSS has two equally distant foreground TSS (one upstream and one downstream), we assign one of them randomly. The measured distance was defined as negative if the foreground TSS was downstream (TSS orientation wise) of the foreground TSS and positive if it was upstream. Associations whose distance $d > |50|$ nt were discarded.

The same method was used to compare 3′ end or acceptor and donor splice site locations between sequencing methods except that no nucleotide shift was performed.

To define GENCODE TSSs, transcription termination sites (TTSs), and acceptor or donor splice sites, we used either a subsetted version of the Gencode v39 annotation (list of biotypes in Supplementary Table S2) or only protein-coding-annotated transcripts. Gencode TSSs were not shifted by 1 nt. When comparing Gencode ends (TSSs, TTSs, and acceptor or donor splice site) with those from other sequencing methods, we used the same methodology as when comparing between different sequencing methods.

Motif and nucleotide content at the 5′ ends, 3′ ends, and splice site locations

Respective sites (TSSs, TTSs, and acceptor or donor splice sites) were given a score equal to the number of reads supporting them. We filtered the sites with the highest score within a sliding window of 40 nt and kept the 750 top sites among all the filtered ones. Nucleotide frequencies were calculated in a [−50, 50]-nt window centered around the sites and normalized by a factor so that the average normalized frequency value of each nucleotide within this window is equal to 1. To plot the logos, we calculated a scaled frequency value $freq_{\text{scaled}}$ as $freq_{\text{scaled}}\ (N)_i = \frac{freq_{\text{norm}}\ (N)_i}{\sum freq_{\text{norm}}\ (M)_{i,M=\{A,C,G,T\}}}$, where N is either A, C, G, or T. This scaled frequency was then used to calculate the information content of the sequence logo [34].

Reproducibility of TLDR-seq libraries and comparisons between amplified/nonamplified libraries

We mapped and quantified amplified and nonamplified libraries according to the "Materials and methods" section "Transcript and gene quantification of TLDR-seq and standard ONT nanopore libraries". We used transcript TPM values, not summarized gene TPM expression estimates. To assess the reproducibility within and across amplified and nonamplified libraries, we either calculated the TPM values from pooled or demultiplexed amplified and nonamplified libraries.

To not limit the comparisons to transcripts, we also compared the coverage of the libraries on 500-bp genomic bins covering the whole genome. To this end, reads were aligned according to the "Genome mapping of TLDR-seq and standard ONT HeLa libraries" section and the coverage (number of reads mapping at a genomic position) was quantified using the SAMtools [35], the BEDTools [36], and the BEDOPS [33]. We summed the coverage over 500-bp bins and normalized this value in each library by its respective total sequencing depth. For plotting, we filtered regions that had at least, on average, one read per bin. As for the transcript expression levels, we used coverage values from the pooled or demultiplexed libraries.

polyA tail length estimation

For the reads derived from the R9 flow cell, FAST5 files including the move table were demultiplexed with the ont_fast5_api tools, according to cutadapt output. The R-package tailfindr [37] (version: tldr-seq_v1.4; branch: https://github.com/adnaniazi/tailfindr/tree/tldr-seq) was used to estimate polyA/T tail length. The standard tailfindr algorithm (as detailed in reference [37]) identifies adapter sequences adjacent to polyA or polyT tails and operates under the assumption that a polyA or polyT tail is always present next to the adapter. This assumption aligns with the standard ONT sequencing protocol, which typically targets polyA+ RNAs for sequencing. However, in the tldr-seq branch, this assumption has been relaxed to accommodate reads that lack a polyA tail. This modification allows the algorithm to process a broader range of RNA sequences, including pA− transcripts. Reads were classified in two groups: with polyA tails (pA+) if their estimated polyA tail was longer than 15 nt, without polyA tail (pA−) otherwise. This cutoff was motivated by a previous use in the literature [38], but we note that a cutoff of >0 produced similar results in terms of types of transcripts detected (Supplementary Fig. S1A).

For reads derived from the R10 flow-cell, polyA estimation code was rewritten so that it implements the tailfindr algorithm for R10 flow cells (https://github.com/adnaniazi/tailfindr/tree/nano3p-seq-r10), while accounting for our adapters and library design. Reads were classified in two groups: with polyA tails (pA+) if their estimated polyA tail was longer than 15 nt, without polyA tail (pA−) otherwise.

Gene expression level comparisons

Comparison of the gene expression levels between TLDR-seq and standard ONT nanopore libraries was done after we quantified TPM gene expression level following the "Materials and methods" section "Transcript and gene quantification of TLDR-seq and standard ONT nanopore libraries". For this specific comparison, only TLDR-seq reads containing a polyA/T tail were considered (see previous paragraph).

When comparing RNA-seq and long-read sequencing methods, gene TPM levels of the RNA-seq libraries were obtained according to the "Materials and methods" section "Mapping and quantification of RNA-seq libraries". As Salmon collapses all groups of identical transcripts into a single transcript, we decided to use the same methodology for the long-read sequencing methods before calculating the gene TPMs, from Bambu transcript quantification. Bambu isoform counts were then summed and normalized for each gene so that we compare them with Salmon TPM.

To compare gene expression between SLIC-CAGE libraries and long-read sequencing methods, we first mapped each library (see the "Mapping of SLIC-CAGE libraries" and "Mapping to genome of TLDR-seq and standard ONT nanopore libraries" sections). SLIC-CAGE, ONT nanopore, and TLDR-seq reads were then trimmed to retain only the first nucleotide at their 5′ ends. Using the *CAGEfightR* package [38], we clustered 5′ ends on the same strand into Tag Clusters (TCs). Each TC was assigned to the nearest gene within 100 bp. Gene expression levels, measured as TPM, were calculated by summing TC counts associated with each gene and normalizing them by library size.

Analysis of associations between alternative TSS usage and RNA polyadenylation

The analysis was performed using pooled amplified R9 TLDR-seq HeLa libraries. Reads were mapped and trimmed to retain only the 5′ terminal nucleotide. They were then classified based on polyA tail status (pA+ or pA−; see the "polyA tail length estimation" section in "Materials and methods" section). To examine the distribution of 5′ ends in pA+ and pA− reads, we used BEDTools to identify reads overlapping the first exon of each gene on the same strand, as defined by GENCODE annotation, and extended them 100 bp upstream and downstream. Since many genes encode multiple isoforms, we selected the first exon of the most highly expressed isoform to avoid multiple instances of 5′-end overlaps (see the "Transcript and gene quantification of TLDR-seq and standard ONT nanopore libraries" section in "Materials and methods" section). Exons with fewer than 10 pA+ or 10 pA− reads were filtered out.

Additionally, since cap-trapping methods can capture 3′ UTR transcript endocleavage byproducts [54], and our focus was on transcription initiation, we excluded mono-exonic transcripts to avoid confounding 5′ ends from both transcription initiation and endocleavage products. Within each promoter region (defined as ± 100 bp around the selected first exons), we compared the distribution of pA+ and pA− 5′ ends using a Wilcoxon test with False Discovery Rate (FDR) correction ($FDR < 0.05$).

For all ribosomal-gene promoters passing this threshold, we constructed core promoter motifs per polyA category. To do so, in each given promoter and for each read polyA status, we selected the TSS with the highest number of reads, along with all the TSS with at least 40% this number. We then used the ±15-bp sequences around these TSSs to calculate the scaled nucleotide frequency per polyA status (see the "Motif and nucleotide content at the 5′ ends, 3′ ends, and splice site locations" section in "Materials and methods" section) and to generate the motifs.

For all promoters passing the $FDR < 0.05$ threshold (not just those associated with ribosomal genes), we calculated the entropy of pA+ and pA− 5′-end distributions. Since entropy calculations are sensitive to sample size and most promoters exhibit an imbalance in pA+/pA− read counts, we downsampled the number of 5′ ends per promoter to match the lower read count in either the pA+ or pA− fraction. This process was repeated 1000 times using bootstrapping, and the average entropy was computed for both fractions per promoter. To assess differences in entropy between the two read fractions, we performed a paired Wilcoxon test.

To compare genomic length, read length, and exon count across polyA statuses, we grouped reads by promoter (if their 5′ end overlapped a given promoter) and by polyA tail status. We then calculated the average genomic length, read length, and exon count per group.

To investigate cases where alternative promoters of the same gene produce differing ratios of pA+ and pA− reads, we reanalyzed reads from the pooled amplified TLDR-seq HeLa libraries. Reads were mapped, trimmed to retain only their 5′ terminal nucleotide, and separated by polyA status. Using the CAGEfightR package [39], we clustered 5′ ends on the same strand into TCs. Since TCs were often broad, we decomposed them into local maxima if these were separated by >30 nt. TCs with <10 reads in either polyA fraction were discarded, and we calculated the $\log_2$ fold change [$\log_2$ {counts (pA+)/counts (pA−)}] for each TC. TCs were classified into three categories: enriched in pA+ ($\log_2 FC > 1$), enriched in pA− ($\log_2 FC < -1$), and nonenriched ($|\log_2 FC| < 0.3$). We then identified genes containing at least two TCs assigned to different categories. TC annotation was performed using CAGEfightR, and a TC was annotated as a TSS if it overlapped a [−50, 200] region around a GENCODE-defined TSS.

PROMPT and enhancer RNA characterization

For promoter upstream transcript (PROMPT) detection, we first pooled mES-RBM7 and mES-ZCCHC8 polyA+ reads. We kept TLDR-seq TSSs supported by at least 10 reads and overlapping with a GENCODE mRNA (mm39, v35). When several TSS appeared in a 3000-bp window, we only kept the most upstream one. We then calculated the coverage of TLDR-seq pA+ and pA− reads around each such TSS, for each cell treatment and flipped the strand orientation when the mRNA TSS were detected on the reverse strand. We assigned negative coverage scores to reads that were oriented opposite to the TSS direction. DNase1-seq (see the "Materials and methods" section, "Processing of ChIP-seq and DNase-seq from mES cells") coverage was then calculated around TSSs, and mRNA TSSs which did not have additional upstream TSSs on the other strand in the −1 to −400-bp region from the mRNA TSSs were filtered out.

Putative enhancer regions were detected starting from intergenic or intronic DNAse hypersensitive site (DHS) peaks based on GENCODE annotation, which we required to overlap at least one TLDR-seq read TSS from mES-RBM7 and mES-ZCCHC8 depletions (pA+ or pA−) TSS on each strand oriented bidirectionally and not overlapping each other, and not separated by >600 bp. When several such enhancer TSSs were detected in a 1000-bp window, we took the most expressed one. Putative enhancer midpoints were defined as the midpoint between the bidirectional TSSs as defined above.

Results

TLDR-seq overview

Briefly, TLDR-seq, described in detail in the "Materials and methods" section above, is a novel nanopore-based method, expanding from ONT cDNA protocols [such as direct cDNA and cDNA-polymerase chain reaction (PCR)], with three key features (Fig. 1A). First, to ensure that true 5′ ends are captured, we employ cap trapping, based on the nAnT-iCAGE and ssCAGE protocols [23, 40], to enrich for m^7G-capped RNA species and selecting against for RNAs with prematurely truncated cDNAs due to incomplete reverse transcription.

Second, this is combined with the ligation of an extended 5′-end A2 adapter, a 58-bp so-called "stuffer" sequence, to avoid the known problem with deteriorating nanopore sequencing quality near the 5′ termini and functioning also as a library barcode to enable multiplexing.

Third, to be able to capture pA− and pA+ RNAs, we used T4 RNA ligase 2 truncated KQ enzyme to ligate 5′ adenylated A1 DNA adapter (58 bases) to all RNAs, and a primer complementary to A1 adapter was used for cDNA synthesis using Induro reverse transcriptase, a group II intron RT with high processivity [41]. T4 RNA ligase 2 truncated KQ enzyme was chosen to avoid adapter–adapter or RNA–RNA concatemerization. As with the 5′-end A2 adapter, the A1 adapter functions as a "stuffer" sequence to avoid deteriorating sequencing

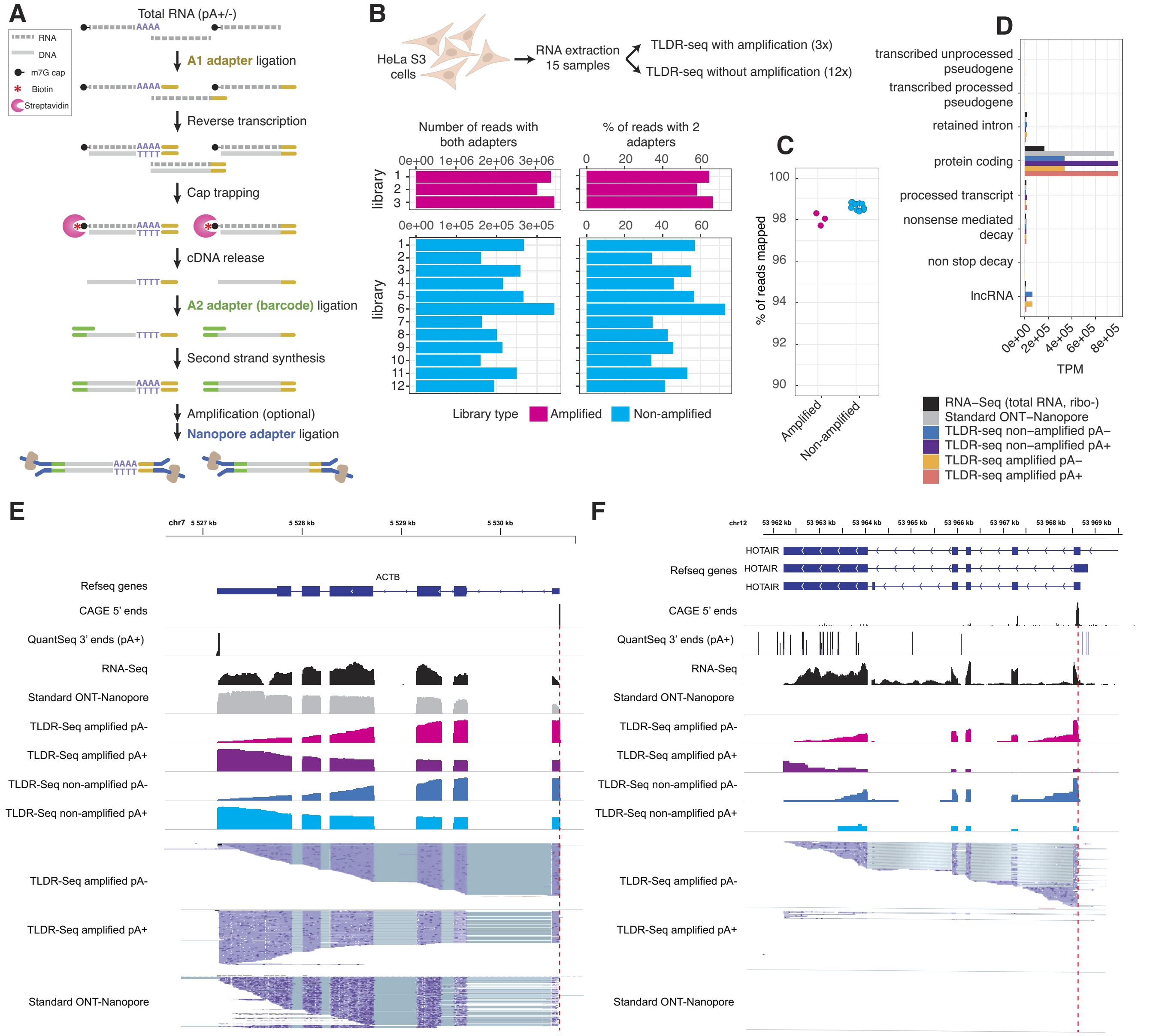


Figure 1. Overview of TLDR-seq and data examples. (**A**) Schematic overview of the TLDR method. With pA+ and pA− total RNA as an input, a 5′ adenylated A1 DNA adapter is ligated to the 3′ end of all RNAs, irrespective of polyA status. Then an oligo complementary to the A1 adapter is used to generate the cDNA by reverse transcription. To enrich for m^{7}G capped RNAs, we employ cap-trapping technology, involving the oxidation and biotinylation of the cap structure followed by pulldown with streptavidin magnetic beads (pink). The cDNA of capped enriched RNAs are released and a barcoded A2 adapter is attached to the 3′ end of the cDNA. Primers complementary to A1 and A2 adapters are used to optionally PCR amplify the library. Finally, the nanopore adapters are attached to the cDNA library using either the LSK110 or LSK114 kit (Oxford Nanopore Technologies) and sequenced on the MinION platform, on R9 or R10 flow cells. (**B**) Experimental design, sequencing depth and mapping rates. Top schematic shows the experimental design used for TLDR-seq experiments. Bar plot shows the number of reads having both adapters (left) and the percent of reads having both adapters (right). Color indicates library type (amplified or nonamplified). (**C**) Mapping rate of libraries. *Y*-axis shows % of reads mapped. *X*-axis shows library type. Dots show individual libraries. (**D**) Mapping to selected GENCODE-annotated gene annotation classes. *X*-axis shows sum of reads mapping to a given gene annotation class (*Y*-axis), TPM-normalized. Colors indicate data type. TLDR-seq reads are divided into pA+ and pA−. (**E**) ACTB genome browser screenshot. Integrative Genomics Viewer (IGV) [68]-based visualization of reads from HeLa-derived total RNA-seq, 5′ and 3′ sequencing methods, standard ONT nanopore cDNA sequencing, and TLDR-seq (split by amplification method and pA ± reads) around the ACTB gene. Refseq gene model is shown on top. Transcription is on the minus strand, right to left. Coverage tracks are shows for all data types. Below, randomly sampled reads from standard ONT nanopore cDNA sequencing and amplified TLDR-seq libraries are shown. Dotted red line shows the location of the main CAGE-defined TSS. (**F**) HOTAIR genome browser screenshot. IGV-based visualization organized as in panel (E) but showing data around the HOTAIR lncRNA locus. All reads from standard ONT nanopore and amplified TLDR-seq libraries are shown.

quality on the ends. The method includes an optional cDNA amplification step: amplified libraries will produce higher read counts, but may skew cDNA length distributions (analyzed further in the next section).

To evaluate TLDR-seq, we employed it on total RNA extracted from HeLa S3 cells (see the “Materials and methods” section), producing 15 libraries where 3 were subjected to cDNA amplification and 12 were not: the reason for sequencing more nonamplified libraries was that such libraries yielded much less cDNA: the three amplified libraries produce enough cDNAs for one ONT MinION flow cell, while for nonamplified libraries 12 libraries are necessary to reach the minimum input required as per the manufacturer’s recommendations for the SQK-LSK110 adapter ligation kit and MinION R9 flow cell loading. Using the MinION R9 flow cell, amplified libraries produced on average 5.2e6 reads, where 63% had both 5′ and 3′ adapters, indicating capture of full-length RNAs, while nonamplified libraries produced an average of 4.7e5 reads, of which 54% had both adapters (Fig. 1B).

The mapping rate of reads having both adapters were 98% (±0.4%) and 98% (±0.3%) in the amplified and the nonamplified libraries, respectively (Fig. 1C). PolyA/T tail lengths were determined with the R package tailfindr [37], using a version specifically tailored to TLDR-seq (see the “Materials and methods” section). Reads were defined as pA+ if having pA tails with >15 adenosines, pA− otherwise. For comparison, we collected ribosomal-depleted (but not pA+-selected) RNA-seq data from [42] and ONT Nanopore direct cDNA sequencing (DCS108) libraries from HeLa cells [24] (referred to as “standard ONT Nanopore” below), and measured the normalized read count overlapping a curated list of GENCODE gene annotations [43] of RNAPII-transcribed transcripts (Fig. 1D). The most striking observation was that TLDR-seq pA− reads had substantially higher number of reads falling into lncRNAs than any other approach, while TLDR-seq pA+ reads numbers were highly similar to that of standard ONT Nanopore within protein-coding genes, with no substantial difference between amplified and nonamplified libraries (further analyzed in next section). Two examples of these observations on gene level are shown as genome-browser screenshots in Fig. 1E and F for the protein-coding ACTB gene and the lncRNA HOTAIR. The same data as in Fig 1D were used, but we also added CAGE 5′-end data from [27] and QuantSeq 3′-end data from [44], both from the HeLa S3 cells (Fig. 1E and F: top panels show coverage of each data type, and lower part shows reads from TLDR-seq and standard ONT Nanopore).

For ACTB (Fig. 1E), it was clear that TLDR-seq reads captured the 5′ ends accurately (compare with the CAGE track and GENCODE annotation). The pA+ TLDR-seq reads generally covered the full gene length, albeit with a subset of reads starting in the 3′ UTR (discussed further below). Conversely, only a small subset of pA− TLDR-seq reads covered the whole gene: most terminated within the second to last exon, which may reflect degradation products, in particular prematurely terminated transcripts. Standard ONT Nanopore reads rarely covered the whole gene: only a fraction of reads started at the annotated 5′ end and a large fraction of reads started within annotated exons. The TLDR-seq 5′ ends residing in the 3′ UTR of ACTB, which were primarily observed in the pA+ data, echoes previous similar observations made using pA+-selected CAGE and RNA-seq data on many genes (e.g. [45–48]). Such 5′ ends of pA+ transcripts residing within 3′ UTRs have been suggested to be cryptic TSSs or more recently proposed to be the result of mRNA endonucleolytic cleavage followed by recapping [49]. Additional examples of lncRNAs detected with isoform resolution by TLDR-seq are shown in Supplementary Fig. S1B–E.

Conversely, for HOTAIR (Fig. 1F), only the RNA-seq and TLDR-seq pA− reads could reliably detect RNAs from the locus, likely due to the fact that most HOTAIR RNAs were pA−. Only a fraction of TLDRseq pA− reads reached the GENCODE-annotated 3′ ends, but this was also in agreement with QuantSeq data which showed dispersed 3′- ends along the annotated gene. The TLDR-seq pA− reads also detected frequent read-through of the first, and to some degree second, intron, indicative of frequent TSS-proximal transcription termination.

Library reproducibility and comparison of amplified versus nonamplified libraries

To assess reproducibility, we first compared the three amplified TLDR-seq libraries to each other, using three approaches. First, we examined read length distributions as the polymerase used in the PCR step is more prone to detachment when amplifying longer cDNA fragments. We observed almost identical read length distribution when comparing the pooled amplified and pooled nonamplified libraries (Fig. 2A shows the average read length distributions between library types, and Supplementary Fig. S2A and B shows read length distributions for individual libraries). This indicates that amplification can be performed without substantial effects of the sequencing of long cDNA fragments.

Next, we assessed the expression reproducibility between libraries using two approaches. First, we binned the genome into 500-bp bins and then counted the sequencing depth in each such bin for each library, followed by calculating Pearson’s correlation of these bin counts between pairs of libraries. In a second approach, we quantified transcript (isoform) expression levels with Bambu [25] (see the “Materials and methods” section), and used the transcript counts to assess the correlation between libraries within the same group (amplified or nonamplified). Fig. 2B shows summaries of Pearson correlation coefficients (Pearson’s *R)* from these comparisons (Supplementary Fig. S2C–F shows individual pairwise comparisons). Pairwise Pearson’s *R* values between amplified libraries were high: on average 0.92 (±0) for both the genome binning and the isoform quantification based method. We obtained slightly lower correlations when comparing nonamplified libraries, with average Pearson’s *R* values of 0.85 (±0.03) and 0.70 (±0.04) for the genome binning and isoform quantification approach, respectively. The slightly lower correlations for nonamplified libraries may be due to the substantially lower sequencing depth in the individual libraries compared with amplified libraries.

Next, we used the same approaches to compare amplified with nonamplified libraries. Because of the large differences in sequencing depth in individual libraries coming from respective groups (discussed above), for the genome binning method, we normalized read counts by the total sequencing depth in each library and for the isoform quantification method, we used TPM-normalized expression values.

Using these normalized values, we compared (i) all individual amplified libraries to all nonamplified libraries individually (Supplementary Fig. S2G and H) and (ii) the pooled reads

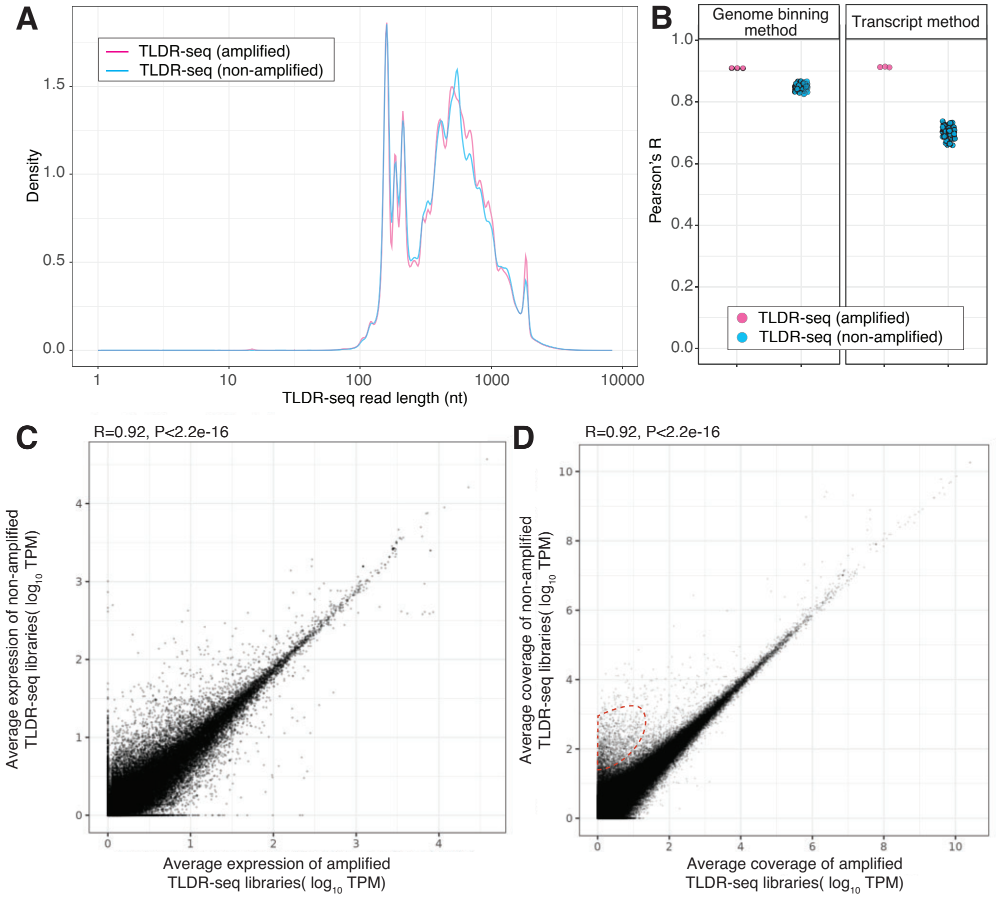


Figure 2. TLDR-seq reproducibility and comparisons between amplified and nonamplified TLDR-seq libraries. (**A**) Read length distributions of amplified and nonamplified TLDR-seq libraries. The *X*-axis shows read length in nt (log scaled). The *Y*-axis shows density. All libraries of respective types are pooled: see Supplementary Fig. S2A and B for distributions of individual libraries. (**B**) Summaries of pairwise correlation analyses within each library type. The *Y*-axis shows Pearson's correlation coefficient values between two libraries. *X*-axis shows library type. Facets show two different methods of comparing libraries: by counting reads in genomics 500-bp bins (left) or by mapping to transcripts (right). Dots correspond to pairwise Pearson's correlation coefficients between each pair of libraries within each group (3 amplified libraries and 12 nonamplified). See Supplementary Fig. S2C–F for pairwise correlation plots. (**C**) Summary of pairwise correlation analyses between library types, using the transcriptome method. The *X*- and *Y*-axes show the average $\log_{10}$ TPM values from pooled amplified (*X*) or non-amplified libraries (*Y*). Each dot corresponds to one transcript. Pearson's correlation coefficients and correlation test *P*-values are shown. See Supplementary Fig. S2G for pairwise correlation plots. (**D**) Summary of pairwise correlation analyses between library types, using the genome binning method. Organized as in panel (C), but axis shows average $\log_{10}$ coverage in 500-bp bins, normalized by library sequencing depth. Each dot shows one bin. See Supplementary Fig. S2H for pairwise correlation plots. A set of regions where nonamplified libraries have higher signal than amplified libraries are highlighted.

from amplified libraries versus pooled reads from nonamplified libraries (Fig. 2C and D). Pooled amplified and nonamplified libraries were highly correlated (Pearson's $R = 0.92$ for the binning method and $R = 0.92$ for the isoform quantification method), and slightly less correlated when comparing individual libraries, with an average Pearson's R of 0.73 (SD $= 0.03$) using the binning method and $R = 0.89$ (SD $= 0.05$) using the isoform quantification method. The Pearson's R decrease when assessing single libraries was expected since the read depth in individual nonamplified libraries was lower and likely had a lower capture rate of lowly expressed transcripts (similar to the correlation decrease when comparing nonamplified libraries with each other).

Most differences between amplified and nonamplified libraries could only be observed using the genome-binning method, and derived from a subset of regions with overall low sequencing depth but higher coverage in nonamplified versus amplified libraries (Fig. 2D, see highlight in red). We found that the regions where nonamplified libraries had substantially higher coverage (based on nonamplified versus amplified fold change) were mostly from the 5′ or 3′ ends of a handful of genes, where the remaining gene body did not show any substantial difference between amplified and nonamplified libraries. The isoform detection method did not detect these cases as it counted the sum of reads within each isoform. Because the discrepancies were mainly located in gene boundaries, we reasoned that the PCR primer efficiency could be affected by local GC content. Therefore, for the 150 regions with the highest nonamplified versus amplified fold change values, we calculated the GC content of the 5′ and 3′ UTRs of the overlapping transcripts (see the "Materials and methods" section), and compared this with corresponding data from the regions with the lowest nonamplified versus amplified fold change values. We found that UTRs associated with higher signal in nonamplified libraries on average had higher GC content (Supplementary Fig. S2I); thus, such transcripts are less likely to be amplified properly in the TLDR amplification step. This is likely due to the formation of stable secondary structures in these high-GC content regions impeding PCR amplification.

As a summary, the TLDR-seq method has satisfying levels of reproducibility. Moreover, these analyses show that with the exception of a few highly GC-rich isoform boundaries, the cDNA libraries can be amplified without introducing substantial bias.

Evaluation of 5′-end detection

An important feature of TLDR-seq is the cap trapping, aiming to accurately detect 5′ ends of successfully capped RNAPII-RNAs. We evaluated 5′-end detection of TLDR-seq using CAGE data from the same cells [27] and GENCODE-annotated [43] TSSs. For comparison, we also evaluated 5′-end detection of a standard ONT cDNA library from HeLa cells from [24]. These sets have different characteristics: first, since the CAGE method uses random priming and thus captures RNAs regardless of polyadenylation status, it should in principle detect the same 5′ ends as TLDR-seq as it was used on the same cells; however, the CAGE libraries contain a total of 88 million mapped reads compared with a total of 5.1 and 4.6 million reads in amplified and nonamplified TLDR-seq libraries, respectively, and would thus be expected to have higher sensitivity, so it may detect lowly expressed TSSs that TLDR-seq cannot. Second, the standard ONT cDNA library contains pA+-selected cDNA fragments resulting in 3.3 million reads mapped to the genome, making it comparable to the pA+ reads of the TLDR-seq libraries, but not pA− reads. Third, GENCODE transcript annotations are derived from large numbers of different cell types and tissues, where some transcripts will not be transcribed in HeLa cells. In addition, while GENCODE annotation includes non-mRNA annotations, its coverage of RNA decay intermediates and lncRNA species is lower than that mRNAs, since the annotations to a large extent are based on sequencing of pA+-selected RNAs.

We first assessed the ability of TLDR-seq to identify TSSs from the same cell type by comparison to HeLa CAGE data. Specifically, we measured the distance of called TSSs from TLDR-seq to the closest CAGE TSS on the same strand. This distance will be positive when a TSS of TLDR-seq falls 5′ of a CAGE TSS, negative if it falls 3′ of a CAGE TSS, and 0 if TSSs overlap on the same bp. To account for the highly different read numbers we defined a TSS for a given set as follows: a TSS must have at least least n reads starting at a given genomic position, where n is defined as (number of reads in the library)/10^6. For comparison, we applied the same analysis to assess TSSs from the standard ONT nanopore library. TLDR-seq and CAGE TSSs showed a high overlap and overall small divergence between TSS positions (Fig. 3A) compared with standard ONT nanopore TSSs, where a substantial fraction fell downstream of CAGE TSSs. The latter likely reflects the known limitation of the standard ONT nanopore protocol to identify 5′ ends, resulting in 5′-truncated transcripts, as discussed in the “Introduction” section.

We also compared TSSs from TLDR-seq, CAGE, and the standard ONT nanopore dataset with GENCODE TSSs. For this analysis, we chose GENCODE transcripts that, based on their annotation, were RNAPII- transcribed and whose 5′ end represented genuine TSS and not processed 5′ ends, e.g. 5′ ends produced by endocleavage (see the “Materials and methods” section and Supplementary Table S2). TLDR-seq and CAGE showed near-identical TSS distance distributions (Fig 3B, three leftmost plots) with a high agreement with GENCODE TSSs. Conversely, as expected, the standard ONT nanopore TSSs were, on average, further away from GENCODE TSS, and often fell downstream for GENCODE TSSs, similar to the analysis above.

Next, we analyzed DNA sequences around TSSs from respective sets to see if typical core promoter motifs could be identified at expected locations. Previous work by us and others have shown that highly expressed genes often use an array of closely spaced TSSs [45, 46, 50, 51]. Therefore, for a given dataset, we identified TSSs as above, and if many such TSSs were within an [−40, 40]-nt window at the same strand, the most expressed TSS was used. We then calculated the nucleotide frequencies in a [−50, 50]-nt window around these TSSs, and sequences were reverse complemented when TSSs were on the reverse strand. Two noticeable sequence patterns stood out around TLDR-seq TSSs (Fig 3C, left). First, a 10-bp wide peak of A/T enrichment at −24 to −32 bp, also shown as a sequence logo (Fig 3D): this sequence pattern likely corresponds to the TATA box. This pattern could also be identified using CAGE TSSs, but not standard nanopore TSSs (Fig. 3C, middle and right). Second, the well known pyrimidine–purine (PyPu) motif that is the core of the Initiator motif (INR, reviewed in [52]) were identified at the −1, +1 position in TLDR-seq and CAGE, but not in standard ONT nanopore (Fig. 3C and E). For standard ONT nanopore TSS, the most prominent sequence pattern was instead a weak enrichment of a G at the +1 nucleotide, likely a bias attributed to the template switching used in the ONT nanopore protocol, as it preferentially selects for mRNAs with a G at the 5′ end.

As a summary, TLDR-seq can accurately determine the 5′ ends of full-length fragments and substantially improves TSS detection compared with the standard ONT nanopore protocol.

Evaluation of 3′-end detection

Similarly to above, we evaluated 3′-end detection using 3′-end sequencing (QuantSeq) from HeLa cells and annotated GENCODE 3′ ends. We first focused on pA+ transcripts and therefore compared TLDR-seq 3′ ends from the pA+ fraction and standard ONT nanopore 3′ ends with pA+-selected QuantSeq 3′ ends, using the same approach as we used for 5′ ends. Regardless of whether amplification was used, TLDR-seq 3′ ends had a substantially higher degree of overlap to QuantSeq 3′ ends than standard ONT nanopore 3′ ends (Fig. 4A), which often fell upstream of QuantSeq 3′ ends. We also compared the same TLDR-seq and standard ONT nanopore sets to GENCODE 3′ ends. Here, for a fair comparison, we focused on 3′ ends of GENCODE-annotated mRNAs, as we expect those to be pA+ . The results were similar as above: regardless of whether amplification was used, there was a high overlap between TLDR-seq 3′ ends and GENCODE mRNA 3′ ends, while 3′ ends from the standard nanopore where frequently displaced upstream from the GENCODE mRNA 3′ ends (Fig. 4B). The same pattern was also evident by comparing GENCODE 3′ ends with TLDR-seq pA+ 3′ ends (Fig 4B).

Because TLDR-seq can also detect pA− transcripts, we compared the full pool of TLDR 3′ ends (including pA+ and pA−) with QuantSeq 3′ ends from a library generated from RNA that had first been subjected to *in vitro* polyadenylation using the polyA polymerase (PAP). For clarity, such

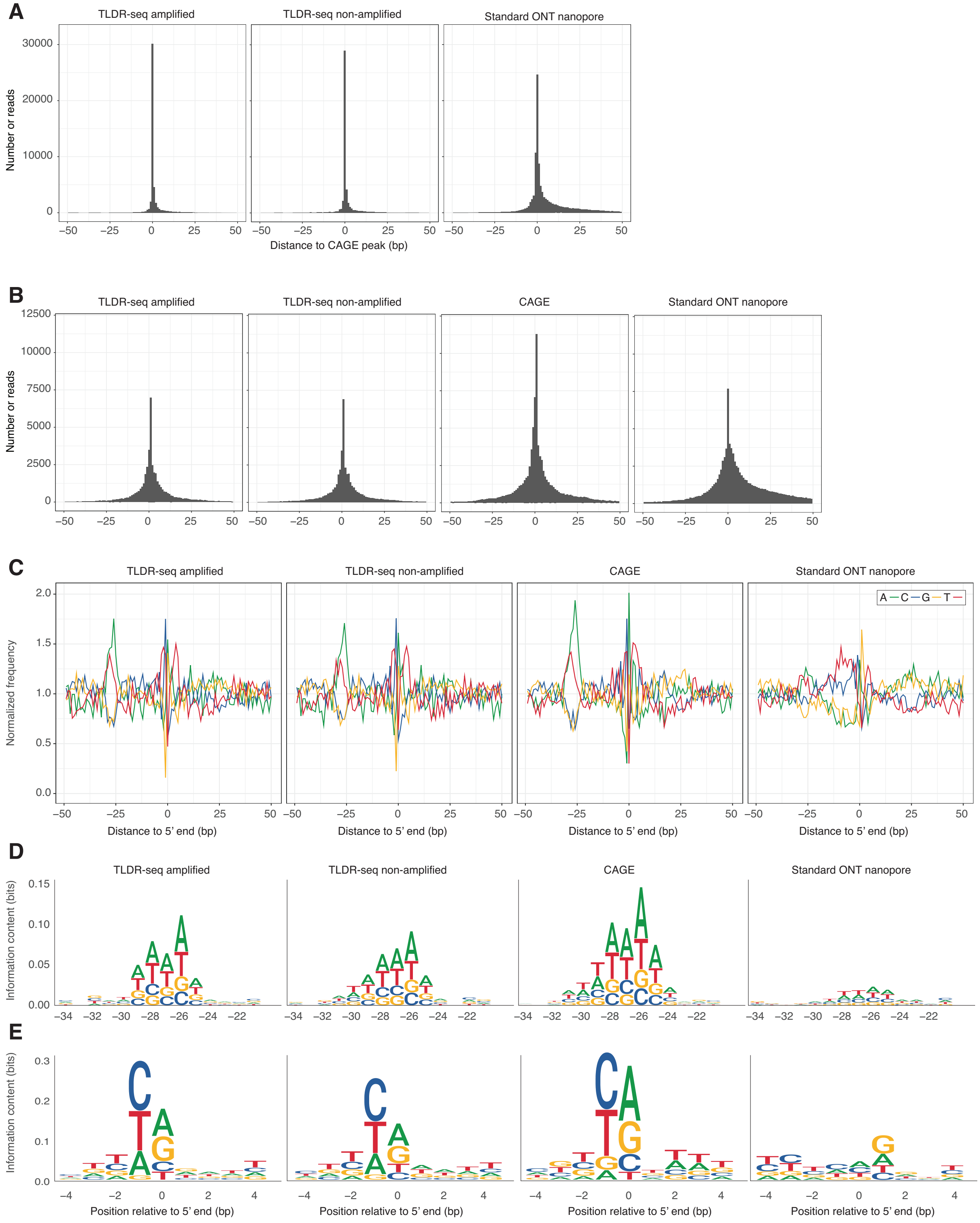


Figure 3. Evaluation of TSS detection. (**A**) Comparison of TSSs detected by TLDR-seq and standard ONT Nanopore cDNA sequencing to CAGE-defined TSSs in HeLa cells. The *X*-axis shows the closest distance in bp to a CAGE-defined TSS (see the "Materials and methods" section) on the same strand: transcription goes left to right. The *Y*-axis shows the number of TSSs from respective datasets that are located at a given distance. Each plot shows one dataset comparison to the same CAGE TSS set. (**B**) Comparison of TSSs detected by TLDR-seq, standard ONT nanopore, and CAGE in HeLa cells with GENCODE-defined TSSs. Plot is organized as in panel (A), but the *X*-axis shows distance to GENCODE-defined TSSs. (**C**) Normalized nucleotide distributions around TSSs defined by TLDR-seq, standard ONT nanopore, and CAGE in HeLa cells. Nucleotide frequencies are scaled so that each nucleotide average frequency is equal to 1 in the displayed window. The *Y*-axis shows the average scaled nucleotide frequencies (see the "Materials and methods" section). *X*-axis shows distance to TSSs defined by respective method. Line colors indicate nucleotides. Each box shows TSSs from one dataset. (**D**) Sequence logos at position −33 to −21 bp relative to TSSs defined by respective methods. The *Y*-axis shows information content in bits. *X*-axis shows relative location to TSS in bp. The TATA box is visible in TLDR-seq and CAGE libraries but not standard ONT nanopore. (**E**) Sequence logos at position −5 to −21 bp relative to TSSs defined by respective methods. Logos are organized as in panel (D). The PyPu motif at −1, 0, part of the Initiator (INR) motif, is visible in TLDR-seq and CAGE libraries but not standard ONT nanopore.

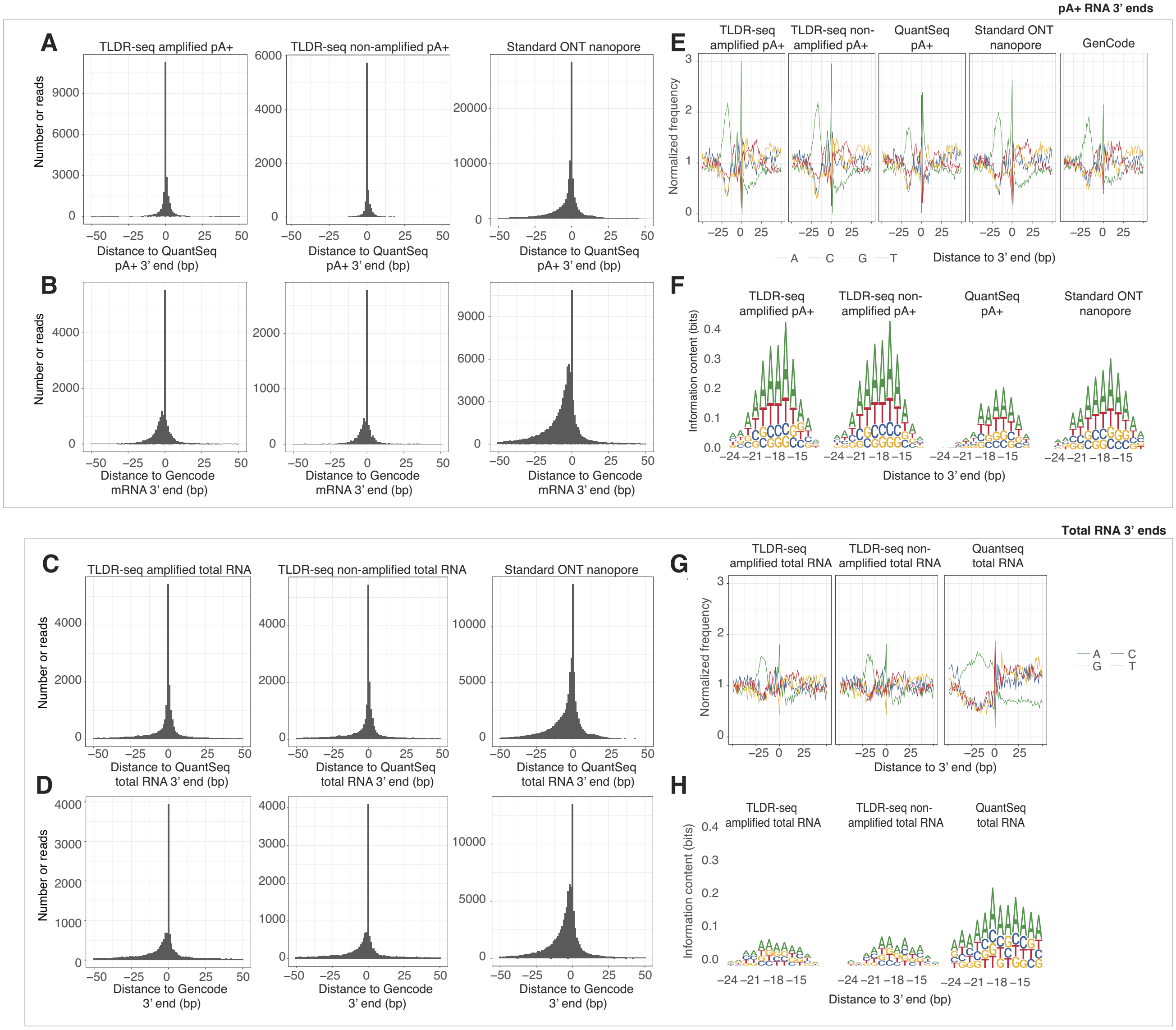


Figure 4. Evaluation of 3′ detection. (**A**) Comparison of pA+ 3′ detected by TLDR-seq and standard nanopore to QuantSeq-defined 3′ ends in HeLa cells. The *X*-axis shows the closest distance in bp to a QuantSeq-defined (pA+) 3′ end (see the "Materials and methods" section) on the same strand: transcription goes left to right. The *Y*-axis shows the number of 3′ ends from respective datasets that are located at a given distance. Each plot shows one dataset comparison to the same QuantSeq set. (**B**) Comparison of pA+ 3′ detected by TLDR-seq and standard nanopore to GENCODE-defined mRNA 3′ ends. Plot is organized as in panel (A), but the *X*-axis shows distance to GENCODE-defined mRNA 3′ ends. (**C**) Comparison of 3′ ends detected by TLDR-seq (all reads) and standard nanopore to QuantSeq-defined 3′ ends from PAP-treated libraries (total RNA) in HeLa cells. The *X*-axis shows the closest distance in bp to a QuantSeq-defined 3′ end, where RNAs were PAP-treated before 3′-end sequencing (see the "Materials and methods" section) on the same strand: transcription goes left to right. The *Y*-axis shows the number of 3′ ends from respective datasets that are located at a given distance. Each plot shows one dataset comparison with the same QuantSeq set. (**D**) Comparison of 3′ ends detected by TLDR-seq (all reads) and standard nanopore to GENCODE-defined 3′ ends. Plot is organized as in panel (B), but the *X*-axis shows distance to GENCODE-defined 3′ ends (same set as in Fig. 3, RNAPII-transcribed RNAs, including pA− transcripts). (**E**) Nucleotide frequency distributions in the −50 to + 50 region around pA+ RNA 3′ ends. The *X*-axis shows nucleotide positions around detected 3′ ends from each approach (one plot per approach, as indicated above the plots: for GENCODE, all RNAs are used). Only pA+ reads are used. The *Y*-axis shows normalized nucleotide frequency (each nucleotide frequency is centered around 1 within the [−50, 50]-bp window, see the "Materials and methods" section). Line colors correspond to nucleotides. (**F**) Sequence logos from the −24 to −12 region upstream of pA+ RNA 3′ ends. *X*-axis as in panel (E). The *Y*-axis shows information content in bits. Nucleotide colors as in panel (E). (**G**) Nucleotide frequency distributions in the −50 to +50 region around total RNA 3′ ends. The plot is organized as in panel (E), but uses all 3′ ends regardless of pA status. Standard ONT nanopore data are not shown since such reads are pA+ only. (**H**) Sequence logos from the −24 to −12 region upstream of total RNA 3′ ends. The plot is organized as in panel (F) but uses all 3′ ends regardless of pA status. Standard ONT nanopore data are not shown since such reads are pA+ only.

3′ ends will constitute a mix of 3′ ends from RNAs that were originally pA+ and pA−, making a comparison with the full TLDR read population relevant. Using these sets, we repeated the analysis described above. We found that as before, most TLDR-seq 3′ ends overlapped exactly with those from the QuantSeq library, arguably with more focused distributions of 3′ ends that were only a few base pair away from the QuantSeq 3′ ends compared with the pA+-only analysis above (Fig. 4C). As before, standard ONT nanopore 3′ ends showed a smaller overlap, although this method is disfavored in this analysis since it is detecting only pA+ transcripts. Highly similar results were obtained by comparing to GENCODE 3′ ends including all GENCODE transcript biotypes used in Fig. 3 (Supplementary Table S2, see the "Materials and methods" section and Fig. 4D).

As with 5′ ends, we analyzed the sequence content upstream and downstream of 3′ ends detected by each method. This showed similar results for all pA+ methods/libraries: an A-rich region was located around −24 to −12 bp upstream of the 3′ ends, likely corresponding to the AWTAA pA motif (Fig. 4E shows nucleotide frequencies in the +50-bp range, and Fig. 4F shows a sequence logo [34] for the −24 to −12 region). As expected, this pattern was not as prevalent when analyzing 3′ ends from the total TLDR-seq transcript pool (also including pA− transcripts: Fig. 4G and H). Similarly, 3′ ends from QuantSeq (PAP-treated library) displayed a more dispersed A-enriched region (Fig. 4G and H). The latter is consistent with previous observations of these libraries: 3′ ends detected by PAP libraries tend to be highly dispersed compared with pA+ 3′ ends [44].

As a summary, TLDR-seq 3′ ends agree well with experimentally defined 3′ ends from 3′ sequencing from the same cells, and GENCODE annotations that derive from many cell types. Somewhat surprisingly, standard nanopore 3′ ends seem to, at times, detect 3′ ends that are placed upstream of 3′ ends detected by any of the other methods, including TLDR-seq.

Evaluation of exon junction detection

One key benefit of full-length sequencing is its ability to accurately characterize full-length isoforms. Aside from identification of transcript 5′ and 3′ ends, as discussed above, precise location of splice junctions is therefore important. Since the difference of TLDR-seq and standard ONT nanopore protocols is only at 5′ and 3′ ends of RNAs/cDNA, we would expect similar splice site detection performance if the same transcripts are detected.

Therefore, we first compared TLDR-seq-derived donor (intron 5′ end) and acceptor (intron 3′ end) splice sites with those in the standard ONT nanopore dataset, and vice versa (Fig 5A–D). We did initially not distinguish between pA+ and pA− transcripts as we do not have a reference splice site compendium for pA− transcripts in HeLa cells. Overall, TLDR-seq donor and acceptor sites were highly supported by corresponding standard ONT nanopore sites, and vice versa. We observed a slightly higher discrepancy when standard ONT nanopore splice site junctions were compared with the TLDR-seq splice site junctions, compared with the reciprocal analysis (TLDR-seqs splice site junctions versus ONT nanopore splice site junctions). We hypothesized this was due to the inclusion of pA− reads in the TLDR-seq data, but when comparing standard ONT nanopore data with pA+ TLDR-seq reads, this small discrepancy was retained (Supplementary Fig. S3A and B).

Subsequently, we compared TLDR-seq and standard ONT nanopore junctions to GENCODE splice sites (GENCODE-transcripts selected as above). This showed near-identical performance between standard ONT nanopore and TLDR-seq data, where the majority of splice sites were right on top of GENCODE splice sites, or a few bp off (Fig. 5E,F). We conclude that TLDR-seq identification of splice sites is comparable to that of the standard ONT nanopore protocol.

Evaluation of expression level estimations compared with short-read approaches

Long-read cDNA sequencing is not primarily aimed at expression quantification, since the number of reads per library is typically much lower than short-read applications. Nevertheless, we reasoned that it would be interesting to assess how expression quantification of TLDR-seq compared with CAGE and RNA-seq data from the same cell line.

To evaluate the TLDR-seq expression quantification accuracy, we used three datasets from the same cell lines: (i) a total RNA-seq library which was ribo-depleted but not pA+-enriched, as used in Fig. 1, (ii) the same CAGE set as above: because CAGE is random-primed, the same population of cap-trapped RNAs are sampled, and (iii) the same standard ONT nanopore libraries as above.

Quantifying gene expression from these heterogeneous datasets required the use of two different methods. First, for comparison with between nanopore-based and short-read RNA-seq sets, we used transcript quantification tools which can estimate the gene expression after mapping the reads to the GENCODE v39 transcriptome: for RNA-seq we used Salmon [9] and for TLDR-seq and standard ONT nanopore we used Bambu [25] after the reads were mapped to the same transcriptome with minimap2 [53]. Second, comparisons between long-read methods (TLDR-seq and ONT nanopore) and CAGE sets cannot be done in the same way, because in CAGE, only the 5′ ends of transcripts are sequenced. Instead, we mapped CAGE, TLDR-seq and ONT Nanopore reads to the hg38 genome and trimmed them to only retain their 5′ terminal nucleotide (denominated as 'tag'). Tags were assigned to their nearest GENCODE gene within 100 bp, and gene tag counts were normalized by library sizes to obtain gene TPM. Using the above methods, we observed high correlation levels between TLDR-seq and RNA-seq (Pearson's $R = 0.90$ for both amplified and nonamplified TLDR-seq) (Fig. 6A) and TLDR-seq versus CAGE (Pearson's $R = 0.84$ for amplified TLDR-seq and $R = 0.86$ for nonamplified TLDR-seq) (Fig. 6B). For comparison, using the same methodology, the correlation between standard ONT nanopore and RNA-seq or CAGE was $R = 0.85$ and $R = 0.79$, respectively (Supplementary Fig. S4A and B). The smaller R values compared with that of TLDR-seq may be due to the pA+ selection in standard ONT nanopore. For reference, as discussed above, between-replicate correlations of TLDR-seq were $R = 0.66$–0.91 depending on correlation method and whether amplification was used (Fig. 2B).

As a summary, we conclude that TLDR-seq expression estimates on transcript or TSS level correlate well with expression estimates from respective short-read approaches based on full RNA samples from the same cell type.

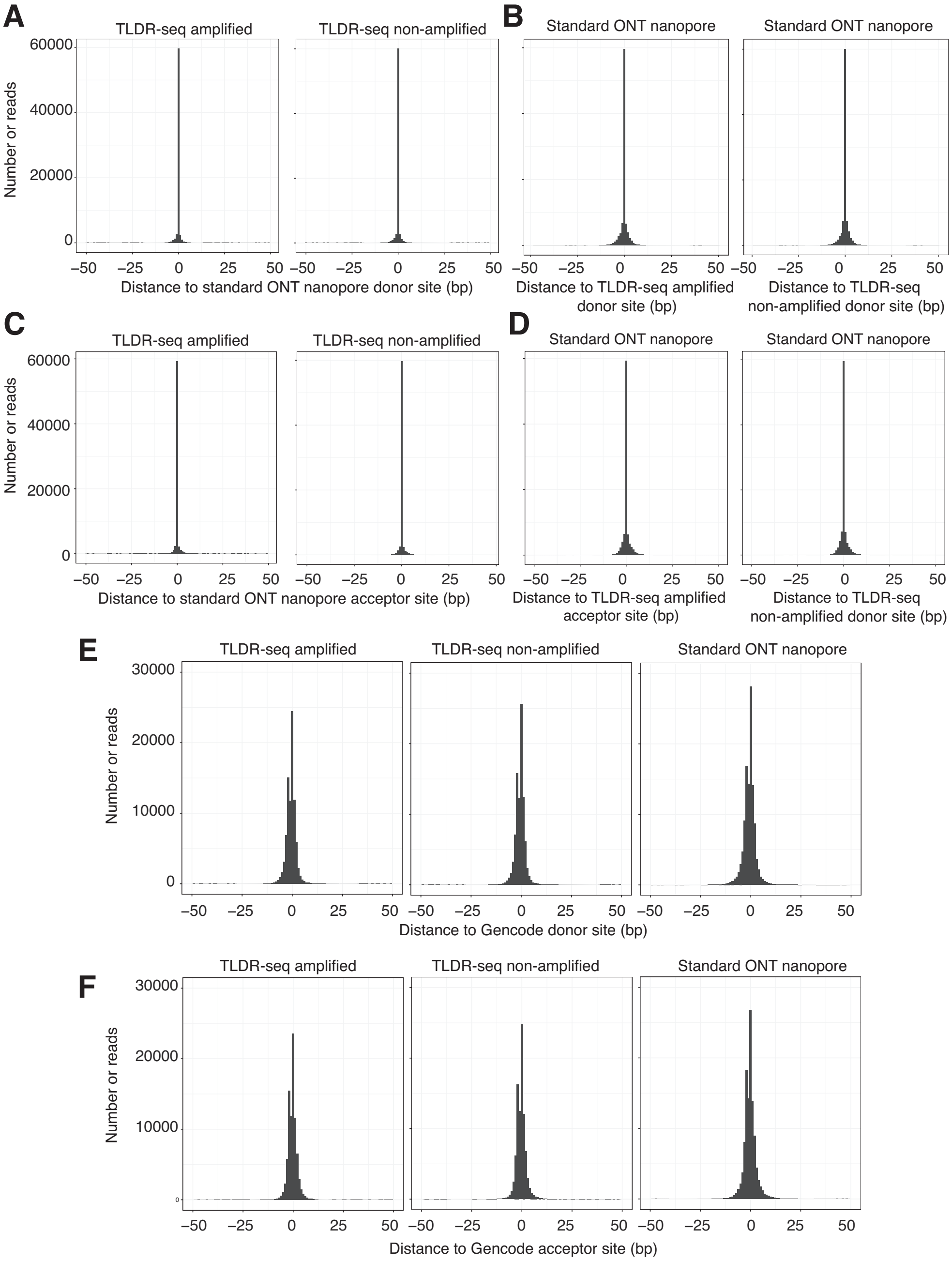


Figure 5. Evaluation of detection of splice sites. (**A**) Comparison of donor splice sites (intron 5′ ends) identified by TLDR-seq versus standard ONT nanopore from HeLa cells. The *X*-axis shows the closest distance in bp on the same strand from a TLDR-seq defined intron 5′ end to a standard ONT nanopore-defined intron 5′ end (see the "Materials and methods" section) in bp. The *Y*-axis shows the number of TLDR-seq reads that are located at a given distance. Plots show results from amplified and nonamplified TLDR-seq libraries as indicated on top. (**B**) Comparison of donor splice sites (intron 5′ ends) identified by the "Materials and methods" section in yourstandard ONT nanopore versus TLDR-seq from HeLa cells. The plot is organized as in panel (A) but the *X*-axis shows the closest distance in bp on the same strand from a standard ONT nanopore-defined intron 5′ end to a TLDR-seq-defined intron 5′ end (see Methods) in bp. The Y axis shows the number of standard ONT nanopore -seq reads that are located at a given distance. Plot shows results of standard ONT nanopore reads compared with amplified and nonamplified TLDR-seq library reads as indicated on the X axis. (**C**) Comparison of acceptor splice sites (intron 3′ ends) identified by TLDR-seq versus standard ONT nanopore from HeLa cells. Figure is organized as in panel (A) but the X axis shows distance to donor splice sites in bp. (**D**) Comparison of acceptor splice sites (intron 3′ ends) identified by standard ONT nanopore versus TLDR-seq from HeLa cells. Figure is organized as in B but the *X*-axis shows distance to donor splice sites in bp. (**E**) Comparison of donor splice sites (intron 5′ ends) identified by TLDR-seq and standard ONT nanopore to GENCODE annotation. Figure is organized as in panel (A) but the *X*-axis shows distance to GENCODE-defined donor splice sites in bp. Each plot shows one dataset comparison to the same GENCODE set. (**F**) Comparison of acceptor splice sites (intron 3′ ends) identified by TLDR-seq and standard ONT nanopore to GENCODE annotation. Figure is organized as in E but the X axis shows distance to GENCODE-defined acceptor splice sites in bp. Each plot shows one dataset comparison to the same GENCODE set.

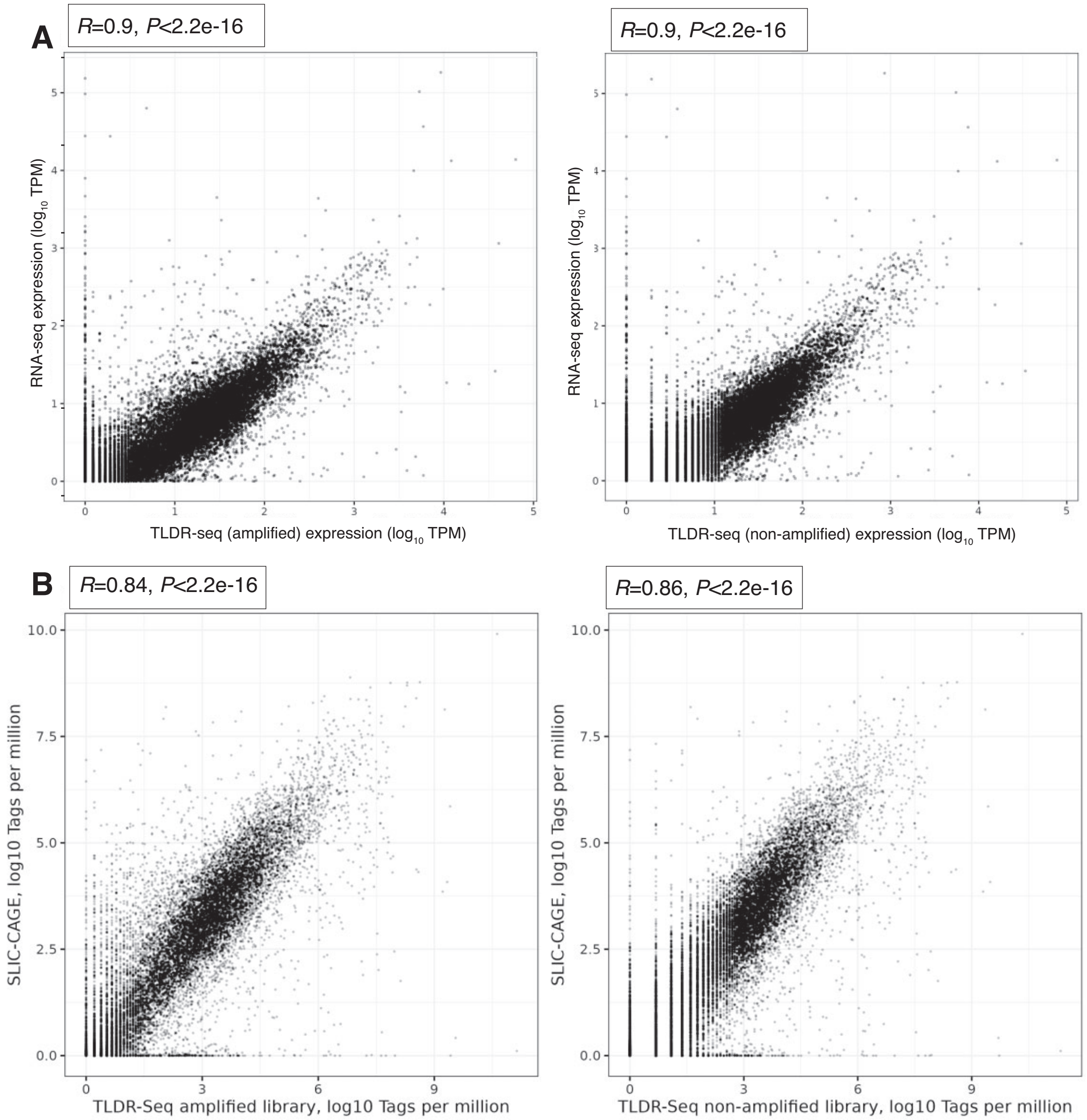


Figure 6. Comparison of transcript level expression levels between TLDR-seq and short-read approaches. (**A**) Comparison of library transcript level expression estimates between TLDR-seq and total RNA-seq from HeLa cells. Axes show $\log_{10}$ TPM expression levels estimated per GENCODE gene from TLDR-seq (*X*-axis) and RNA-seq (*Y*-axis). The left plot shows amplified TLDR-seq libraries, the right plot shows nonamplified TLDR-seq libraries. Pearsons's *R* values and associated *P*-values are shown. (**B**) Comparison of TSS-level expression estimates between TLDR-seq and CAGE from HeLa cells. Axes show $\log_{10}$ TPM expression levels estimated per GENCODE genes from TLDR-seq (X) and CAGE (Y). Left plot shows amplified TLDR-seq libraries, the right plot shows nonamplified TLDR-seq libraries. The TLDR-seq range is different between panels (A) and (B): this is because the calculation method is different to make it comparable to RNA-seq and CAGE, respectively (see the "Materials and methods" section).

TLDR-seq can be used with R9 or R10 flow cells without modifications

The above analyses were made using the R9 Oxford Nanopore flow cell. To ensure that the method also works with the newer R10 flow cell, we re-sequenced the three amplified HeLa libraries using an R10 flow cell (three barcoded libraries in one flow cell, as in our R9 flow cell experiment). This produced on average 11e6 reads per library, of which 65% had both adapters, of which 96.5% (±0.04 between libraries) mapped to the genome. This was roughly twice the number of reads compared with the R9 flow cells, with similar adapter inclusion and mapping rate (in R9, 63% and 98%, respectively). We then assessed the resulting reads in the same way as the R9 reads (as in Figs. 1–5) The distribution of R10 reads among annotated RNA classes was highly similar to that of R9 reads (Supplementary Fig. S5A). The reproducibility between R10 libraries was high and on similar levels as the corresponding R9 library analyses (Pearsons's *R* of 0.803–0.813 for bin-based comparison and 0.705–0.711 for transcript-based comparisons; Supplementary Fig. S5B and C), and expression quantification in genome bins or per transcript were highly similar between R9 and R10 (Supplementary Fig. S5D and E). The detection of 5′ ends (compared with CAGE and GENCODE annotation, Supplementary Fig. S5F,G), 3′ ends (compared QuantSeq and GENCODE annotation, with and without selection for pA+, Supplementary Fig. S5H-K) and splice sites (compared with standard ONT Nanopore and GENCODE, Supplementary Fig. S5L-Q) were highly similar between R10 and R9 libraries. Additional genome browser visualizations showing reads from both flow cells are available in Supplementary Fig. S5R-T.

Analysis of associations between alternative TSS usage and RNA polyadenylation using TLDR-seq

We and others, using short-read 5′-end sequencing techniques such as CAGE, have previously established that most mammalian core promoters use an array of closely located TSSs—typically covering a region up to ∼100 bp, rather than a single TSS (reviewed in [54]). It has previously been difficult to ascertain whether the choice of transcription initiation sites on this local scale makes a difference for the RNA elongation and/or processing (e.g. splicing and polyadenylation) of the resulting transcript. Since TLDR-seq is the only method to our knowledge that has the ability to characterize full-length transcripts regardless of pA status with highly accurate TSS locations, we used the pooled amplified HeLa library data to address whether there are core promoters for known genes where local TSS location choice is associated with polyadenylation status. Specifically, we identified TLDR-seq TSSs occurring within, and on the same strand of, annotated first exons including ± 100-bp flanking regions (3253 cases), and for each such case tested whether the distribution of TSSs from pA+ and pA− reads were significantly shifted positionally (Wilcox test, FDR < 0.05, see the "Materials and methods" section). This identified 64 first exons with significant shifts. Two genome-browser examples of such TSS shifts are shown in Fig. 7A and B. Somewhat surprisingly, when TSS shifts associated with pA status occurred, pA+ TSS distributions were on average significantly more dispersed than pA− TSS distributions, as measured by positional entropy ($P = 5\text{e}{-}3$, Wilcoxon paired test, Fig. 7C).

Next, we asked whether the pA+ and pA− reads originating from such shifts represented substantially different transcripts in terms of splicing patterns and read lengths. We found genomic lengths of pA+ reads were typically larger than pA− reads, but the difference was not large (median 398 bp, Fig. 7D). Similarly the pA+ reads were typically longer than pA− reads but the difference was minor (median 202 nt, Fig. 7D), and the pA+ reads typically had the same number of exons than pA− reds, or one more exon (Supplementary Fig. S6A). Together with manual inspections, we interpret these results as follows: in most cases, the splicing patterns were similar or identical up to the second-last or last exon, where the shorter read length in the pA− fraction was likely capturing 3′ degradation following the lack of polyadenylation.

Genes that had one or more of these local TSS shifts associated to pA status were overrepresented in GO terms related to translation, where 23 of 64 genes featuring significant shifts were GO (Supplementary Fig. S6B). This is interesting as ribosome protein genes have a specific core promoter structure, featuring a ribosomal protein-specific initiator sequence, the so-called TCT motif [55]. Although it is possible that this motif structure may contribute to the shifts observed, pA− and pA+ TSSs in ribosomal protein promoters with substantial TSS shifts both had the TCT motif, although, perhaps related to the larger diversity of TSSs for the pA+ reads, the TCT average motif strength appeared weaker for pA+ read TSSs (Supplementary Fig. S6C).

We reasoned that the same phenomenon may occur at larger scales, where two distinct alternative promoters for the same gene produce substantially different ratios of pA+ and pA− reads. To identify such cases, we identified genes that had two or more distinct clusters of TLDR-seq derived TSSs on the same strand (see the "Materials and methods" section). Using TSS clusters that had at least 10 pA+ or 10 pA− reads, we classified them into three categories based on their pA+/pA− $\log_2$ fold change ($\log_2$FC) : enriched in pA+ ($\log_2\text{FC} > 1$), enriched in pA− ($\log_2\text{FC} < -1$), and nonenriched ($|\log_2\text{FC}| < 0.3$). We then selected only genes containing at least two TCs that had at least two different categories. This identified 306 genes, out of which 107 had TCs that overlapped GENCODE annotated TSS. One example of this is shown in Supplementary Fig. S6D. Somewhat surprisingly, out of these TSS clusters, pA− biased clusters were more often overlapping annotated TSSs, while the pA+ biased TCs were overrepresented in 3′ UTRs (Supplementary Fig. S6E): the latter is likely the same observation as discussed in Fig. 1E, and previously described [45–48].

As a summary, although the causality is unclear, our observations strongly argues that the TSS selection, either at distinct promoters or by local TSS shifts on bp resolution level in the same core promoter, can in some core promoters be strongly correlated to transcript processing and fate. Notably, this observation requires the unique capabilities of TLDR-seq, as it requires base-pair level TSS identification and simultaneous characterization of pA+ and pA transcripts.

Detection and characterization of PROMPTs and enhancer RNAs using TLDR-seq

As exemplified in Fig. 1 D–F, TLDR-seq can capture both pA+ and pA− transcripts, where the latter category includes lncRNAs. We and others have reported that most mRNA promoters initiate transcription bidirectionally, transcribing short, capped, pA− lncRNAs in the opposite strand relative to the mRNA TSSs which are targeted by the nuclear exosome, referred to as PROMPTs or asRNAs/uaRNAs [56–60]. Similarly, active enhancer regions also initiated bidirectional transcription [61], where both strands produce capped, pA− enhancer RNAs (eRNAs) that share many features with PROMPTs, including their rapid degradation by the nuclear RNA exosome complex [62, 63]. The RNA features of PROMPTS and eRNAs have previously only been characterized by short-read approaches (e.g. CAGE [62] and nascent RNA approaches [64]).

We found that TLDR-seq data from WT HeLa cells could identify clear cases of PROMPTs and bidirectional eRNAs, despite having functional RNA decay mechanisms (Supplementary Fig. S7A and B). Because of their rapid degradation, such RNAs are easier to capture if one or more components of the nuclear exosome or its adaptor complexes are depleted, previously shown [56, 57]. Therefore, to explore whether TLDR-seq can be used to characterize PROMPTS and eRNAs in greater detail, we used a dTAG-based degron system [65] to rapidly deplete RBM7 and ZCCHC8 in mES cells, key components of the nuclear exosome targeting complex [66]) that targets nuclear pA− RNAs including PROMPTs and eRNAs for decay via the exosome [44, 67]. Specifically, endogenous RBM7 or ZCCHC8 alleles were tagged with 2xHA-FKBP-V (HA-dTAG) by CRISPR/Cas9, allowing for the inducible depletion of tagged proteins using dTAG$^{\text{V}}$-1. We isolated total RNA from WT (ctrl), HA-dTAG-RBM7, [21] and HA-dTAG-ZCCHC8 cells following 4 h of dTAG$^{\text{V}}$-1 treatment and applied (amplified) TLDR-seq. TLDR-seq reads were mapped as above but on the mm39 genome. The sequencing depth was 6.1·10e6, 4.5·10e6 and

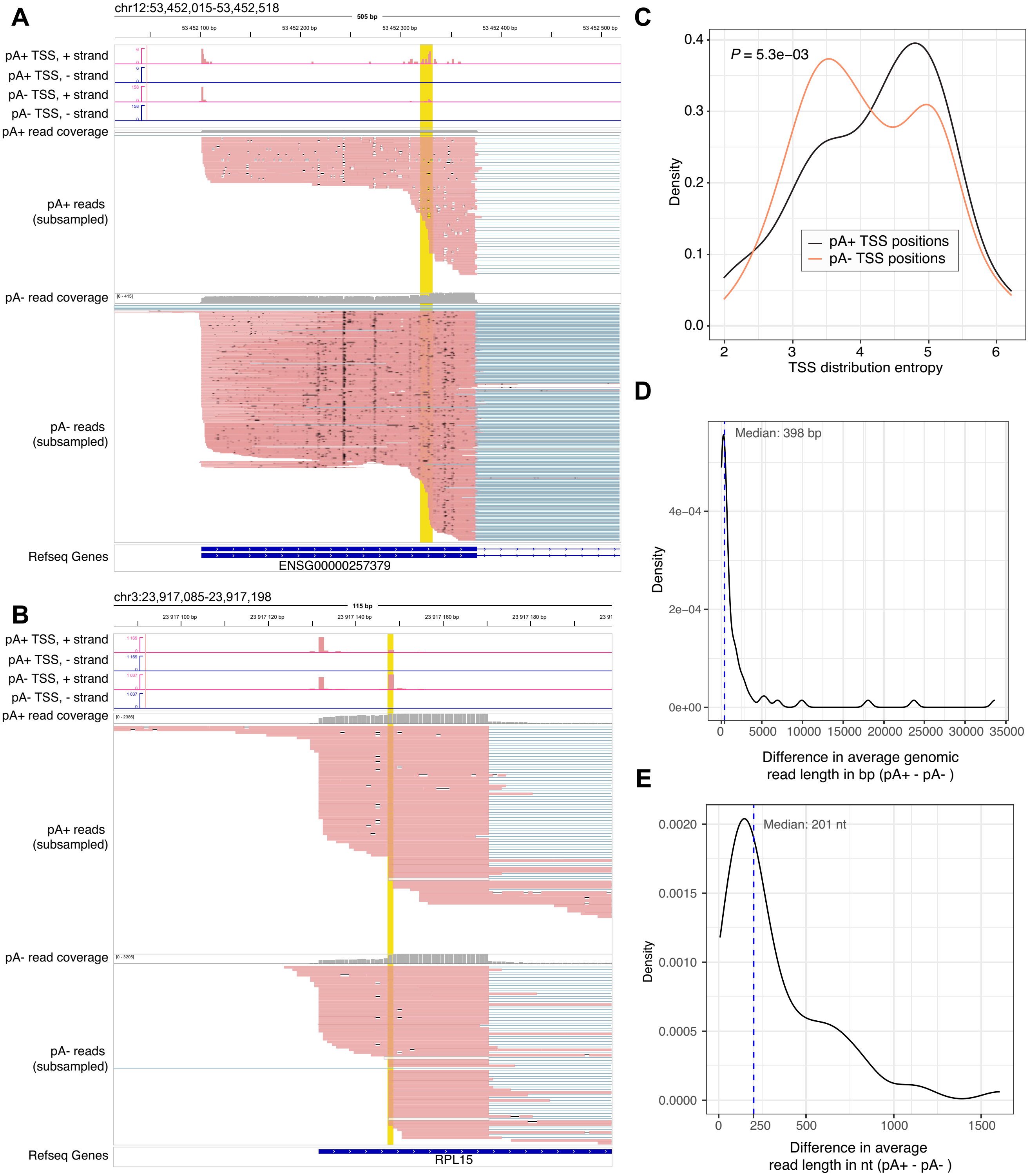


Figure 7. Analysis of local TSS shifts in core promoters associated with polyadenylation status. (**A**, **B**) Genome browser examples of TSS shifts in the ENSG00000257379 and RPL15 loci using TLDR-seq data in HeLa cells. Top tracks show counts of 5′ ends of TLDR-seq reads (pooled amplified libraries in HeLa cells), stratified by strand and whether reads are pA+ or pA−. Middle tracks show TLDR-seq reads and overall coverage of reads (bar plot on top), stratified by whether reads are pA+ or pA−. Note that shown reads are randomly subsampled. Blue color indicates minus strand reads, red indicates plus strand reads. Vertical highlights show the main TSS(s) that are shifted between pA+ and pA− reads. RefSeq gene annotations are shown at the bottom. Genome browser graphics based on the IGV browser [68]. Coordinates refer to the hg38 assembly. (**C**) TSS dispersion distributions of pA+ and pA− read TSS in core promoters featuring polyadenylation-associated TSS shifts. The plot shows the distributions of TSS position dispersion for pA+ and pA− reads in core promoters featuring TSS shifts (64 cases), measured by entropy (*X*-axis). *P*-value from Wilcoxon paired test between distributions. (**D**) Distribution of average difference in genomic alignment lengths of pA+ and pA− reads originating from core promoters with TSS shifts. *Y*-axis shows density, *X*-axis shows the difference between average genomic lengths of pA+ reads and average genomic lengths of pA+ reads in bp, for each analyzed case (*N* = 64). Dotted line shows the distribution median, with the median value highlighted. (**E**) Distribution of average difference in read lengths of pA+ and pA− reads originating from core promoters with TSS shifts. Arranged as in D, but plots read length, not genomic length.

4.2·10e6 mapped reads for Ctrl, HA-dTAG-RBM7-, and HA-dTAG-ZCCHC8- libraries.

As in HeLa cells, it was possible to detect of PROMPTs and eRNAs in mES cells even when the RNA decay mechanism was unperturbed (Ctrl cells from above), but as expected, the number of reads corresponding to these RNAs were overall higher in RBM7 and ZCCHC8 depletions (Fig. 8A,B and Supplementary Fig. S7C-H)

To characterize PROMPTs in more detail, we identified GENCODE-annotated mRNA TSSs that were highly transcribed and produced polyA+ RNAs in Ctrl mES cells, as detected by TLDR-seq (see the "Materials and methods" section). We then plotted the TLDR-seq coverage on both strands in the −1500 to +500 region in Ctrl and both depletions for each locus, stratified by whether reads were pA+ or pA− (Fig. 8C). We could identify 483 clear cases of mRNA PROMPTs for highly transcribed mRNAs. In the Ctrl library, we could identify 158 of these cases (Supplementary Fig. S7I). The large majority of PROMPTs initiated 500 bp or less upstream of the mRNA TSS, in a PROMPT TSS pattern echoing results from previous CAGE-based studies [57, 63], flanking DNAse hypersensitive site edges (Fig. 8C and Supplementary Fig. S7J). PROMPTs were rarely spliced: 98% of PROMPT reads had one exon (average across cell treatments), compared with mRNA-overlapping reads which had on average 2.6 exons (Fig. 8D). The average length of PROMPTs (see the "Materials and methods" section) was 459 nt, covering an average of 485 genomic bp (including introns, Fig. 8E). For reference, pA+ mRNA-associated reads had an average read length of 695 nt, covering an average of 4218 genomic bp (Fig. 8E). Thus, the majority of PROMPTs are pA−, unspliced, and on average 450-nt long.

We identified 1398 candidate eRNA loci, identified by DNAse hypersensitive sites from mES cells from from the encode dataset ENCFF048DWN [28] in intergenic or intronic regions which had bidirectional TLDR-seq read TSSs from RBM7 and/or ZCCHC8 depletions (see Methods): 204 of these were also detected in Ctrl cells (Supplementary Fig. S7K) and plotted TLDR-seq coverage around each DHS on both strands (±2000 bp) (Fig. 8F). This showed clear bidirectional TSSs signatures placed at the edges of the DHS peaks (Supplementary Fig. S7L), echoing similar CAGE-base analysis [63]. eRNAs were detected strongly in the pA− fraction and to a limited degree also in the pA+ fraction of RBM7 and ZCCHC8 depletions (Fig. 8F). Similar to PROMPTs, eRNAs were mono-exonic in 95% of cases (Fig. 8D), and had an average read length of 473 nt and genomic length of 646 bp including introns (Fig. 8E). Interestingly, the heat map visualization in Fig. 7F indicated that in most loci, most eRNAs were short (around 500 nt), but a small subset were substantially longer (>1000 nt), suggesting that most loci can produce a diverse collection of eRNAs.

Overall, these analyses exemplify the ability of TLDR-seq to identify pA− lncRNAs and characterize their splicing patterns and lengths.

Discussion

Here, we present TLDR-seq, a method to sequence full-length cDNAs from 5′ to 3′ ends regardless of polyadenylation status. The method selects for m^{7}G-capped RNAs, so no ribosomal depletion is necessary. We show that splice sites, 5′ and 3′ ends are accurately identified by comparing with other methods employed in HeLa cells, or genome-wide annotations, and we furthermore show that expression quantifications using the method are highly correlated to that of short-read approaches. In particular, compared with data from the standard ONT cDNA protocol, TLDR-seq is substantially better at identifying correct 5′ ends, and is able to detect pA− RNA species such as diverse lncRNAs that cannot be detected by the standard ONT nanopore cDNA protocol.

The method has certain limitations. First, as with all long-read methods, sensitivity will be lower than corresponding short-read methods (e.g. TLDR-seq versus CAGE for capturing 5′ ends) due to lower effective sequencing depth (fewer, but longer, reads). Thus, low copy number transcripts, such as some lncRNAs, or RNAs that are only expressed in a small subset of cells, may be challenging to detect reliably, and quantification of rare detected transcripts will be less reliable. For example, while we can detect eRNAs, which typically have low copy numbers due to their rapid degradation, with the sequencing depths used in this paper in both HeLa and mES cells, the number of eRNA loci detected were limited. This can be remedied by using multiple sequencing runs, or larger flow cells, if needed. For instance, in a pilot experiment, a ONT PromethION platform run of amplified TLDR-seq libraries produced ∼10 times the number of reads compared with the ONT MinION runs used here, per flow cell, and we also note that the R10 MinIon flow cell has roughly double the yield of the R9 flow cell used for most analyses above. TLDR-seq can also be used in parallel with high-depth short-read methods, where TLDR-seq would be used to identify transcripts while shorter read methods would be used for quantifications by mapping short reads to such TLDR-detected transcripts.

Second, an inherent kinetics-based bias in all ONT nanopore approaches is that shorter cDNAs take less time to sequence than longer, so there is a capture bias for such short reads. We believe this is not a large issue since this will be true for all TLDR-seq runs, enabling comparisons between runs, and that the primary use with ONT nanopore cDNA-based methods is not expression quantification but transcript discovery.

Third, as in all other long read cDNA-based approaches, reverse transcription and amplification steps in TLDR-seq may introduce biases for shorter cDNAs with lower GC content or truncation of the polyA tail. On top of this, due to the conversion of RNA information to a cDNA library, epitranscriptomic modifications of the native RNAs cannot be measured.

Fourth, a relatively large amount (8 μg) of RNA is required for the method and although this can be easily obtained from cultured cells as in this study, it can be challenging for clinical samples or rare cell types. An avenue for improving this aspect is the supplementation of synthetic carrier RNAs to allow the use of lower input RNA for TLDR-seq, as done in SLIC-CAGE [50].

Fifth, an in-built limitation of TLDR-seq is the selection for capped transcripts. While this allows for accurate TSS detection and removes the need for ribosomal depletions, uncapped RNAs, resulting from e.g. endonucleolytic cleavage, will not be detected. This means that many processed noncoding RNAs such as mature miRNAs, will not be detectable. This can be viewed as an advantage or disadvantage depending on the biological questions asked.

Regardless, we believe this method will be useful for transcript discovery and characterization, because it enables

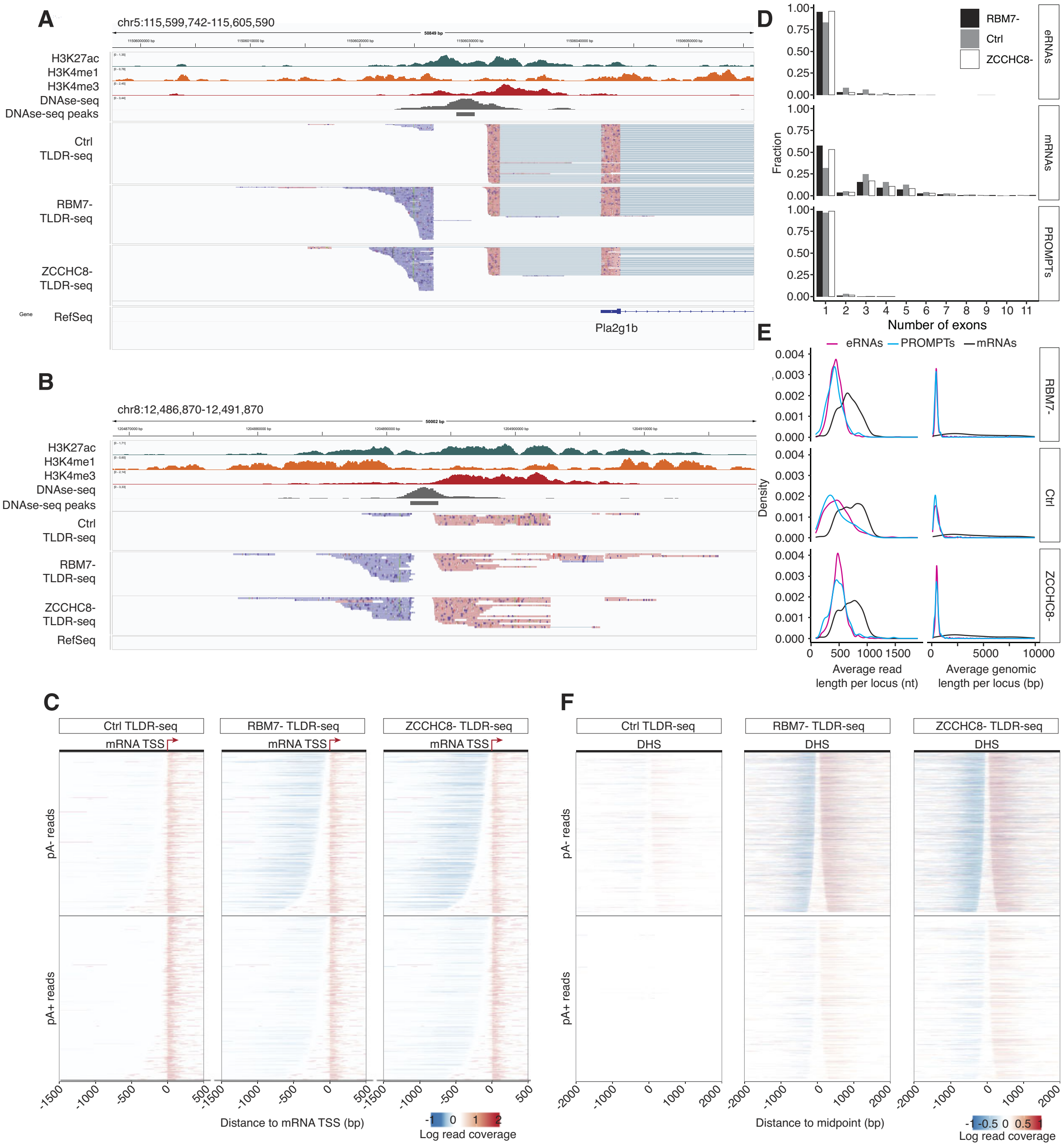


Figure 8. Analysis of PROMPTs and eRNAs in mES cells using TLDR-seq. (**A**) Genome browser example of detection of a PROMPT-mRNA locus using TLDR-seq in mES cells. Top tracks show coverage of Chromatin Immunoprecipitation followed by Sequencing (ChIP-seq) for H3K27ac, H3K4me1, H3K4me3, and DNase1-seq. DNase peaks (DHSs) are shown below. Middle tracks show TLDR-seq reads from Ctrl, RBM7 depleted and ZCCHC8 depleted cells. Blue color indicates minus strand reads, red indicates plus strand reads. RefSeq gene annotations are shown at the bottom. Genome browser graphics based on the IGV browser [68]. Coordinates refer to the GRCm39 assembly. (**B**) Genome browser example of enhancer RNAs using TLDR-seq in mES cells. Organized as in panel A but showing a DHS with bidirectional transcription. (**C**) Detection of mRNA-PROMPT loci using TLDR-seq in mES cells. The X axis indicates position in bp relative to an mRNA TSS (X = 0), where mRNA transcription goes from left to right (also indicated in schematic on the top of each subplot). Each row on the Y axis indicates one mRNA TSS locus, organized in the same order in all subplots. Colored horizontal lines indicate TLDR-seq coverage, where higher color intensity indicates higher numbers of mapped reads: red lines indicate reads on the same strand as the annotated mRNA, blue lines indicate reads on the opposite strand. Subplots with TLDR-seq reads from Ctrl, RBM7 depleted and ZCCHC8 depleted cells are shown, where reads are split whether reads are pA+ or pA−. Also see Supplementary Fig. S7J for the same TLDR-seq data with overlaid DNase-seq coverage. (**D**) Number of exons detected in PROMPTs, enhancer RNAs and mRNAs by TLDR-seq in mES cells. The Y axis shows the fraction of TLDR-seq reads having a given number of exons (X axis). Color indicates cell treatment. Each subplot shows data from one RNA type. (**E**) Length distributions of PROMPTs, enhancer RNAs and mRNAs by TLDR-seq in mES cells. For each locus, an average RNA length was calculated across all reads. Plots show the distribution of this statistic across all RNAs of a given type from a given cell. Y axis show density, X axis show read length in nt (left) and length of genomic mapping of the read, including introns in bp (right). (**F**) Detection of enhancer RNAs using TLDR-seq in mES cells. Each row on the Y axis indicates one DHS locus, organized in the same order in all subplots. The X axis indicates position in bp relative to the midpoint between TLDR-seq TSSs (see Methods). Subplots show TLDR-seq coverage as in panel D, showing data from Ctrl, RBM7 depleted and ZCCHC8 depleted cells, where reads are split by pA+/pA− status. Also see Supplementary Fig. S7L for the same TLDR-seq data with overlaid DNase-seq coverage.

linkage between transcription initiation and termination sites, splice sites and polyadenylation, and because it can capture pA− RNAs, including lncRNAs and pA+ RNAs simultaneously. In particular, the method will be useful for linking TSS locations and RNA processing events (as in Fig. 7) and characterizing premature RNAs and RNAs that are in the process of being degraded, and more transient RNAs such as eRNAs, especially in combination with relevant protein depletions (as exemplified in Fig. 8, where TLDR-seq was used to characterize PROMPTs and eRNAs). Such characterizations are important since lncRNA classification in general has been based on technical (e.g. length-based) rather than biochemical features that are hard to capture with short-read methods.

Acknowledgements

We thank Hazuki Takahashi at RIKEN, Center for Integrative Medical Sciences, Laboratory for Transcriptome Technology, for protocol discussions; Nadia Laurs Schmidt for excellent technical assistance; and Robin Andersson, Hjorleifur Einarsson, James Bryson, Sarah Rennie, and group members for excellent analysis discussions.

Author contributions: Jamie Auxillos (Conceptualization, Formal analysis, Methodology, Validation, Visualization, Writing—original draft), Arnaud Stigliani (Formal analysis, Visualization, Writing—original draft), Christian Skov Vaagensø(Methodology, Validation, Writing—review & editing), Albin Sandelin (Conceptualization, Writing—original draft, Visualization, Funding acquisition, Supervision), Eivind Valen (Supervision, Writing—review & editing), Adnan Muhammed Niazi (Software, Writing—review & editing), William Garland (Methodology, Validation, Writing—review & editing), and Torben Heick Jensen (Supervision, Writing—review & editing, Funding acquisition)

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

No conflicts of interest.

Funding

Work in the AS group was supported by the Novo Nordisk Foundation [NNF20OC0059951]. Work in the THJ group was supported by Independent Research Fund Denmark—Medical Sciences [Grant 1030-00020B] and the Novo Nordisk Foundation [ExoAdapt Grant 31199]. Funding to pay the Open Access publication charges for this article was provided by Novo Nordisk Foundation [NNF20OC0059951].

Data availability

The R9 TLDR-seq libraries generated for this paper are deposited at GEO under accession number GSE279131. R10 TLDR-seq libraries are deposited at GEO under accession number GSE289428. Third-party data analyzed in the paper are as follows: Data from HeLa cells: CAGE data (GEO GSE147655), RNA-seq data (GEO GSE84172), standard ONT cDNA data (ENA PRJEB44747), and QuantSeq (GEO GSE137612). DNase-seq, H3K4me1/3, and H3H27a bigWig files were downloaded from the ENCODE portal (accession numbers ENCFF977IGB, ENCFF889EXI, ENCFF884DQE, and ENCFF023UNN). Data from mES cells: ChIP data for H3K4me1/3 and H3K27ac (GEO GSE137491). DNase-seq bigWig files were downloaded from the ENCODE portal (accession numbers ENCFF672DJH and ENCFF962TCT). Scripts used for processing and analysis are available at GitHub: https://github.com/ArnaudStigliani/TLDR-Seq.

References

1. Pan Q, Shai O, Lee LJ *et al.* Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 2008;**40**:1413–5. https://doi.org/10.1038/ng.259
2. Yang X, Coulombe-Huntington J, Kang S *et al.* Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* 2016;**164**:805–17. https://doi.org/10.1016/j.cell.2016.01.029
3. Arner E, Daub CO, Vitting-Seerup K *et al.* Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* 2015;**347**:1010–4. https://doi.org/10.1126/science.1259418
4. Danckwardt S, Hentze MW, Kulozik AE. 3′ end mRNA processing: molecular mechanisms and implications for health and disease. *EMBO J* 2008;**27**:482–98. https://doi.org/10.1038/sj.emboj.7601932
5. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74. https://doi.org/10.1038/nature11247
6. Bentley DL. Coupling mRNA processing with transcription in time and space. *Nat Rev Genet* 2014;**15**:163–75. https://doi.org/10.1038/nrg3662
7. Norbury CJ. Cytoplasmic RNA: a case of the tail wagging the dog. *Nat Rev Mol Cell Biol* 2013;**14**:643–53. https://doi.org/10.1038/nrm3645
8. Raghavan V, Kraft L, Mesny F *et al.* A simple guide to de novo transcriptome assembly and annotation. *Brief Bioinform* 2022;**23**:bbab563. https://doi.org/10.1093/bib/bbab563
9. Patro R, Duggal G, Love MI *et al.* Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017;**14**:417–9. https://doi.org/10.1038/nmeth.4197
10. Bray NL, Pimentel H, Melsted P *et al.* Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016;**34**:525–7. https://doi.org/10.1038/nbt.3519
11. Kim D, Paggi JM, Park C *et al.* Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 2019;**37**:907–15. https://doi.org/10.1038/s41587-019-0201-4
12. Kodzius R, Kojima M, Nishiyori H *et al.* CAGE: cap analysis of gene expression. *Nat Methods* 2006; **3**:211–22.
13. Moll P, Ante M, Seitz A *et al.* QuantSeq 3′ mRNA sequencing for RNA quantification. *Nat Methods* 2014;**11**:i–iii. https://doi.org/10.1038/nmeth.f.376
14. Oikonomopoulos S, Bayega A, Fahiminiya S *et al.* Methodologies for transcript profiling using long-read technologies. *Front Genet* 2020;**11**:606. https://doi.org/10.3389/fgene.2020.00606
15. Wulf MG, Maguire S, Humbert P *et al.* Non-templated addition and template switching by Moloney murine leukemia virus (MMLV)-based reverse transcriptases co-occur and compete with each other. *J Biol Chem* 2019;**294**:18220–31. https://doi.org/10.1074/jbc.RA119.010676
16. Grapotte M, Saraswat M, Bessière C *et al.* Discovery of widespread transcription initiation at microsatellites predictable by sequence-based deep neural network. *Nat Commun* 2021;**12**:3297. https://doi.org/10.1038/s41467-021-23143-7

17. Maeng JH, Jang HJ, Du AY *et al.* Using long-read CAGE sequencing to profile cryptic-promoter-derived transcripts and their contribution to the immunopeptidome. *Genome Res* 2023;**33**:2143–55. https://doi.org/10.1101/gr.277061.122
18. Begik O, Diensthuber G, Liu H *et al.* Nano3P-seq: transcriptome-wide analysis of gene expression and tail dynamics using end-capture nanopore cDNA sequencing. *Nat Methods* 2023;**20**:75–85. https://doi.org/10.1038/s41592-022-01714-w
19. Ibrahim F, Oppelt J, Maragkakis M *et al.* TERA-Seq: true end-to-end sequencing of native RNA molecules for transcriptome characterization. *Nucleic Acids Res* 2021;**49**:e115. https://doi.org/10.1093/nar/gkab713
20. Ugolini C, Mulroney L, Leger A *et al.* Nanopore ReCappable sequencing maps SARS-CoV-2 5′ capping sites and provides new insights into the structure of sgRNAs. *Nucleic Acids Res* 2022;**50**:3475–89. https://doi.org/10.1093/nar/gkac144
21. Kruse T, Garvanska DH, Varga JK *et al.* Substrate recognition principles for the PP2A-B55 protein phosphatase. *Sci Adv* 2024;**10**:eadp5491. https://doi.org/10.1126/sciadv.adp5491
22. Yeung PY, Zhao J, Chow EY-C *et al.* Systematic evaluation and optimization of the experimental steps in RNA G-quadruplex structure sequencing. *Sci Rep* 2019;**9**:8091. https://doi.org/10.1038/s41598-019-44541-4
23. Murata M, Nishiyori-Sueki H, Kojima-Ishiyama M *et al.* Detecting expressed genes using CAGE. *Methods Mol Biol* 2014;**1164**:67–85. https://doi.org/10.1007/978-1-4939-0805-9_7
24. Karousis ED, Gypas F, Zavolan M *et al.* Nanopore sequencing reveals endogenous NMD-targeted isoforms in human cells. *Genome Biol* 2021;**22**:223. https://doi.org/10.1186/s13059-021-02439-3
25. Chen Y, Sim A, Wan YK *et al.* Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nat Methods* 2023;**20**:1187–95. https://doi.org/10.1038/s41592-023-01908-w
26. Dobin A, Davis CA, Schlesinger F *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21. https://doi.org/10.1093/bioinformatics/bts635
27. Wu M, Karadoulama E, Lloret-Llinares M *et al.* The RNA exosome shapes the expression of key protein-coding genes. *Nucleic Acids Res* 2020;**48**:8509–28. https://doi.org/10.1093/nar/gkaa594
28. Luo Y, Hitz BC, Gabdank I *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* 2020;**48**:D882–9. https://doi.org/10.1093/nar/gkz1062
29. Garland W, Comet I, Wu M *et al.* A functional link between nuclear RNA decay and transcriptional control mediated by the polycomb repressive complex 2. *Cell Rep* 2019;**29**:1800–11.e6. https://doi.org/10.1016/j.celrep.2019.10.011
30. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 2012;**28**:2184–5. https://doi.org/10.1093/bioinformatics/bts356
31. Raney BJ, Barber GP, Benet-Pagès A *et al.* The UCSC Genome Browser database: 2024 update. *Nucleic Acids Res* 2024;**52**:D1082–8. https://doi.org/10.1093/nar/gkad987
32. Kent WJ, Zweig AS, Barber G *et al.* BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* 2010;**26**:2204–7. https://doi.org/10.1093/bioinformatics/btq351
33. Neph S, Kuehn MS, Reynolds AP *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* 2012;**28**:1919–20. https://doi.org/10.1093/bioinformatics/bts277
34. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucl Acids Res* 1990;**18**:6097–100. https://doi.org/10.1093/nar/18.20.6097
35. Danecek P, Bonfield JK, Liddle J *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* 2021;**10**:giab008. https://doi.org/10.1093/gigascience/giab008
36. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2. https://doi.org/10.1093/bioinformatics/btq033
37. Krause M, Niazi AM, Labun K *et al.* Alignment-free poly(A) length measurement for Oxford Nanopore RNA and DNA sequencing. *RNA* 2019;**25**:1229–41. https://doi.org/10.1261/rna.071332.119
38. Long Y, Jia J, Mo W *et al.* FLEP-seq: simultaneous detection of RNA polymerase II position, splicing status, polyadenylation site and poly(A) tail length at genome-wide scale by single-molecule nascent RNA sequencing. *Nat Protoc* 2021;**16**:4355–81. https://doi.org/10.1038/s41596-021-00581-7
39. Thodberg M, Thieffry A, Vitting-Seerup K *et al.* CAGEfightR: analysis of 5′-end data using R/Bioconductor. *BMC Bioinformatics* 2019;**20**:487. https://doi.org/10.1186/s12859-019-3029-5
40. Takahashi H, Nishiyori-Sueki H, Ramilowski JA *et al.* Low-quantity single strand CAGE (LQ-ssCAGE) maps regulatory enhancers and promoters. *Methods Mol Biol* 2021;**2351**:67–90. https://doi.org/10.1007/978-1-0716-1597-3_4
41. Unlu I, Maguire S, Guan S *et al.* Induro-RT mediated circRNA-sequencing (IMCR-seq) enables comprehensive profiling of full-length and long circular RNAs from low input total RNA. *Nucleic Acids Res* 2024;**52**:e55. https://doi.org/10.1093/nar/gkae465
42. Meola N, Domanski M, Karadoulama E *et al.* Identification of a nuclear exosome decay pathway for processed transcripts. *Mol Cell* 2016;**64**:520–33. https://doi.org/10.1016/j.molcel.2016.09.025
43. Frankish A, Carbonell-Sala S, Diekhans M *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res* 2023;**51**:D942–9. https://doi.org/10.1093/nar/gkac1071
44. Wu P, Schmid M, Rib L *et al.* A two-layered targeting mechanism underlies nuclear RNA sorting by the human exosome. *Cell Rep* 2020;**30**:2387–2401. https://doi.org/10.1016/j.celrep.2020.01.068
45. Carninci P, Sandelin A, Lenhard B *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 2006;**38**:626–35. https://doi.org/10.1038/ng1789
46. Carninci P, Kasukawa T, Katayama S *et al.* The transcriptional landscape of the mammalian genome. *Science* 2005;**309**:1559–63. https://doi.org/10.1126/science.1112014
47. Kocabas A, Duarte T, Kumar S *et al.* Widespread differential expression of coding region and 3′ UTR sequences in neurons and other tissues. *Neuron* 2015;**88**:1149–56. https://doi.org/10.1016/j.neuron.2015.10.048
48. Kiss DL, Oman K, Bundschuh R *et al.* Uncapped 5′ ends of mRNAs targeted by cytoplasmic capping map to the vicinity of downstream CAGE tags. *FEBS Lett* 2015;**589**:279–84. https://doi.org/10.1016/j.febslet.2014.12.009
49. Haberman N, Digby H, Faraway R *et al.* Abundant capped RNAs are derived from mRNA cleavage at 3’UTR G-quadruplexes. *BMC Biology* 2024;**22**:254. https://doi.org/10.1101/2023.04.27.538568
50. Cvetesic N, Leitch HG, Borkowska M *et al.* SLIC-CAGE: high-resolution transcription start site mapping using nanogram-levels of total RNA. *Genome Res* 2018;**28**:1943–56. https://doi.org/10.1101/gr.235937.118
51. Kawaji H, Frith MC, Katayama S *et al.* Dynamic usage of transcription start sites within core promoters. *Genome Biol* 2005;**7**:R118. https://doi.org/10.1186/gb-2006-7-12-r118
52. Kadonaga JT. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip Rev Dev Biol* 2012;**1**:40–51. https://doi.org/10.1002/wdev.21
53. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100. https://doi.org/10.1093/bioinformatics/bty191
54. Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* 2012;**13**:233–45. https://doi.org/10.1038/nrg3163
55. Parry TJ, Theisen JWM, Hsu J-Y *et al.* The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes Dev* 2010;**24**:2013–8. https://doi.org/10.1101/gad.1951110

56. Preker P, Nielsen J, Kammler S *et al.* RNA exosome depletion reveals transcription upstream of active human promoters. *Science* 2008;**322**:1851–4. https://doi.org/10.1126/science.1164096
57. Ntini E, Järvelin AI, Bornholdt J *et al.* Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nat Struct Mol Biol* 2013;**20**:923–8. https://doi.org/10.1038/nsmb.2640
58. Flynn RA, Almada AE, Zamudio JR *et al.* Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *Proc Natl Acad Sci USA* 2011;**108**:10460–5. https://doi.org/10.1073/pnas.1106630108
59. Seila AC, Calabrese JM, Levine SS *et al.* Divergent transcription from active promoters. *Science* 2008;**322**:1849–51. https://doi.org/10.1126/science.1162253
60. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 2008;**322**:1845–8. https://doi.org/10.1126/science.1162228
61. Kim T-K, Hemberg M, Gray JM *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* 2010;**465**:182–7. https://doi.org/10.1038/nature09033
62. Andersson R, Gebhard C, Miguel-Escalada I *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* 2014;**507**:455–61. https://doi.org/10.1038/nature12787
63. Chen Y, Pai AA, Herudek J *et al.* Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters. *Nat Genet* 2016;**48**:984–94. https://doi.org/10.1038/ng.3616
64. Core LJ, Martins AL, Danko CG *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* 2014;**46**:1311–20. https://doi.org/10.1038/ng.3142
65. Nabet B, Roberts JM, Buckley DL *et al.* The dTAG system for immediate and target-specific protein degradation. *Nat Chem Biol* 2018;**14**:431–41. https://doi.org/10.1038/s41589-018-0021-8
66. Lubas M, Christensen MS, Kristiansen MS *et al.* Interaction profiling identifies the human nuclear exosome targeting complex. *Mol Cell* 2011;**43**:624–37. https://doi.org/10.1016/j.molcel.2011.06.028
67. Lubas M, Andersen PR, Schein A *et al.* The human nuclear exosome targeting complex is loaded onto newly synthesized RNA to direct early ribonucleolysis. *Cell Rep* 2015;**10**:178–92. https://doi.org/10.1016/j.celrep.2014.12.026
68. Robinson JT, Thorvaldsdóttir H, Winckler W *et al.* Integrative genomics viewer. *Nat Biotechnol* 2011;**29**:24–6. https://doi.org/10.1038/nbt.1754

Received: October 22, 2024. **Revised:** February 20, 2025. **Editorial Decision:** March 11, 2025. **Accepted:** March 14, 2025