

Perspective

Challenges in identifying mRNA transcript starts and ends from long-read sequencing data

Ezequiel Calvo-Roitberg, Rachel F. Daniels, and Athma A. Pai

RNA Therapeutics Institute, University of Massachusetts Chan Medical School, Worcester, Massachusetts 01605, USA

Long-read sequencing (LRS) technologies have the potential to revolutionize scientific discoveries in RNA biology through the comprehensive identification and quantification of full-length mRNA isoforms. Despite great promise, challenges remain in the widespread implementation of LRS technologies for RNA-based applications, including concerns about low coverage, high sequencing error, and robust computational pipelines. Although much focus has been placed on defining mRNA exon composition and structure with LRS data, less careful characterization has been done of the ability to assess the terminal ends of isoforms, specifically, transcription start and end sites. Such characterization is crucial for completely delineating full mRNA molecules and regulatory consequences. However, there are substantial inconsistencies in both start and end coordinates of LRS reads spanning a gene, such that LRS reads often fail to accurately recapitulate annotated or empirically derived terminal ends of mRNA molecules. Here, we describe the specific challenges of identifying and quantifying mRNA terminal ends with LRS technologies and how these issues influence biological interpretations of LRS data. We then review recent experimental and computational advances designed to alleviate these problems, with ideal use cases for each approach. Finally, we outline anticipated developments and necessary improvements for the characterization of terminal ends from LRS data.

[Supplemental material is available for this article.]

The development of long-read sequencing (LRS) technologies has heralded a new era that advances the ability to interrogate RNA at single-molecule resolution, characterize the full composition of exons on individual mRNAs (Tilgner et al. 2014; Tian et al. 2021; Wright et al. 2022b), and determine the connections between protein-coding and untranslated regions (UTRs) that regulate mRNA functions and expression. Furthermore, the quantification of mRNA isoforms using LRS can provide data on their relative usage across cell types (Au et al. 2013; Sharon et al. 2013; Byrne et al. 2017; Gupta et al. 2018), cellular contexts (Wright et al. 2022b), and disease conditions (Rausch et al. 2023) to inform research related to disease mechanisms and public health (Maestri et al. 2020; Beyter et al. 2021). Thus, LRS has the potential to enable a functional understanding of how and why specific full-length mRNA isoforms are expressed. LRS technologies have been applied for the study of many organisms, spanning from the description of very long bacterial operons (Grünberger et al. 2022) to the uncovering of the full diversity of human mRNAs.

However, there remain challenges in the widespread implementation of LRS for RNA-based applications. Although much focus has been placed on the advantages and challenges of LRS in defining exon composition and structure in mRNA isoforms (Tang et al. 2020; Chen et al. 2023; Pardo-Palacios et al. 2024b), the ability to identify and quantify the terminal ends of mRNA molecules—specifically, the transcript start and end (polyadenylation) sites (TSSs and PASs, respectively)—has been less carefully investigated. Such characterization is crucial for completely delineating the full mRNA molecule and regulatory functions encoded in UTRs.

Herein, we discuss the importance of characterizing mRNA terminal ends with LRS and, through a systematic re-analysis of publicly available LRS data sets, illustrate the limitations of terminal end identification and quantification. We outline experimental and/or computational issues that may underlie these biases and review current approaches that aim to alleviate them. Finally, we provide a long-term perspective on necessary advances to improve LRS in RNA-based applications.

mRNA terminal ends are regions of great biological importance

The biogenesis of eukaryotic mRNA molecules involves multiple molecular mechanisms, each of which can be alternatively regulated. Messenger RNAs are synthesized by RNA Polymerase II (RNAPII), which is recruited to TSS (Juven-Gershon et al. 2008). As nascent RNAs emerge from the RNAPII exit channel, a seven-methyl-guanosine cap is added to protect the 5' end (Galloway and Cowling 2019). Splice sites on the pre-mRNA are recognized by spliceosomal complexes to excise introns, often as the molecule is actively being elongated (Carrocci and Neugebauer 2019; Zhang et al. 2021). Finally, recognition of a polyadenylation signal (PAS) by the RNA cleavage complex initiates cleavage and polyadenylation (addition of a poly(A) tail) to form the 3' end of the mRNA molecule (Tian and Manley 2016). Each of these events—transcription initiation, pre-mRNA splicing, and 3' end cleavage and polyadenylation—can be alternatively regulated to create a multitude of mRNA isoforms from a single gene region (Fig. 1A). Importantly, splicing occurs at specific splice sites on an mRNA (delineated by strong sequence motifs), but there is more stochasticity in the choice of specific TSS and PAS sites, where these sites

Corresponding author: athma.pai@umassmed.edu

Article, supplemental material, and publication date are at <https://www.genome.org/cgi/doi/10.1101/gr.279559.124>. Freely available online through the *Genome Research* Open Access option

© 2024 Calvo-Roitberg et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

are often thought to occur in clusters rather than single positions (Geisberg et al. 2022).

Although much attention has been given to alternative splicing of internal exons, alternative regulation at the 5' and 3' ends of the transcript contributes just as much, if not more, to isoform diversity (Shabalina et al. 2014; Reyes and Huber 2018). There are thought to be more than 37,000 and more than 85,000 known alternative TSSs and PASs, respectively, in the human genome (Wang et al. 2018; Herrmann et al. 2019; Moore et al. 2022), with an average of five and four alternative TSSs and PASs per gene, respectively (Wang et al. 2018; Abugessaisa et al. 2021). Alternative terminal ends can expand or truncate protein sequences, but more frequently, they do not change coding regions and in-

stead determine the composition of the 5' and 3' UTRs of an mRNA molecule. UTRs play important roles in post-transcriptional regulation by containing sequence features that influence mRNA stability, localization, and translation efficiency (for review, see Di Giammartino et al. 2011). TSS usage delineates the 5' UTR, which is involved in ribosome loading and translation efficiency and can influence the expressed open reading frame (Leppek et al. 2018).

At the other end, alternative PASs dictate the 3' UTR, which is involved in the regulation of mRNA stability and localization and can also lead to protein truncation (Di Giammartino et al. 2011; Lianoglou et al. 2013; Taliaferro et al. 2016; Alfonso-Gonzalez and Hilgers 2024). Furthermore, lack of proper 3' end cleavage

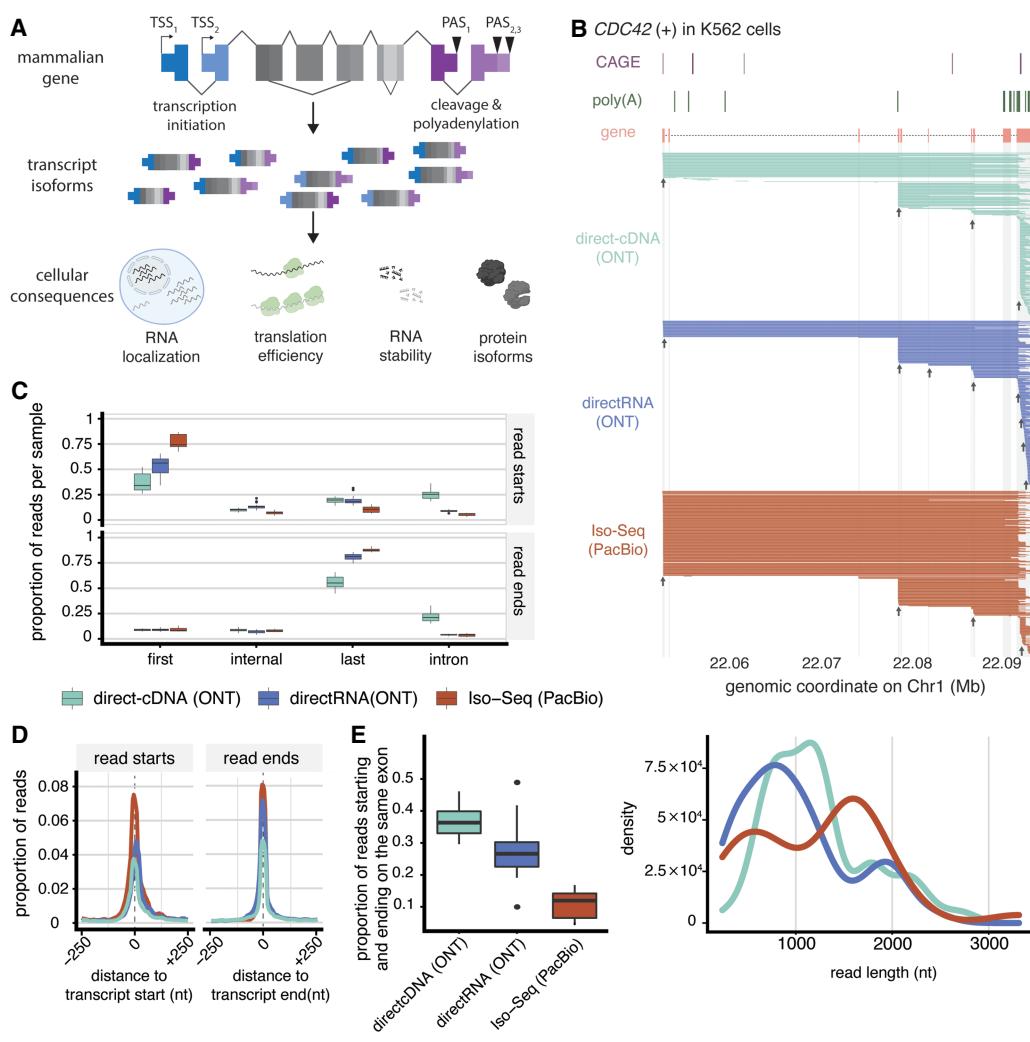


Figure 1. LRS reads show large variability in start and end coordinates. (A) Mammalian genes can have multiple alternative transcript starts (TSSs) and ends (PASs), leading to different isoforms expressed from the same gene with variable cellular consequences. (B) Representative example of long-read RNA sequencing reads for *CDC42* in K562 cells. To aid visualization, 280 randomly sampled reads from each sequencing technology are shown. At the top are annotated features of *CDC42*, including CAGE peaks (violet), poly(A) peaks (dark green), and annotated exons (orange). At the bottom are read coverage plots, in which each horizontal line represents the span between the first and last coordinates of a read, for the ONT direct-cDNA (teal) (Chen et al. 2021), ONT directRNA (blue) (Chen et al. 2021), and PacBio Iso-Seq (red) (Luo et al. 2020) data sets (Supplemental Methods; Supplemental Table S1). Arrows mark TSS clusters. The bottom panel shows the distribution of read length across sequencing technologies for *CDC42*. (C) Distributions of the proportion of reads that start (top) or end (bottom) in HITIndex-classified (Fiszbein et al. 2022) first, internal, or last exons or introns across from three LRS technologies, using data from the A549, Hct116, HepG2, K562, and MCF-7 cell lines. (D) The distribution of read starts and ends around annotated transcription start (left) or end (right) sites across three LRS technologies. The y-axis represents the proportion of reads per sample, calculated using a sliding window of 0.01 kb around the feature. (E) Distributions of the proportion of single-exon reads that start and end within the same exonic feature across three LRS technologies. Note that B–E use published data sets, and the metrics presented here may change when estimated with other data sets or protocols.

induced by cellular stresses or other perturbations can cause transcriptional readthrough, which alters the 3' ends of mRNAs (for reviews, see Vilborg and Steitz 2017; Rosa-Mercado and Steitz 2022). Thus, through a mix of effects on protein composition, biogenesis, and levels, alternative terminal ends may significantly impact protein diversity (Wang et al. 2018; Abugessaia et al. 2021; Carrión et al. 2023).

Although the regulation of alternative TSSs, splicing, or PASs on an mRNA have all historically been thought to occur independently, increasing evidence suggests extensive coordination to regulate the expression of full-length isoforms (Alfonso-Gonzalez and Hilgers 2024). For instance, recent studies have identified coupling between alternative TSSs and splicing (Anvar et al. 2018; Fiszbein et al. 2019; Alfonso-Gonzalez et al. 2023), splicing and alternative TESs (Anvar et al. 2018; Alfonso-Gonzalez et al. 2023), and alternative TSS and PAS mechanisms (Anvar et al. 2018; Alfonso-Gonzalez et al. 2023; Calvo-Roitberg et al. 2024). Furthermore, alternative splice sites or PASs may compete for usage, and splicing or 3' polyadenylation decisions may influence each other within an mRNA molecule (Coulon et al. 2014; Reimer et al. 2021; Choquet et al. 2023; Zhang et al. 2023). The coordination between these processes highlights the need to move beyond characterization of individual exons toward studying alternative RNA processing events in the isoform context. Therefore, LRS provides an attractive approach for studying full-length RNA molecules.

Challenges of terminal end identification with long-read RNA sequencing

Short-read sequencing (SRS) technologies are the gold standard for high-throughput transcriptome profiling (Wang et al. 2009). RNA sequencing (RNA-seq) has been used to investigate gene expression (The ENCODE Project Consortium 2012), canonical and alternative mRNA splicing (Wright et al. 2022a), noncoding RNAs, and post-transcriptional modifications (Jia et al. 2020) among many other applications. However, SRS technologies are usually limited by their inability to sequence more than 100–300 contiguous nucleotides (nt). Although SRS offers robust and accurate solutions for detecting and characterizing local features within the sequenced fragment, these data cannot inform exon connectivity across a full isoform. The average transcript length in most species is much longer than the average SRS read length (i.e., human transcripts are ~2000 nt on average), leading to computational approaches to infer full transcript levels and compositions (Trapnell et al. 2010; Bray et al. 2016; Patro et al. 2017; Hölzer and Marz 2019). These methods often anchor their analyses on known sequences or databases of previously observed transcripts. Therefore, although they succeed at gene expression quantification (in which gene boundaries are well identified) and transcript quantification of known isoform structures, these approaches cannot accurately identify and quantify new isoforms that involve previously uncharacterized connections between exons (Conesa et al. 2016). This is especially problematic for the identification and quantification of exons at the terminal ends of transcripts because there are no common features to define reads from these regions (Qin et al. 2018; Cass and Xiao 2019; Fiszbein et al. 2022). The challenge of characterizing mRNA terminal ends in SRS data is also magnified by fragmentation and size selection steps during library preparation, which lead to an “edge-effect” in which there is a depletion of reads from the ends of transcripts (Carralot et al. 2012; Birol et al. 2015).

LRS technologies promise the ability to directly sequence full-length mRNA molecules without fragmentation and subsequent computational inference of the relationships between individual exons. The LRS technology space is currently dominated by Pacific Biosciences (PacBio; <https://www.pacb.com/products-and-services/applications/rna-sequencing/> [accessed April 25, 2024]) and Oxford Nanopore Technologies (ONT) (Box 1). PacBio utilizes single-molecule real-time (SMRT) sequencing, in which circularized cDNAs are sequenced by detecting pulse waves resulting from the excitation of fluorescent dNTPs during the synthesis of a complementary strand (Iizuka et al. 2022). In contrast, ONT nanopore motors directly pull molecular fragments (double-stranded cDNA or “native” RNA) through a pore as a constant electrical current is applied. Each nucleotide (or modified nucleotide) within the stretch of nucleotides in the nanopore results in a characteristic voltage signal, allowing the sequencing of a molecule without requiring a synthesis step (Pugh 2023).

Although advances in LRS methods have improved accuracy and consistency in sequencing of full-length molecules in recent years, these techniques still face technical and computational hurdles, which vary according to LRS technology (Pardo-Palacios et al. 2024b). These include relatively low genomic coverage (compared with short-read technologies), which limits the annotation or quantification of lowly expressed genes or isoforms (Boldogkői et al. 2019), and persistent concerns regarding insertion, deletion, or misalignment errors, leading to problems with the accurate identification of single-nucleotide variations (SNVs) and internal splice sites (Amarasinghe et al. 2020; Cui et al. 2020; Dohm et al. 2020; Mikheenko et al. 2022). These challenges affect all applications of LRS for RNA profiling, including the characterization of mRNA terminal ends. Although analytical tools have been proposed to correct for some of these challenges when characterizing internal exons (Amarasinghe et al. 2020; Tang et al. 2020; Chen et al. 2023), it remains unclear how these and/or other technical challenges may uniquely impact terminal end characterization. Although low accuracy of LRS reads may affect the ability to identify precise mRNA 5' and 3' ends, which, unlike splice sites, are not delineated by common sequence elements (Birol et al. 2015; Arefeen et al. 2019; Dudnyk et al. 2024), this is not likely to be the major challenge for terminal end identification (as reviewed below).

Previous studies have observed decreased LRS read coverage toward the ends of mRNA molecules, suggesting limited power to characterize terminal ends (Wright et al. 2022b). Moreover, the read start and end positions, denoting TSSs and transcript end sites, are highly variable across reads (Ibrahim et al. 2021). This is evident when looking at aligned reads for an individual gene across LRS approaches and platforms. As a representative example, there is large variability in the number of TSSs identified across high-coverage LRS data sets for the *CDC42* gene in K562 cells (Supplemental Methods; Supplemental Code; Supplemental Table S1; Fig. 1B), with four, eight, and four start site clusters supported by at least 10 read starts that are <50 nt away from each other across the direct-cDNA, directRNA (Chen et al. 2021), and Iso-Seq (Luo et al. 2020), respectively. We observe less variability across methods in transcript end site identification, with three, four, and three end site clusters for *CDC42* in K562 cells in the direct-cDNA, directRNA, and Iso-Seq data sets, respectively. Notably, especially for read starts, many of the identified terminal positions do not overlap with the terminal coordinates of annotated *CDC42* isoforms or with empirically derived 5' or 3' sites (using CAGE and PAS-seq peaks, respectively). This is consistent with early reports

Box 1. Overview of current long-read RNA sequencing technologies

Two main companies, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) offer commercial solutions for long-read RNA sequencing.

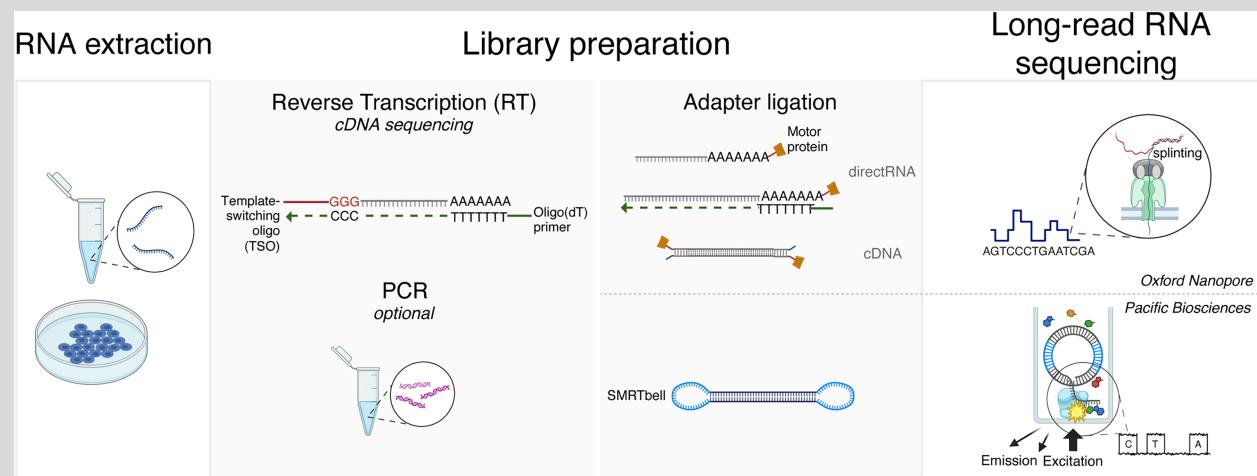
PacBio. In this technology, a single-stranded circularized cDNA molecule is generated by ligating hairpin adaptors to double-stranded cDNA (SMRTbell library preparation) and applied to small wells embedded in proprietary SMRT cells. Each well contains a single immobilized ϕ 29 DNA polymerase molecule and a pool of phospholinked (four-color) fluorescent dNTPs. Single-molecule sequencing is performed by reading fluorescent signals produced in real-time through base-pairing events between nucleotides in the template strand with complementary dNTPs. Correct base-pairing results in a longer dwell time in the polymerase, corresponding to a characteristic pulse wave detected by the instrument during this synthesis phase. PacBio's high-fidelity long-read (HiFi) technology uses an enzyme with processivity greater than the insert length; thus, the circularized template is sequenced multiple times to obtain circular consensus reads with high accuracy when combining information across iterative sequencing instances.

ONT. In this technology, nanopores are embedded in an electro-resistant membrane and suspended between two chambers with ionic solutions, through which an electric current is passed. Single- or double-stranded molecules with ligated adaptors are bound to an enzyme at the mouth of the pore that "unzips" the strands as applied electric potential causes a single strand to be electrokinetically pulled through the pore. Stretches of 10–15 nt travel through the pore, causing a temporary blockage of the current, with electrical signatures that are specific to each nucleotide (or modified nucleotide). The RNA applications of ONT include direct RNA sequencing (which preserves base modification information lost during traditional cDNA synthesis), in which the RNA molecules are either ligated directly to the RNA molecule or subjected to "splinting" using single-strand reverse transcription to generate a hybrid DNA-RNA molecule that is more efficiently processed by the pore to read the RNA strand of this molecule (Garalde et al. 2018). ONT can also be used to sequence cDNA molecules (Grünberger et al. 2022; Wongsurawat et al. 2022).

LRS library preparation approaches. RNA isolation and purification are shared between all LRS sequencing methods. Similarly, library construction involving cDNA synthesis is fundamentally similar between ONT and PacBio. Reverse transcription is initiated by primers that enrich for specific target RNA populations, commonly either oligo(dT) or random hexamers, depending on the specific application. Oligo(dT) primers enrich for mature RNAs transcribed by RNAPII by annealing to the polyadenylated tail present at the 3' end of most eukaryotic mRNA molecules. Random hexamers anneal to various sites along the RNA molecules, allowing for reverse transcription of the entire RNA population starting from different sites.

The reverse transcriptase enzyme then uses the RNA molecule as a template to synthesize a cDNA strand by extending the primer to synthesize the first strand of cDNA (fs-cDNA). Long-read library preparation methods commonly use template-switching reverse transcriptase enzymes, which add a nontemplated "CCC" triplet at the 5' end of the fs-cDNA. A template-switching oligo (TSO) with a complementary "GGG" can then bind the "CCC" and allow the reverse transcriptase enzyme to immediately begin synthesizing a double-stranded second-strand cDNA (ss-cDNA) using the first cDNA strand as a template.

Following cDNA synthesis, technology-specific adapters are ligated to the completed molecules. For ONT, these adapters include a sequencing adapter attached to a motor protein that drives the translocation of the cDNA molecule through the pore. The adapter additionally concentrates DNA substrates at the membrane surface near the nanopore to increase DNA capture. The SMRTbell adapters in PacBio technology include two adapters, each of which includes a double-stranded stem and a single-stranded hairpin loop. The two adapters may have identical (symmetric) or different (asymmetric) sequences. The ligation of these adapters to double-stranded cDNA allows the formation of a single-stranded circular DNA template that is read by the embedded polymerase.



Box Figure 1. Overview of current long-read RNA sequencing technologies. Depiction of RNA preparation, library preparation, and sequencing steps for ONT and PacBio technologies. On the left are shared RNA extraction and library preparation steps, in which reverse transcriptase is only required for the cDNA sequencing methods (both PacBio and ONT) and PCR is required for PacBio but optional for ONT. On the right are depictions of the varying adapter ligation and sequencing for ONT (top) and PacBio (bottom) methods.

that only 60% of PacBio read starts and ends map to annotated first and last exons, respectively (Sharon et al. 2013). Although it is very likely that the reference databases do not contain all alternative terminal sites (Tang et al. 2020; Wyman et al. 2020; Tian et al. 2021), this cannot explain the large variability between different

LRS approaches. For techniques that use adapter sequences to tag the 5' and 3' ends of molecules during library preparation (see below), one report in the nascent data observed that as few as 7% of reads actually contain these adapters (Reimer et al. 2021). Additionally, in an approach that uses both 5' and 3' adapters,

80% of the 3' ends have the correct adapter, but only as many as 32% of the 5' ends have the correct adapter (Ibrahim et al. 2021). When molecules have the 5' adapter, they tend to also have the 3' adapter. Similar analyses looking for nontemplated 3' adenosines have seen that reads without poly(A) tails often do not match known 3' ends (Wright et al. 2022b).

More systematically, when evaluating how many read start and end positions occur in empirically derived first, internal, or last exons, we see that a substantial proportion of read starts and ends do not fall within the expected first or last exons, respectively (Fig. 1C). This is especially true for read starts in the ONT approaches, in which >50% and >40% of read starts in the direct-cDNA and directRNA libraries, respectively, do not fall within first exons. Consistently, for ONT methods, we find that a higher proportion of read starts are found within last exons. Similar trends are observed for read start and end distributions around annotated TSSs or PASs. We see that all methods have a slight shift toward read start positions downstream from annotated TSSs, with a higher density of read starts centered around the annotated TSSs for PacBio Iso-Seq data (Fig. 1D). The positioning of read ends for all sequencing techniques is much more tightly distributed around annotated PAS, consistent with previous observations that although the 3' ends of LRS RNA-seq data tend to fall within ~5 nt of an annotated transcript end, there is a wider distribution (~15–100 nt) of the 5' end positioning around annotated TSSs (Sharon et al. 2013). A final surprising observation is that many LRS reads appear to be derived from a single exon. For instance, 36%, 27%, and 10% of reads across the direct-cDNA, directRNA, and Iso-Seq data sets, respectively, exhibited no splicing and were fully contained within the 3' most exon of *CDC42*, despite no annotated single-exon isoforms. More broadly, >25% of reads generated by ONT methods can be classified as single-exon reads (Fig. 1E). Given the prevalence of these types of reads, recent computational approaches have included a “monoexon” category that classifies putative transcripts assembled from these reads as incomplete (Pardo-Palacios et al. 2024a). Although it is not possible to rule out that LRS approaches may be especially sensitive to uncover extremely widespread stochasticity in terminal site precision and novel isoform usage driven by alternative terminal ends, it is more likely that LRS reads often have spurious terminal ends.

Altogether, these observations suggest widespread 5' truncation in LRS reads, particularly in those from ONT approaches (Mikheenko et al. 2022). A relatively low proportion of read terminal ends across LRS approaches fall in introns; thus, these reads likely represent mature mRNA molecules rather than DNA contamination, and 5' truncation may occur during library preparation, sequencing, or processing of the data. The presence of truncation is also supported by the relatively shorter reads generated by ONT directRNA and direct-cDNA approaches (Pardo-Palacios et al. 2024b). However, it is still difficult to disambiguate whether these terminal end inconsistencies occur owing to technical or biological causes. One option to better understand these observations is to leverage biological expectations, in which single-exon genes without any alternative isoform regulation should be less likely to have variability in 5' end read positions. Incomplete LRS coverage is positively correlated with the length of single-exon genes, such that longer single-exon genes show more variability in 5' read position (Workman et al. 2019). Similarly, transcripts >2 kb in length tend to have lower coverage in ONT directRNA sequencing data sets (Soneson et al. 2019; Workman et al. 2019). To probe technical biases, Mikheenko et al. (2022) devised a strategy to use unique molecular identifiers (UMIs) to tag individual mole-

cules during reverse transcription (RT) and sequence the same cDNA library with PacBio and ONT approaches. This allowed the direct assessment of terminal end identification across platforms. After stringent filtering for reads within 100 nt of an annotated poly(A) site or CAGE peak and at least one read within a set of reads sharing a UMI being assigned directly to a TSS or PAS, the authors observed that 95% of PacBio and ONT reads had the same 3' end, but only 87% had the same 5' end. The representation of known and novel isoforms in LRS data is influenced by both sample preparation and sequencing technology. For example, technologies with limited read length may fail to capture the full diversity of transcript ends, whereas low read depth can hinder the detection of low-abundance isoforms (Pardo-Palacios et al. 2024b). As LRS technologies continue to advance, it is important to balance these considerations when selecting an appropriate method.

mRNA terminal ends are affected by LRS experimental biases

Recent observations indicate that LRS reads exhibit truncation at the 5' and, to a lesser extent, the 3' ends of mRNA transcripts (Depledge et al. 2019; Workman et al. 2019; Ibrahim et al. 2021; Mikheenko et al. 2022). Thus, there is a general shift in the distribution of LRS read lengths toward shorter lengths relative to the presumed true distribution of mRNA lengths (Mikheenko et al. 2022). Many experimental steps could introduce biases in LRS read lengths and, thus, lead to incorrect identification and quantification of mRNA terminal ends. Below, we explore how general RNA handling and extraction practices, library preparation steps, and inherent sequencing biases can contribute to these effects.

Several aspects of RNA handling before and during library preparation can significantly impact the representation of full-length mRNA molecules in LRS data (Fig. 2A). Common RNA extraction and clean-up procedures using either purification columns or nucleic acid affinity beads lead to a bias in the size of RNA molecules retained, as these procedures deplete very short (<200 nt) molecules or selectively enrich for a particular length distribution depending on the ratio of nucleic acids to affinity beads (Shi et al. 2021). These biases persist throughout library preparation, affecting the length distribution of cDNA molecules during the clean-up and downstream steps (Byrne et al. 2019). Furthermore, given the fragility of RNA molecules, they are highly subject to random fragmentation during any molecular manipulations (Fig. 2A). This fragmentation can be caused by mechanical (e.g., shearing during pipetting), enzymatic (e.g., nuclease contamination), or chemical (e.g., guanidinium) factors, all leading to systematic shortening of RNA or cDNA molecules (Byrne et al. 2019; Amarasinghe et al. 2020). Because random fragmentation should equally affect both ends of RNA or cDNA molecules, the increased prevalence of 5' truncation in LRS reads suggests that other causes might be the primary drivers of biased representation of mRNA terminal ends. For instance, in methods that use the 3' end to synthesize cDNA or initiate sequencing, molecules truncated near the 3' end would be underrepresented in the final sequencing data set, thus accentuating the 5' truncation bias.

Limitations during RT

Most LRS protocols involve RT at some stage during library preparation. However, RT enzymes are inherently limited in their ability to successfully generate full-length cDNA reflective of the full-length mRNA molecule (Fig. 2A; Minshall and Git 2020; Verwilt et al. 2023). In particular, the processivity of RT enzymes (the number of nucleotides an RT enzyme can incorporate into a cDNA

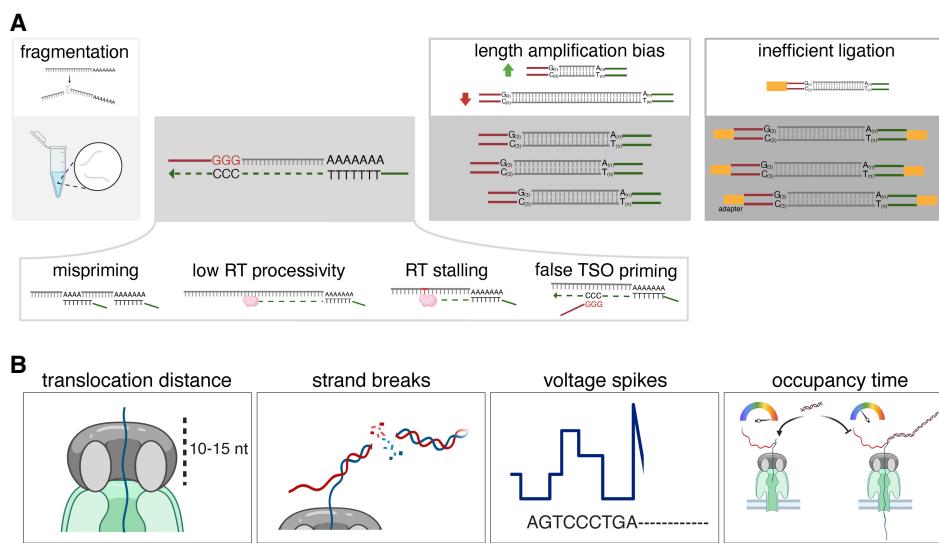


Figure 2. Technical biases affecting the accuracy of terminal end sequences. (A) Multiple events during RNA handling and library preparation can introduce biases that influence the accuracy of terminal ends in LRS reads. (B) Nanopores present unique inherent hardware and software challenges that affect accurate sequencing, especially at the terminal ends of molecules.

strand before dissociating from the RNA template) is often not sufficient to reach the true 5' end of an mRNA molecule. RT processivity highly varies, and typical RTs can synthesize 5–7 kb amplicons under standard conditions (Martín-Alonso et al. 2021), but more persistent enzymes have been developed recently (e.g., Marathon [Zhao et al. 2018] and TGIRT [Mohr et al. 2013]) (for review, see Martín-Alonso et al. 2021; Osorbin and Filipenko 2021). Limited processivity and/or stalling of the enzyme can lead to premature termination of cDNA synthesis and truncated molecules with 5' ends that do not represent the true mRNA 5' end (Wyman et al. 2020).

RT processivity can be impacted by many factors, including sequence (i.e., stretches of GC-rich regions) (Minshall and Git 2020), RNA secondary structure (Verwilt et al. 2023), and RNA modifications (Kellner et al. 2010). Mispriming can also occur during RT, leading to the representation of erroneous 3' ends. Specifically, oligo(dT) primers, which are designed to anneal to the poly(A) tail, can bind to and prime off of internal stretches of adenines within the mRNA, resulting in cDNA 3' ends that fall upstream of the true poly(A) site (Nam et al. 2002; Balázs et al. 2019; Wyman et al. 2020). This issue can be mitigated by the ligation of a 3' adapter ligation step before RT (see below). At a lower frequency, mispriming can also occur during template-switching RT, whereby template-switching oligos can be mispositioned owing to an internal CCC sequence, leading to erroneous 5' ends with an adapter that suggests complete RT (Cocquet et al. 2006; Tang et al. 2013; Moldován et al. 2018). In addition, the sequence or chemical modification at the 5' end of mRNAs may affect the efficiency of template switching (Wulf et al. 2019, 2022). However, two pieces of evidence support the idea that RT processivity or mispriming only contributes in part to the inaccurate representation of terminal sites. First, the average LRS read length is slightly shorter than the average molecule length of cDNA after RT (Soares et al. 2022; Shiau et al. 2023). Second, there is a higher prevalence of 5' truncations in directRNA ONT sequencing data, which have an optional RT step relative to cDNA sequencing methods, but there is no ob-

servable difference in the prevalence of 3' truncations (Pardo-Palacios et al. 2024b).

Limitations during library preparation

The process of converting RNA molecules to cDNA molecules involves many molecular steps and nucleic acid clean-ups during which fragmentation and size selection could impact the length of retained RNA or cDNA molecules. The largest shift in length, however, is likely introduced by polymerase chain reaction (PCR) steps at the end of library preparation, which favors the amplification of shorter fragments (Fig. 2A; Shagin et al. 1999; Parekh et al. 2016). This bias should again equally affect misrepresentation of both the 5' and 3' ends. However, because PCR involves priming off adapter sequences that are either added to RNA molecules during RT or ligated to cDNA molecules, any biases in the positioning of these adapters will be propagated during PCR. Further, any molecule that does not include an adapter sequence at one or both ends will fail to be amplified, which is more likely to negatively impact longer RNA molecules that were incompletely reverse-transcribed (Fig. 2A).

Limitations during sequencing

During both nanopore and well-based sequencing methods, shorter molecules are more likely to get sequenced because they clear the pore or pass through the polymerase more often (Fig. 2B; Baslan et al. 2021). This equally affects molecules with truncated 5' or 3' ends but leads to increased coverage of shorter molecules with truncated ends, propagating all the biases described above. Furthermore, DNA strands might be broken as they are translocated through the pores (particularly during the mechanical forces applied during ONT sequencing), resulting in truncation events that are more likely to occur at the 5' end, because molecules are sequenced in the 3'-to-5' direction (Fig. 2B; Workman et al. 2019). After 36 h of sequencing, 5% of mitochondrial RNA reads show decreased length, suggesting that strand breaks are not very abundant (Workman et al. 2019). The characteristics of

nanopores may affect the accuracy of terminal ends by influencing the translocation speed at which nucleic acid molecules, both RNA or cDNA, pass through the pore. In ONT sequencing, the motor protein that facilitates molecular translation is positioned 10–15 nt away from the terminal end, resulting in the final nucleotides being pulled through the pore rapidly as they are released by the enzyme (Fig. 2B). This rapid movement limits the ability to accurately identify these final nucleotides (Workman et al. 2019; Parker et al. 2020; Ibrahim et al. 2021).

Basecalling inaccuracies during nanopore sequencing are also influenced by motor protein stalling or voltage spikes (Workman et al. 2019). Motor protein stalling occurs when the protein responsible for translocating the nucleic acid molecule through the nanopore encounters obstacles or excessive resistance, causing it to pause or terminate sequencing (White et al. 2023). Similarly, voltage spikes in the electrical signal as the RNA or cDNA molecule travels through the pore can lead to basecalling issues (Fig. 2B). These sudden fluctuations can be caused by modified nucleotides and/or background electrical noise inherent to the system and can be misinterpreted as the end of the molecule, thus terminating basecalling (Jain et al. 2022). Both stalling and voltage spikes can result in reads with truncated terminal ends, with a bias toward more 5' truncation in direct RNA 3'-to-5' sequencing. Although these issues are widespread, how they are generated remains unknown. In particular, it is unclear to what extent these are random events or triggered by specific molecular features.

Experimental approaches to enrich for full-length RNA molecules

In light of the difficulties involved in natively sequencing full-length RNA molecules, many experimental approaches have been developed to tackle this challenge. These approaches broadly focus on enriching for or tagging 5' and 3' ends of RNA molecules to ensure full-length sequencing.

In principle, nanopore directRNA sequencing allows for full-length sequencing by bypassing terminal end or length biases created during library preparation. Additionally, directRNA sequencing can identify RNA modifications (e.g., the 5' cap at the terminal end) (Mulroney et al. 2022; Ugolini et al. 2022) that are lost during cDNA synthesis and PCR. However, directRNA sequencing has the highest amount of 5' truncation across LRS technologies, likely owing to issues during sequencing as outlined above. This suggests that biases generated during library preparation steps may not be the primary cause of 5' truncation in LRS data.

Enrichment of longer molecules

As discussed above, one of the major biases in LRS is the increased probability of sequencing shorter molecules, which likely amplifies any terminal truncation effects. One strategy proposed to mitigate this issue is to enrich for longer RNA molecules before and during library preparation (Fig. 3A; Alfonso-Gonzalez et al. 2023). However, this enrichment does not entirely supersede other truncation effects created during library preparation or

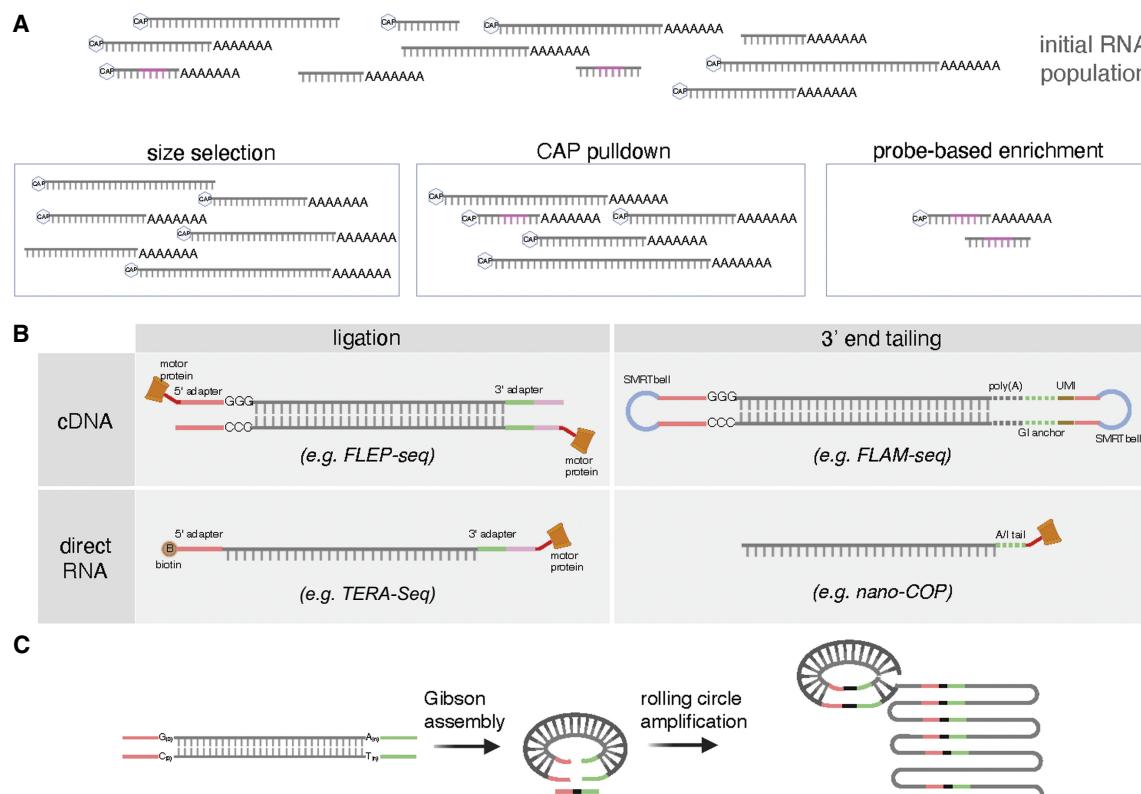


Figure 3. Experimental strategies to enrich for full-length RNA molecules in LRS data. (A) Methods that specifically enrich for 5' ends or full-length molecules (based on size or sequence) can be used to increase the chances of sequencing full-length mRNAs. (B) Methods that add known sequences to the terminal ends of mRNAs by ligation (left) or tailing (right) are an unbiased way to recognize full-length molecules in both cDNA (top) and directRNA (bottom) sequencing. (C) Circularization of mRNAs or cDNAs during library preparation allows for multiple rounds of terminal adapter and target molecule sequencing.

sequencing. The enrichment of predetermined sizes of RNA molecules may result in reads that do not represent the true distribution of molecular lengths present in the original RNA population. Thus, this process can introduce its own biases, and biological conclusions drawn from the data should be considered carefully.

Selective enrichment of terminal ends

To enrich for full-length RNAs, several approaches selectively isolate molecules with intact terminal end markers. RT with oligo (dT) priming targeting a poly(A) tail (as described above) represents one such strategy to selectively carry forward polyadenylated molecules. Strategies to enrich for intact 5' ends usually involve capturing or tagging molecules with a 5' cap (Fig. 3A). 5' Ends can be pulled down using antibodies targeting cap modifications or a cap-binding protein (Ogami et al. 2023) or following biotinylation of the 5' end (Jiang et al. 2019; Parker et al. 2020; Carbonell-Sala et al. 2024), with 5' and 3' adapter ligations that recognize the cap and poly(A) tails, respectively (e.g., CapTrap-seq) (Carbonell-Sala et al. 2024). Alternatively, the 5' cap can be enzymatically replaced with a biotinylated RNA adapter, followed by pulldown of these cap-containing RNAs and RT primed off of a poly(A) tail to ensure full-length molecules (e.g., 5'-Cap capturing) (Jiang et al. 2019). These approaches generally generate high-quality reads, but analysis of synthetic RNAs showed fragmentation during 5' enrichment, leading to reduced read coverage in the middle of molecules (Pardo-Palacios et al. 2024b). This may impact the characterization of terminal ends in the context of full-length isoforms and, in particular, the discovery and quantification of alternatively spliced isoforms (Maeng et al. 2023).

In principle, another enrichment-based approach to increase the sequencing of full-length isoforms involves selectively isolating transcripts from specific genes using complementary biotinylated probes (Fig. 3A; e.g., RNA capture long-read sequencing [Lagarde et al. 2017] and others [Sheynkman et al. 2020; Wang et al. 2023; Zhang et al. 2023]). This increases the coverage (number of reads) for specific genes and would probabilistically increase the number of full-length molecules sequenced. However, these libraries would still suffer from all truncation issues mentioned above without any tags or selective enrichment of terminal ends.

Sequence-based tagging of terminal ends

Several approaches add known adapter sequences to the 5' and/or 3' ends of RNA molecules to enrich for and identify full-length cDNA or RNA molecules (Fig. 3B). The simplest version is the use of a template-switching RT with a known 5' TSO sequence, as implemented in commercial kits aiming to sequence polyadenylated RNA (e.g., TeloPrime [Balázs et al. 2019], SMARTer cDNA amplification [Cartolano et al. 2016; Legnini et al. 2019]). A more flexible approach that is agnostic to polyadenylation status involves ligation of an adapter to the 3' end of the RNA molecule, followed by a template-switching RT with a known TSO sequence (e.g., FLEP-seq) (Herzel et al. 2018; Reimer and Neugebauer 2020; Long et al. 2021). However, because ligation is often inefficient, complementary approaches often use 3' end guanosine and/or inosine tailing of the RNA molecule to synthesize a tail that can be used to prime a template-switching RT reaction (e.g., FLAM-seq) (Legnini et al. 2019). Approaches with similar conceptual frameworks have also been developed for directRNA nanopore sequencing. For instance, it is possible to directly ligate 5' and/or 3' adapters onto an RNA molecule (agnostic of polyadenylation status), which can be used to mediate ligation of the ONT sequencing adapters (e.g., TERA-

Seq) (Ibrahim et al. 2021). Alternatively, 3' end adenine or inosine tailing of the RNA molecule can be used to mediate adapter ligation directly to RNA (e.g., nano-COP) (Drexler et al. 2021).

Together, these methods all add at least one terminal tag that can be used to confirm full-length sequencing, which allows for unbiased terminal end identification and quantification without relying on independent terminal site annotations. Many of these methods can address the ONT translocation issue that prevents accurate sequencing of the final 10–15 nt by ligating longer adapter sequences. Longer adapter sequences are also helpful to prevent potential internal mispriming events at either the 5' or 3' end as described above. The use of longer adapters or in vitro nucleoside tailing provides a more defined target for primer binding with higher fidelity and enhances the specificity of terminal end identification, particularly at the 3' end (Ibrahim et al. 2021). Finally, any fragmentation or length biases that occur during RT (i.e., owing to RT processivity limitations for distances longer than ~10 kb) (Zhao et al. 2018) would be propagated during these sequence-based terminal end tagging approaches. Similarly, these methods would still suffer from all of the artifacts introduced during sequencing (e.g., strand breaks) that lead to 5' truncation of reads, although only retaining reads with a 5' adapter would increase the probability that reads delineate true 5' RNA ends.

Sequencing of full-length molecules after circularization

Circularizing molecules after the addition of terminal adapters protects the ends from exonuclease degradation and enhances the accuracy of terminal-end sequencing by directly colocalizing 5' and 3' ends of molecules (Wang et al. 2020; Scarano et al. 2024). This is routinely done during PacBio library preparation with the addition of SMRTbell adapters that enable circularization and continuous sequencing of the circle (<https://www.pacb.com/products-and-services/applications/rna-sequencing/> [accessed April 25, 2024]). Circularization of full-length cDNA followed by rolling circle amplification to generate multiple concatemerized sequences of the same molecule enables a similar method to be applied with ONT sequencing (e.g., R2C2) (Fig. 3C; Volden et al. 2018). These methods help to both increase read length (thus decreasing truncation effects) and reduce basecalling errors because circularization allows each molecule to be sequenced multiple times. Although these data often achieve some of the highest RNA read lengths (Pardo-Palacios et al. 2024b), these approaches are still limited by the previously mentioned biases that affect terminal end truncation.

Computational approaches to identify and quantify full-length isoforms

Despite the development of experimental approaches to more accurately sequence full-length RNA molecules, there are still analytical hurdles to identifying and quantifying complete isoforms. Thus, many computational approaches have been developed to account and/or correct for technical biases (Table 1). Although most of these pipelines were specifically designed to analyze alternative splicing events using LRS, we will focus on how they consider constitutive and alternative terminal ends.

Overall, computational analyses have shown that LRS data can improve the identification of new isoforms and new terminal ends, especially in poorly annotated species. For instance, Iso-Seq data from stickleback fish found that the identified isoforms began

Challenges in identifying RNA terminal ends

Table 1. Computational approaches to identify and quantify mRNA terminal ends in LRS data

Approach	Known annotations	De novo identification	Terminal end selection	Orthogonal features	Full-length status	Reference
TAPIS	Optional	Clustering	Longest isoform with shared splice sites	Terminal end A's, terminal end adapters	No	(Abdel-Ghany et al. 2016)
StringTie2	Optional	Graph-based	Unknown	No	No	(Kovaka et al. 2019)
FLAIR	Yes	Clustering	Read density	Terminal end adapters	Read level	(Tang et al. 2020)
TALON	Yes	Clustering	Clustering	Terminal end A's	Transcript level, FSM/ISM/NIC categories	(Wyman et al. 2020)
FLAMES	Yes	Clustering	Annotations	SRS RNA-seq	No	(Tian et al. 2021)
Bambu	Yes	Probabilistic ML model	Categorization and modeling	Terminal end A's	Read level and transcript level	(Chen et al. 2023)
ESPRESSO	Yes	Clustering	Splice junctions of terminal exons	No	Read level and transcript level with FSM/ISM/NIC categories	(Gao et al. 2023)
IsoQuant	Optional	Graph-based	Clustering	Terminal end A's	No	(Prjibelski et al. 2023)
IsoTools	Yes	Graph-based	Peak calling, read density	No	No	(Lienhard et al. 2023)
SQANTI3	Recommended	Probabilistic ML model	Categorization and modeling	Terminal end A's, terminal end adapters, SRS RNA-seq, complementary SRS data sets	Transcript level, FSM/ISM/NIC categories	(Pardo-Palacios et al. 2024a)

99 nt upstream of previously annotated TSSs and ended 500 nt downstream from annotated PAs, on average (Naftaly et al. 2021). More broadly, high-confidence 5' and 3' terminal sites identified from LRS have increased downstream or upstream LRS read coverage, respectively (Naftaly et al. 2021; Kainth et al. 2023; Pardo-Palacios et al. 2024a). This is unlikely if most of these sites were derived from random truncation events, providing confidence in the identification and quantification of terminal ends from LRS. Therefore, the challenge for computational analyses of LRS data is to filter out spurious reads or sites and determine which novel terminal ends should be considered high confidence.

Comparing to reference transcriptomes

One underlying assumption of these analyses is that most genes and isoforms in a cell should already have been discovered and, thus, contained in a reference transcriptome (Box 2; Tang et al. 2020; Wyman et al. 2020; Tian et al. 2021; Chen et al. 2023; Lienhard et al. 2023; Pardo-Palacios et al. 2024a). Therefore, the default mode for most methods is to intrinsically rely on reference transcriptomes and annotation databases to map and/or define transcripts from annotated genes. Thus, the most conservative analyses of LRS data directly compare groups of reads based on their alignment to previously annotated isoforms and choose terminal sites within read classes based on splicing patterns of known isoforms (Abdel-Ghany et al. 2016; Tang et al. 2020). High-confidence isoforms are those that share all splice sites with a known isoform and have terminal ends within 100 nt of the annotated isoform ends and/or the 5' read ends within the first exon of the annotated gene (Abdel-Ghany et al. 2016; Kovaka et al. 2019). However, because reference databases often do not contain all alternative terminal sites, using databases to infer true terminal ends can lead to errors. For instance, isoforms identified by tools that use annotations to correct ends were less likely to overlap with those identified by orthogonal experimental ap-

proaches that directly identify terminal ends (Pardo-Palacios et al. 2024b).

De novo isoform discovery

Most methods also provide an option to conduct de novo isoform identification to discover novel isoforms. These approaches generally fall into three broad categories: clustering reads, implementing graph theory, and probabilistic modeling. The first category involves generating de novo transcript models by self-aligning or clustering LRS reads and merging them to generate one transcript model per cluster (Tilgner et al. 2014; Gordon et al. 2015; Wang et al. 2016; Komor et al. 2017; Tang et al. 2020; Wyman et al. 2020; Gao et al. 2023). This clustering is often done based on the consistency of internal splice sites, in which each read is compared to previously analyzed reads on the basis of splice junctions (Tang et al. 2020; Wyman et al. 2020; Gao et al. 2023). The second category of de novo isoform discovery approaches leverages graph theory, such that isoforms are determined by charting possible paths through splice sites or introns (using inexact matching algorithms to account for splice site shifts) (Kovaka et al. 2019; Prjibelski et al. 2023). For both the clustering and graph theory approaches, terminal sites are determined either by choosing the most represented site in the transcript cluster (Tang et al. 2020; Wyman et al. 2020) or by specifically clustering read starts and ends and then choosing the most-represented site or most likely site based on annotations (Kovaka et al. 2019; Prjibelski et al. 2023).

The third category of de novo isoform discovery methods involves probabilistic modeling to identify and classify high-confidence and/or full-length isoforms (Chen et al. 2023; Pardo-Palacios et al. 2024a). These approaches train machine learning models on read QC metrics and define a set of parameters by which to correct and filter reads. These parameters can then be used to

Box 2. Defining the ground truth for full-length transcript quantification and discovery

When developing and evaluating new LRS sequencing approaches or analytical frameworks, it is critical to objectively assess the accuracy of the insights gained from the data. For terminal end identification and quantification, it is particularly crucial to assess whether the reads are full length. Several different biological features or synthetic constructs have been put forth as positive controls to define the ground truth for RNA terminal ends, but each of these options have strengths and limitations.

Reference transcriptomes. Comprehensive reference annotations of known isoforms (including exon, intron, and junction connectivity annotations) are often used as the default ground truth with the assumptions that (1) most isoforms have already been observed and (2) isoforms in the database are validated (Tang et al. 2020; Wyman et al. 2020; Tian et al. 2021). Transcriptome annotations are easy to obtain and are well maintained for model organisms or commonly studied species, with the caveat that the maintenance or quality of the annotations varies across species. However, annotations may contain spurious isoforms that have only been observed once, may be incomplete (especially for lowly expressed or condition-specific isoforms), and/or may have inaccurate terminal ends. This can hamper novel isoform discovery, especially for machine learning models that may be trained to recognize features of ubiquitously expressed isoforms. Similarly, the precise estimation of terminal ends may be constrained by known isoform structures for methods that rely on annotated internal exon connectivity to select terminal ends.

Single-exon genes. Genes with no known splice sites or alternative isoforms, estimated to be 2285 genes in the human genome (Jorquera et al. 2021), may provide an internal control for LRS analyses. Reads from these genes allow the assessment of how biases created during RNA extraction, library preparation, and sequencing influence terminal-end identification in situations in which no alternative events are present (Workman et al. 2019). However, their status as single-exon genes is again dependent on known annotations, which might be incorrect, and their use as a control set is conditional on sufficient expression levels within the data set. Finally, it is hard to train models on the limited number of single-exon genes (which often express shorter than average transcripts) (Movassat et al. 2019), and they are not useful for evaluating novel isoform discovery.

Orthogonal experimental data sets. Direct experimental probing of 5' and 3' terminal ends can be used to complement LRS-based identification and quantification of terminal ends. One of the most readily available experimental approaches, SRS RNA-seq, can be used to increase confidence in RNA quantification by providing higher sequencing depth compared with that of LRS approaches. SRS data sets can be generated or downloaded easily, can provide higher statistical power for confidence in novel sites, and can feature well-established methods for correcting biases in these data sets. Thus, these data have been used to empirically determine a ground truth for splice junctions and terminal ends (Depledge et al. 2019; Kainth et al. 2023), which is of particular use for lowly expressed transcripts or splice junctions. However, SRS data sets often suffer from edge effects that limit the ability to directly identify and quantify terminal ends. Although computational approaches exist to overcome these biases (Carralot et al. 2012; Birol et al. 2015; Fiszbein et al. 2022), they do not provide a direct assessment of terminal ends.

Instead, alternative sequencing protocols can be used to directly enrich and characterize terminal ends. For instance, Cap Analysis of Gene Expression (CAGE) (Takahashi et al. 2012b), POINT-seq (Sousa-Luis et al. 2021), RAMPAGE (Batut and Gingeras 2013), and 5' RACE (Scotto-Lavino et al. 2006) are common approaches to do targeted sequencing of 5' ends, whereas QuantSeq, 3'READS, and 3pSeq are common approaches for 3' end sequencing. In addition, because TSSs are known to fall within open chromatin regions, assay for transposase-accessible chromatin (ATAC-seq) data can be used as an orthogonal confirmation for 5' ends (Pardo-Palacios et al. 2024a). Data from these methods set a ground truth to evaluate the accuracy of terminal ends derived from LRS reads or consensus isoforms (Pardo-Palacios et al. 2024b). Additionally, these orthogonal data can be used to filter out reads whose terminal ends are not consistent with empirically derived sites (Anvar et al. 2018; Alfonso-Gonzalez et al. 2023; Calvo-Roitberg et al. 2024). However, care is required when using these data sets, because the results are often tissue- or context-specific and thus need to be collected for every cell type or condition being analyzed. Furthermore, there are several reports of spurious or biologically unclear CAGE peaks within introns and last exons from 5' cap pulldown methods, potentially owing to post-transcriptional cleavage events (Mercer et al. 2010, 2011; Malka et al. 2017; Haberman et al. 2023). These may not reflect true 5' ends of steady-state mRNAs and thus can lead to the inclusion of truncated reads as false positives.

Synthetic RNA pools. To evaluate technical biases introduced during library preparation and sequencing, it has become common to "spike-in" pools of synthetic RNA molecules with known sequences and defined structures. The two most commonly used sets are the External RNA Controls Consortium (ERCC) mix (Pine et al. 2016; Movassat et al. 2019) and the Spike-in RNA Variants (SIRV) (Paul et al. 2016) libraries, which include hundreds of molecules with variable, but known sequences, lengths, and expression levels. SIRV pools include molecules with isoform diversity, and one SIRV pool also contains extremely long (up to 13 kb) monoexonic molecules. Thus, any diversity in read length, sequence (i.e., SNPs), splice sites, or terminal sites in LRS reads can confidently be ascribed to technical biases or sequencing errors. However, these synthetic molecules do not reflect the true biological diversity of sequence or isoform usage and thus limit the ability to assess the accuracy of novel isoform discovery.

Simulated data sets. *In silico* simulated reads data sets offer customizable ground-truth data sets that can incorporate known transcript structures, variable molecule lengths, high-complexity libraries, and a range of known expression levels. Furthermore, simulations can explicitly model truncated reads or other errors during sequencing. Because every step of simulated read generation is computationally programmed and the true value of each parameter can be stored, simulations represent the ultimate ground truth. Tools like Trans-NanoSim (Hafezqorani et al. 2020), IsoSeqSim (<https://github.com/yunhaowang/IsoSeqSim> [accessed July 24, 2024]), and PBSIM3 (Ono et al. 2022) can be used to simulate long reads from a reference transcriptome. SQANTI-SIM (Mestre-Tomás et al. 2023) can also simulate novel transcripts and orthogonal data (e.g., corresponding CAGE peaks). Finally, DeepSimulator (Li et al. 2020) can generate both reads and nanopore raw electrical signals. However, simulations can only be designed to incorporate known biases with clear expected outcomes or distributions and thus are limited by previous knowledge of technical and biological biases.

assign confidence to transcripts and/or identify artefactual transcripts (Chen et al. 2023; Pardo-Palacios et al. 2024a). These models are trained on annotations, simulations, or curated LRS data sets (see Box 2), but in the absence of such data, they can be initiated with user-defined criteria. To derive a set of parameters, the models use sequence (e.g., genome sequence indicative of mis-priming events, poly(A) density at the 3' end, adapter sequences),

coverage (e.g., isoform expression, variability in 5' and 3' site support), splice site information (e.g., known junctions, splice sites), and/or orthogonal data sets (e.g., SRS RNA-seq, experimental methods to identify 5' and 3' sites) to identify transcripts, determine statistical error rates, and quantify isoform expression levels after read or transcript filtering (Chen et al. 2023; Pardo-Palacios et al. 2024a). An advantage of these models is the ability to

calculate the probability with which each read is likely to match a complete transcript and whether it can be uniquely assigned to a single transcript, which then allows for probabilistic estimation of transcript abundance using standard expectation maximization algorithms (Chen et al. 2023). The learning of informative parameters from 5' and 3' features enables these models to classify reads and transcripts by their full-length status.

Leveraging orthogonal data sources

The integration of LRS and SRS data sets can increase confidence by leveraging exon connectivity and increased coverage provided by each method, respectively (Kainth et al. 2023). This is helpful not only for gene or transcript abundance estimates but also for creating putative transcript databases that support the discovery of novel splice sites or terminal ends (Depledge et al. 2019; Fiszbein et al. 2022; Kainth et al. 2023; Calvo-Roitberg et al. 2024; Pardo-Palacios et al. 2024a). For instance, SQANTI3 recently introduced a “TSS ratio” metric, which calculates the ratio of SRS read coverage downstream from TSSs derived from LRS reads relative to the upstream coverage, where true TSSs are expected to be depleted of upstream reads. This feature can be used to filter putative spurious TSSs and can be a feature in the downstream machine learning algorithm. Furthermore, SRS data from orthogonal experimental methods (e.g., CAGE [Takahashi et al. 2012a], ATAC-seq [Buenrostro et al. 2013], 5' RAMPAGE [Takahashi et al. 2012a; Batut and Gingeras 2013], etc., for 5' ends and Quant-seq [Moll et al. 2014], 3p-seq [Jan et al. 2011], etc., for 3' ends) (see Box 2) can be also be used to fine-tune or filter terminal ends derived from LRS data. Although most approaches validate their isoform identification pipelines by comparing with data from these methods (Wyman et al. 2020; Tian et al. 2021; Lienhard et al. 2023; Pardo-Palacios et al. 2024a), some LRS approaches now allow for integrated analyses of these data sets to provide confidence for novel terminal end discovery. SQANTI3 can process complementary data sets from CAGE or Quant-seq experiments or annotations from the PolyASite database (Buenrostro et al. 2013; Herrmann et al. 2019) and reports the distance between terminal regions supported by empirical evidence and LRS terminal sites (Pardo-Palacios et al. 2024a).

More directly, several computational approaches take advantage of the sequence features intrinsic to terminal ends or terminal end adapters added during library preparation as outlined above. At the 3' end, reads with nontemplated poly(A) tails can be used to cluster read ends and identify high-density sites as bona fide 3' ends (Abdel-Ghany et al. 2016; Celik and Mortazavi 2022; Chen et al. 2023; Prjibelski et al. 2023; Pardo-Palacios et al. 2024a). In contrast, the enrichment of templated adenines or thymines in the first or last 20 nt of a read or a cluster of reads may indicate internal mispriming events and can be used to discount those sites as true 3' ends (Wyman et al. 2020; Chen et al. 2023). Finally, several approaches identify 5' or 3' primers added during library preparation (either by ligation or during RT steps) to provide confidence in the identification of terminal sites from reads or consensus transcripts (Abdel-Ghany et al. 2016; Tian et al. 2021; Chen et al. 2023; Pardo-Palacios et al. 2024a). This allows approaches to selectively filter those reads (Tian et al. 2021) or perform a dual analysis with or without those reads (Chen et al. 2023).

Categorizing full-length vs. incomplete isoforms

An alternative way to consider LRS analyses is to retain all data and instead categorize reads or potential consensus isoforms based on

their likelihood of being full length. This is primarily done by assigning each isoform to a “completeness” class based on some or all of the features listed above, often using machine learning models that curate 3' and 5' descriptors to fine-tune classifiers (Wyman et al. 2020; Pardo-Palacios et al. 2024a). Transcripts with splice junctions that perfectly match known isoform models (generally based on annotations) are termed “full splice matches” (FSMs), with some flexibility at the 5' and 3' ends (FSM alt 5' end or FSM alt 3' end). Transcripts that overlap some junctions of an existing isoform are termed “incomplete splice matches” (ISMs), with flexibility at the 5' and/or 3' end or intron inclusion (ISM 5' fragment, ISM 3' fragment, ISM internal fragment, or ISM intron retention). Transcripts with no junctions are termed “monoexon,” and those with novel junctions or splice sites are termed “novel-in-catalog” (NIC) or “novel not-in-catalog” (NNC), respectively (where “in-catalog” refers to annotated sites). These categories provide information about potentially incomplete isoforms without rigid filtering and flexibility to decide which isoforms to include in downstream analyses. Approaches that use probabilistic modeling can also calculate the likelihood of individual reads matching complete isoforms (based on annotations) and, thus, provide two abundance estimates per isoform: one from reads that are likely to be full length and another that is more inclusive of reads that may be incomplete (Chen et al. 2023). This approach found that abundance estimates were more reproducible when only using reads that are likely to be full length (Chen et al. 2023). This framework also allows for analyses that require individual read information, such as those that investigate coordinated site usage across individual molecules.

What's next? Remaining challenges for characterizing mRNA terminal ends with LRS technologies

Although LRS continues to offer new possibilities for deeper characterization of essential biological and genomic processes, it is important to be cautious when selecting the specific application and interpreting data, as technical and analytical biases may affect the results and/or lead to misleading findings. This risk is especially prevalent for the discovery or quantification of the ends of mRNA transcripts. As outlined above, LRS reads are likely to delineate spurious RNA terminal ends owing to numerous biases or issues introduced during the library preparation or sequencing stages. Despite advances in experimental and computational approaches to overcome these challenges, several key developments are needed for improved LRS-based discovery and characterization of mRNA terminal ends.

Many experimental steps during the generation of LRS libraries can be optimized, starting from the input material. Preservation of full-length transcripts is critical; therefore, care is needed during RNA extraction and library preparation to avoid shearing effects by pipetting too quickly or mixing reagents too aggressively. As mentioned above, during library preparation, several steps are potentially problematic; thus, further optimization of RT and the subsequent PCR amplification is needed to ensure the complete processing of the entire molecule and avoid the introduction of length biases, respectively. Increasing adapter lengths or more widespread use of UMIs at both ends of the molecule might help to resolve technical artifacts (Wyman et al. 2020; Ibrahim et al. 2021). Finally, although several approaches have been described to select and enrich full-length RNA molecules, further development is anticipated to enhance the robustness and throughput of these applications. One area of particular interest is the

development of methods to directly identify 5' end mRNA modifications (e.g., 5' 7-methyl-guanosine caps) or sequences during direct RNA sequencing, which would likely require advances in both biological insight and technical capabilities (for reviews, see Kong et al. 2023; van Dijk et al. 2023).

Following library preparation, LRS sequencing instruments themselves require additional design considerations, particularly for ONT flow cells and/or instruments. More accurate detection methods for signal spikes or aberrant pauses can reduce error rates and improve basecalling accuracy. ONT can also improve the pore characteristics and control translocation specific for RNA applications, including the development of solid-state pores with more consistent characteristics and performance (Fragasso et al. 2020; Fried et al. 2021), appropriate translocation speed, and software ability to hold a strand within the pore and read the strand several times, which may allow both time for more accurate basecalling and repeated sensing of the terminal ends to improve coverage and increase confidence in basecalls. Additional robust methods are needed to address the challenges of accurate basecalling of the terminal 10–15 nt passing through the ONT pore, as well as continued basecalling following a spike in electrical current. Both PacBio and ONT continue to develop updated methods to boost throughput to increase coverage, which will allow higher basecalling accuracy (Jain et al. 2022; Al'Khafaji et al. 2024).

Finally, there is still room for development and optimization of LRS analyses approaches. Despite progress in the development of methods to correct or model biases, accurately identifying terminal ends remains an unsolved problem. Although filtering or classifying reads can improve confidence in terminal-end observations, these steps come at the cost of read coverage. Moreover, the reliance on external (often not sample-matched) data sets reduces the ability to confidently make novel observations. The current tools for isoform discovery generally involve either clustering, graph theory, and probabilistic modeling, and they estimate confidence using either reference annotations or orthogonal data sets. An inherent challenge across all of these analytical tools is the definition of a comprehensive “ground-truth” data set for training data or validating findings. The establishment of these data would enable further development of data-informed thresholds for calling true isoforms to better understand the trade-offs between sensitivity and accuracy across these approaches (particularly for the detection of rare isoforms) (Chen et al. 2023). When specifically thinking about terminal ends, it appears that the widespread anchoring of isoform models on splicing junctions prevents current methods from identifying isoforms that only differ in their start or end coordinates. Explicit modeling of technical processes that cause read truncation and incorporation of isoform-specific transcription and/or degradation rates may allow methods to overcome these limitations (Chen et al. 2023). Finally, there is a need for additional characterization of the statistical power for isoform discovery and quantification afforded across different LRS sequencing depths to better understand how many LRS reads are needed for different biological analyses.

Recent benchmarking studies have made significant strides toward standardizing depth and read length expectations in the field (Chen et al. 2021; Dong et al. 2023; Pardo-Palacios et al. 2024b), as well as evaluating both library preparation strategies and computational pipelines for transcript/gene discovery and quantification. The results of these studies have generally shown that PacBio delivers higher accuracy and longer reads compared with ONT, yet the cost per read is significantly lower for ONT (Mikheenko et al. 2022; Pardo-Palacios et al. 2024b).

Additionally, the accuracy of both technologies is still lower than that for SRS (Pardo-Palacios et al. 2024b). Thus, the choice of which technology likely varies depended on specific research objectives and practical considerations. For projects aiming to quantify known RNA isoforms with high-throughput, the affordability of ONT technologies might be preferable. Conversely, the accuracy and longer reads provided by PacBio might be better suited for research focused on discovering novel isoforms or quantifying alternative terminal site usage. Importantly, these benchmarking studies continue to standardize assessment methods to ensure comparability in results across studies, identify systematic technical and computational challenges that must be addressed, and provide practical guidance for researchers in this field. However, rapid developments in LRS technology, both biologically and computationally, raise the need for caution when considering the conclusions of such benchmarking papers, as changes in chemistry and software versions can quickly make comparisons obsolete. Thus, continuous efforts are needed to ensure rapid assessments of these updates.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Ana Fiszbein, Christine Carroll, Zach Wakefield, Steven T. Mick, and the Pai laboratory for help with databases, discussions, and feedback on the manuscript. Funding was by the National Institute of General Medical Sciences (R35GM133762) to E.C.-R., R.F.D., and A.A.P.; the National Human Genome Research Institute (R01HG012967) to E.C.-R. and A.A.P.; and the National Institute of Allergy and Infectious Diseases (R21AI166281) to R.F.D.

Author contributions: Conceptualization was by E.C.-R., R.F.D., and A.A.P. Formal analysis was by E.C.-R. Investigation was by E.C.-R. Supervision was by A.A.P. Visualization was by E.C.-R. and R.F.D. Writing of the original draft preparation was by E.C.-R., R.F.D., and A.A.P.

References

- Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, Ben-Hur A, Reddy ASN. 2016. A survey of the sorghum transcriptome using single-molecule long reads. *Nat Commun* **7**: 11706. doi:10.1038/ncomms11706
- Abugessaisa I, Ramilowski JA, Lizio M, Severin J, Hasegawa A, Harshbarger J, Kondo A, Noguchi S, Yip CW, Ooi JLC, et al. 2021. FANTOM enters 20th year: expansion of transcriptomic atlases and functional annotation of non-coding RNAs. *Nucleic Acids Res* **49**: D892–D898. doi:10.1093/nar/gkaa1054
- Alfonso-Gonzalez C, Hilgers V. 2024. (Alternative) transcription start sites as regulators of RNA processing. *Trends Cell Biol* doi:10.1016/j.tcb.2024.02.010
- Alfonso-Gonzalez C, Legnini I, Holec S, Arrigoni L, Ozbulut HC, Mateos F, Koppstein D, Rybak-Wolf A, Bönisch U, Rajewsky N, et al. 2023. Sites of transcription initiation drive mRNA isoform selection. *Cell* **186**: 2438–2455.e22. doi:10.1016/j.cell.2023.04.012
- Al'Khafaji AM, Smith JT, Garimella KV, Babadi M, Popic V, Sade-Feldman M, Gatzem M, Sarkizova S, Schwartz MA, Blaum EM, et al. 2024. High-throughput RNA isoform sequencing using programmed cDNA concatenation. *Nat Biotechnol* **42**: 582–586. doi:10.1038/s41587-023-01815-7
- Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21**: 30. doi:10.1186/s13059-020-1935-5
- Anvar SY, Allard G, Tseng E, Sheykmann GM, de Klerk E, Vermaat M, Yin RH, Johansson HE, Ariyurek Y, den Dunnen JT, et al. 2018. Full-length mRNA sequencing uncovers a widespread coupling between

Challenges in identifying RNA terminal ends

- transcription initiation and mRNA processing. *Genome Biol* **19**: 46. doi:10.1186/s13059-018-1418-0
- Arefeen A, Xiao X, Jiang T. 2019. DeepPASTA: deep neural network based polyadenylation site analysis. *Bioinformatics* **35**: 4577–4585. doi:10.1093/bioinformatics/btz283
- Au KF, Sebastian V, Afshar PT, Durruthy JD, Lee L, Williams BA, van Bakel H, Schadt EE, Reijo-Pera RA, Underwood JG, et al. 2013. Characterization of the human ESC transcriptome by hybrid sequencing. *Proc Natl Acad Sci* **110**: E4821–E4830.
- Balázs Z, Tombácz D, Csabai Z, Moldován N, Snyder M, Boldogkői Z. 2019. Template-switching artifacts resemble alternative polyadenylation. *BMC Genomics* **20**: 824. doi:10.1186/s12864-019-6199-7
- Baslan T, Kovaka S, Sedlacek HJ, Zhang Y, Wappel R, Tian S, Lowe SW, Goodwin S, Schatz MC. 2021. High resolution copy number inference in cancer using short-molecule nanopore sequencing. *Nucleic Acids Res* **49**: e124. doi:10.1093/nar/gkab812
- Batut P, Gingras TR. 2013. RAMPAGE: promoter activity profiling by paired-end sequencing of 5'-complete cDNAs. *Curr Protoc Mol Biol* **104**: Unit 25B.11. doi:10.1002/0471142727.mb25b11s104
- Beyer D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, Atlasson BA, Kristmundsdottir S, Mehringer S, Hardarson MT, et al. 2021. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat Genet* **53**: 779–786. doi:10.1038/s41588-021-00865-4
- Birol I, Raymond A, Chiu R, Nip KM, Jackman SD, Kreitzman M, Docking TR, Ennis CA, Robertson AG, Karsan A. 2015. Kleat: cleavage site analysis of transcriptomes. *Pac Symp Biocomput* 347–358. doi:10.1142/9789814644730_0034
- Boldogkői Z, Moldován N, Balázs Z, Snyder M, Tombácz D. 2019. Long-read sequencing: a powerful tool in viral transcriptome research. *Trends Microbiol* **27**: 578–592. doi:10.1016/j.tim.2019.01.010
- Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525–527. doi:10.1038/nbt.3519
- Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**: 1213–1218. doi:10.1038/nmeth.2688
- Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, DuBois RM, Forsberg EC, Akeson M, Vollmers C. 2017. Nanopore long-read RNASeq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat Commun* **8**: 16027. doi:10.1038/ncomms16027
- Byrne A, Cole C, Volden R, Vollmers C. 2019. Realizing the potential of full-length transcriptome sequencing. *Philos Trans R Soc Lond B Biol Sci* **374**: 20190097. doi:10.1098/rstb.2019.0097
- Calvo-Roitberg E, Carroll CL, Venec SV, Kim G, Mick ST, Dekker J, Fiszbein A, Pai AA. 2024. mRNA initiation and termination are spatially coordinated. bioRxiv doi:10.1101/2024.01.05.574404
- Carbonell-Sala S, Perfettella T, Lagarde J, Nishiyori H, Palumbo E, Arnan C, Takahashi H, Carninci P, Uszczynska-Ratajczak B, Guigó R. 2024. CapTrap-seq: a platform-agnostic and quantitative approach for high-fidelity full-length RNA sequencing. *Nat Commun* **15**: 5278. doi:10.1038/s41467-024-49523-3
- Carralot JP, Ogier A, Boese A, Genovesio A, Brodin P, Sommer P, Dorval T. 2012. A novel specific edge effect correction method for RNA interference screenings. *Bioinformatics* **28**: 261–268. doi:10.1093/bioinformatics/btr648
- Carrión SA, Michal JJ, Jiang Z. 2023. Alternative transcripts diversify genome function for genome relevance to health and diseases. *Genes (Basel)* **14**: 2051. doi:10.3390/genes14112051
- Carroccia TJ, Neugebauer KM. 2019. Pre-mRNA splicing in the nuclear landscape. *Cold Spring Harb Symp Quant Biol* **84**: 11–20. doi:10.1101/sqb.2019.84.040402
- Cartolano M, Huettel B, Hartwig B, Reinhardt R, Schnieberger K. 2016. cDNA library enrichment of full length transcripts for SMRT long read sequencing. *PLoS One* **11**: e0157779. doi:10.1371/journal.pone.0157779
- Cass AA, Xiao X. 2019. mountainClimber identifies alternative transcription start and polyadenylation sites in RNA-seq. *Cell Syst* **9**: 393–400.e6. doi:10.1016/j.cels.2019.07.011
- Celik MH, Mortazavi A. 2022. Analysis of alternative polyadenylation from long-read or short-read RNA-seq with LAPA. bioRxiv doi:10.1101/2022.11.08.515683
- Chen Y, Davidson NM, Wan YK, Patel H, Yao F, Low HM, Hendra C, Watten L, Sim A, Sawyer C, et al. 2021. A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines. bioRxiv doi:10.1101/2021.04.21.440736 (The SG-NEx data were accessed in May 2023 at registry.opendata.aws/sg-nex-data)
- Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, Love MI, Göke J. 2023. Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nat Methods* **20**: 1187–1195. doi:10.1038/s41592-023-01908-w
- Choquet K, Baxter-Koenigs AR, Dulk S-L, Smalec BM, Rouskin S, Churchman LS. 2023. Pre-mRNA splicing order is predetermined and maintains splicing fidelity across multi-intronic transcripts. *Nat Struct Mol Biol* **30**: 1064–1076. doi:10.1038/s41594-023-01035-2
- Cocquet J, Chong A, Zhang G, Veitia RA. 2006. Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88**: 127–131. doi:10.1016/j.ygeno.2005.12.013
- Conesa A, Madrigal P, Tarazona S, Gomez-Cabrer D, Cervera A, McPherson A, Szczęśniak MW, Gaffney DJ, Elo LL, Zhang X, et al. 2016. A survey of best practices for RNA-seq data analysis. *Genome Biol* **17**: 13. doi:10.1186/s13059-016-0881-8
- Coulon A, Ferguson ML, de Turris V, Palangat M, Chow CC, Larson DR. 2014. Kinetic competition during the transcription cycle results in stochastic RNA processing. *eLife* **3**: e03939. doi:10.7554/eLife.03939
- Cui J, Shen N, Lu Z, Xu G, Wang Y, Jin B. 2020. Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the *Arabidopsis* transcriptome. *Plant Methods* **16**: 85. doi:10.1186/s13007-020-00629-x
- Depledge DP, Srinivas KP, Sadaoka T, Bready D, Mori Y, Placantonakis DG, Mohr I, Wilson AC. 2019. Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nat Commun* **10**: 754. doi:10.1038/s41467-019-08734-9
- Di Giannattasio DC, Nishida K, Manley JL. 2011. Mechanisms and consequences of alternative polyadenylation. *Mol Cell* **43**: 853–866. doi:10.1016/j.molcel.2011.08.017
- Dohm JC, Peters P, Stralis-Pavese N, Himmelbauer H. 2020. Benchmarking of long-read correction methods. *NAR Genom Bioinform* **2**: lqaa037. doi:10.1093/nargab/lqaa037
- Dong X, Du MRM, Gouil Q, Tian L, Jabbari JS, Bowden R, Baldoni PL, Chen Y, Smyth GK, Amarasinghe SL, et al. 2023. Benchmarking long-read RNA-sequencing analysis tools using *in silico* mixtures. *Nat Methods* **20**: 1810–1821. doi:10.1038/s41592-023-02026-3
- Drexler HL, Choquet K, Merens HE, Tang PS, Simpson JT, Churchman LS. 2021. Revealing nascent RNA processing dynamics with nano-COP. *Nat Protoc* **16**: 1343–1375. doi:10.1038/s41596-020-00469-y
- Dudnyk K, Cai D, Shi C, Xu J, Zhou J. 2024. Sequence basis of transcription initiation in the human genome. *Science* **384**: eadj0116. doi:10.1126/science.adaj0116
- The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74. doi:10.1038/nature11247
- Fiszbein A, Krick KS, Begg BE, Burge CB. 2019. Exon-mediated activation of transcription starts. *Cell* **179**: 1551–1565.e17. doi:10.1016/j.cell.2019.11.002
- Fiszbein A, McGurk M, Calvo-Roitberg E, Kim G, Burge CB, Pai AA. 2022. Widespread occurrence of hybrid internal-terminal exons in human transcriptomes. *Sci Adv* **8**: eabk1752. doi:10.1126/sciadv.abk1752
- Fragasso A, Schmid S, Dekker C. 2020. Comparing current noise in biological and solid-state nanopores. *ACS Nano* **14**: 1338–1349. doi:10.1021/acsnano.9b09353
- Fried JP, Swett JL, Nadappuram BP, Fedosyuk A, Sousa PM, Briggs DP, Ivanov AP, Edel JB, Mol JA, Yates JR. 2021. Understanding electrical conduction and nanopore formation during controlled breakdown. *Small* **17**: e2102543. doi:10.1002/smll.202102543
- Galloway A, Cowling VH. 2019. mRNA cap regulation in mammalian cell function and fate. *Biochim Biophys Acta Gene Regul Mech* **1862**: 270–279. doi:10.1016/j.bbagr.2018.09.011
- Gao Y, Wang F, Wang R, Kutschera E, Xu Y, Xie S, Wang Y, Kadash-Edmondson KE, Lin L, Xing Y. 2023. ESPRESSO: robust discovery and quantification of transcript isoforms from error-prone long-read RNA-seq data. *Sci Adv* **9**: eabq5072. doi:10.1126/sciadv.abq5072
- Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, et al. 2018. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* **15**: 201–206. doi:10.1038/nmeth.4577
- Geisberg JV, Moqtaderi Z, Fong N, Erickson B, Bentley DL, Struhl K. 2022. Nucleotide-level linkage of transcriptional elongation and polyadenylation. *eLife* **11**: e83153. doi:10.7554/eLife.83153
- Gordon SP, Tseng E, Salamov A, Zhang J, Meng X, Zhao Z, Kang D, Underwood J, Grigoriev IV, Figueroa M, et al. 2015. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLoS One* **10**: e0132628. doi:10.1371/journal.pone.0132628
- Grünberger F, Ferreira-Cerca S, Grohmann D. 2022. Nanopore sequencing of RNA and cDNA molecules in *Escherichia coli*. *RNA* **28**: 400–417. doi:10.1261/rna.078937.121
- Gupta I, Collier PG, Haase B, Mahfouz A, Joglekar A, Floyd T, Koopmans F, Barres B, Smit AB, Sloan SA, et al. 2018. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol* **36**: 1197–1202. doi:10.1038/nbt.4259

- Haberman N, Digby H, Faraway R, Cheung R, Jobbins AM, Parr C, Yasuzawa K, Kasukawa T, Yip CW, Kato M, et al. 2023. Abundant capped RNAs are derived from mRNA cleavage at 3'UTR G-quadruplexes. *bioRxiv* doi:10.1101/2023.04.27.538568
- Hafezqorani S, Yang C, Lo T, Nip KM, Warren RL, Birol I. 2020. Trans-NanoSim characterizes and simulates nanopore RNA-sequencing data. *GigaScience* **9**: gaa061. doi:10.1093/gigascience/gaa061
- Herrmann CJ, Schmidt R, Kanitz A, Artimo P, Gruber AJ, Zavolan M. 2019. PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3' end sequencing. *Nucleic Acids Res* **48**: D174–D179. doi:10.1093/nar/gkz918
- Herzel L, Straube K, Neugebauer KM. 2018. Long-read sequencing of nascent RNA reveals coupling among RNA processing events. *Genome Res* **28**: 1008–1019. doi:10.1101/gr.232025.117
- Hölzer M, Marz M. 2019. De novo transcriptome assembly: a comprehensive cross-species comparison of short-read RNA-seq assemblers. *GigaScience* **8**: giz039. doi:10.1093/gigascience/giz039
- Ibrahim F, Oppelt J, Maragkakis M, Mourelatos Z. 2021. TERA-Seq: true end-to-end sequencing of native RNA molecules for transcriptome characterization. *Nucleic Acids Res* **49**: e1115. doi:10.1093/nar/gkab713
- Iizuka R, Yamazaki H, Uemura S. 2022. Zero-mode waveguides and nanopore-based sequencing technologies accelerate single-molecule studies. *Biophys Physicobiol* **19**: e190032. doi:10.2142/biophysico.bppb-v19.0032
- Jain M, Abu-Shumays R, Olsen HE, Akeson M. 2022. Advances in nanopore direct RNA sequencing. *Nat Methods* **19**: 1160–1164. doi:10.1038/s41592-022-01633-w
- Jan CH, Friedman RC, Ruby JG, Bartel DP. 2011. Formation, regulation and evolution of *Caenorhabditis elegans* 3'UTRs. *Nature* **469**: 97–101. doi:10.1038/nature09616
- Jia J, Long Y, Zhang H, Li Z, Liu Z, Zhao Y, Lu D, Jin X, Deng X, Xia R, et al. 2020. Post-transcriptional splicing of nascent RNA contributes to widespread intron retention in plants. *Nat Plants* **6**: 780–788. doi:10.1038/s41477-020-0688-1
- Jiang F, Zhang J, Liu Q, Liu X, Wang H, He J, Kang L. 2019. Long-read direct RNA sequencing by 5'-cap capturing reveals the impact of Piwi on the widespread exonization of transposable elements in locusts. *RNA Biol* **16**: 950–959. doi:10.1080/15476286.2019.1602437
- Jorquera R, González C, Clausen PTLC, Petersen B, Holmes DS. 2021. Sinex DB 2.0 update 2020: database for eukaryotic single-exon coding sequences. *Database* **2021**: baab002. doi:10.1093/database/baab002
- Juven-Gershon T, Hsu J-Y, Theisen JW, Kadonaga JT. 2008. The RNA polymerase II core promoter: the gateway to transcription. *Curr Opin Cell Biol* **20**: 253–259. doi:10.1016/j.ceb.2008.03.003
- Kainth AS, Haddad GA, Hall JM, Ruthenburg AJ. 2023. Merging short and stranded long reads improves transcript assembly. *PLoS Comput Biol* **19**: e1011576. doi:10.1371/journal.pcbi.1011576
- Kellner S, Burhenne J, Helm M. 2010. Detection of RNA modifications. *RNA Biol* **7**: 237–247. doi:10.4161/rna.7.2.11468
- Komor MA, Pham TV, Hiemstra AC, Pierwsa SR, Bolijn AS, Schelfhorst T, Delis-van Diemen PM, Tijsse M, Sebra RP, Ashby M, et al. 2017. Identification of differentially expressed splice variants by the proteogenomic pipeline Splicify. *Mol Cell Proteomics* **16**: 1850–1863. doi:10.1074/mcp.TIR117.000056
- Kong Y, Mead EA, Fang G. 2023. Navigating the pitfalls of mapping DNA and RNA modifications. *Nat Rev Genet* **24**: 363–381. doi:10.1038/s41576-022-00559-5
- Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20**: 278. doi:10.1186/s13059-019-1910-1
- Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Pérez-Lluch S, Abad A, Davis C, Gingeras TR, Frankish A, Harrow J, Guigo R, et al. 2017. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat Genet* **49**: 1731–1740. doi:10.1038/ng.3988
- Legnini I, Alles J, Karaiskos N, Ayoub S, Rajewsky N. 2019. FLAM-seq: full-length mRNA sequencing reveals principles of poly(A) tail length control. *Nat Methods* **16**: 879–886. doi:10.1038/s41592-019-0503-y
- Leppek K, Das R, Barna M. 2018. Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat Rev Mol Cell Biol* **19**: 158–174. doi:10.1038/nrm.2017.103
- Li Y, Wang S, Bi C, Qiu Z, Li M, Gao X. 2020. DeepSimulator1.5: a more powerful, quicker and lighter simulator for Nanopore sequencing. *Bioinformatics* **36**: 2578–2580. doi:10.1093/bioinformatics/btz963
- Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. 2013. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev* **27**: 2380–2396. doi:10.1101/gad.229328.113
- Lienhard M, van den Beucken T, Timmermann B, Hochradel M, Börno S, Caïment F, Vingron M, Herwig R. 2023. IsoTools: a flexible workflow for long-read transcriptome sequencing analysis. *Bioinformatics* **39**: btad364. doi:10.1093/bioinformatics/btad364
- Long Y, Jia J, Mo W, Jin X, Zhai J. 2021. 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* **16**: 4355–4381. doi:10.1038/s41596-021-00581-7
- Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, Myers Z, Sud P, Jou J, Lin K, et al. 2020. New developments on the encyclopedia of DNA elements (ENCODE) data portal. *Nucleic Acids Res* **48**: D882–D889. doi:10.1093/nar/gkz1062
- Maeng JH, Jang HJ, Du AY, Tzeng S-C, Wang T. 2023. Using long-read CAGE sequencing to profile cryptic-promoter-derived transcripts and their contribution to the immunopeptidome. *Genome Res* **33**: 2143–2155. doi:10.1101/gr.277061.122
- Maestri S, Maturo MG, Cosentino E, Marcolungo L, Iadarola B, Fortunati E, Rossato M, DelleDonne M. 2020. A long-read sequencing approach for direct haplotype phasing in clinical settings. *Int J Mol Sci* **21**: 9177. doi:10.3390/ijms21239177
- Malka Y, Steiman-Shimony A, Rosenthal E, Argaman L, Cohen-Daniel L, Arbib E, Margalit H, Kaplan T, Berger M. 2017. Post-transcriptional 3'-UTR cleavage of mRNA transcripts generates thousands of stable uncapped autonomous RNA fragments. *Nat Commun* **8**: 2029. doi:10.1038/s41467-017-02099-7
- Martín-Alonso S, Frutos-Beltrán E, Menéndez-Arias L. 2021. Reverse transcriptase: from transcriptomics to genome editing. *Trends Biotechnol* **39**: 194–210. doi:10.1016/j.tibtech.2020.06.008
- Mercer TR, Dinger ME, Bracken CP, Kolle G, Szubert JM, Korbie DJ, Askarian-Amiri ME, Gardiner BB, Goodall GJ, Grimmond SM, et al. 2010. Regulated post-transcriptional RNA cleavage diversifies the eukaryotic transcriptome. *Genome Res* **20**: 1639–1650. doi:10.1101/gr.112128.110
- Mercer TR, Wilhelm D, Dinger ME, Soldà G, Korbie DJ, Glazov EA, Truong V, Schwenke M, Simons C, Matthaei KI, et al. 2011. Expression of distinct RNAs from 3' untranslated regions. *Nucleic Acids Res* **39**: 2393–2403. doi:10.1093/nar/gkq1158
- Mestre-Tomás J, Liu T, Pardo-Palacios F, Conesa A. 2023. SQANTI-SIM: a simulator of controlled transcript novelty for lRNA-seq benchmark. *Genome Biol* **24**: 286. doi:10.1186/s13059-023-03127-0
- Mikheenko A, Prijibelski AD, Joglekar A, Tilgner HU. 2022. Sequencing of individual barcoded cDNAs using pacific biosciences and Oxford nanopore technologies reveals platform-specific error patterns. *Genome Res* **32**: 726–737. doi:10.1101/gr.276405.121
- Minshall N, Git A. 2020. Enzyme- and gene-specific biases in reverse transcription of RNA raise concerns for evaluating gene expression. *Sci Rep* **10**: 8151. doi:10.1038/s41598-020-65005-0
- Mohr S, Ghanem E, Smith W, Sheeter D, Qin Y, King O, Polioudakis D, Iyer VR, Hunnicke-Smith S, Swamy S, et al. 2013. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA* **19**: 958–970. doi:10.1261/ma.039743.113
- Moldován N, Tombácz D, Szűcs A, Csabai Z, Balázsz Z, Kis E, Molnár J, Boldogkői Z. 2018. Third-generation sequencing reveals extensive polycistrion and transcriptional overlapping in a baculovirus. *Sci Rep* **8**: 8604. doi:10.1038/s41598-018-26955-8
- Moll P, Ante M, Seitz A, Reda T. 2014. QuantSeq 3' mRNA sequencing for RNA quantification. *Nat Methods* **11**: i–iii. doi:10.1038/nmeth.f.376
- Moore JE, Zhang X-O, Elhajjajy SI, Fan K, Pratt HE, Reese F, Mortazavi A, Weng Z. 2022. Integration of high-resolution promoter profiling assays reveals novel, cell type-specific transcription start sites across 115 human cell and tissue types. *Genome Res* **32**: 389–402. doi:10.1101/gr.275723.121
- Movassat M, Forouzmand E, Reese F, Hertel KJ. 2019. Exon size and sequence conservation improves identification of splice-altering nucleotides. *RNA* **25**: 1793–1805. doi:10.1261/rna.070987.119
- Mulrone L, Wulf MG, Schildkraut I, Tzertzinis G, Buswell J, Jain M, Olsen H, Diekhans M, Corrêa IR, Akeson M, et al. 2022. Identification of high-confidence human poly(A) RNA isoform scaffolds using nanopore sequencing. *RNA* **28**: 162–176. doi:10.1261/rna.078703.121
- Naftaly AS, Pau S, White MA. 2021. Long-read RNA sequencing reveals widespread sex-specific alternative splicing in threespine stickleback fish. *Genome Res* **31**: 1486–1497. doi:10.1101/gr.274282.120
- Nam DK, Lee S, Zhou G, Cao X, Wang C, Clark T, Chen J, Rowley JD, Wang SM. 2002. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proc Natl Acad Sci* **99**: 6152–6156. doi:10.1073/pnas.092140899
- Ogami K, Oishi Y, Hoshino S-I. 2023. Protocol for analyzing intact mRNA poly(A) tail length using nanopore direct RNA sequencing. *STAR Protoc* **4**: 102340. doi:10.1016/j.xpro.2023.102340
- Ono Y, Hamada M, Asai K. 2022. PBSIM3: a simulator for all types of PacBio and ONT long reads. *NAR Genom Bioinform* **4**: lqac092. doi:10.1093/nar/gab/lqac092

- Oscorbin IP, Filipenko ML. 2021. M-MuLV reverse transcriptase: selected properties and improved mutants. *Comput Struct Biotechnol J* **19**: 6315–6327. doi:10.1016/j.csbj.2021.11.030
- Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomás J, Amorín R, Estevan-Morió E, Liu T, Nanni A, McIntyre L, et al. 2024a. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nat Methods* **21**: 793–797. doi:10.1038/s41592-024-02229-2
- Pardo-Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, Loveland JE, De Maria M, Adams MS, Balderrama-Gutiérrez G, et al. 2024b. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat Methods* **21**: 1349–1363. doi:10.1038/s41592-024-02298-3
- Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. 2016. The impact of amplification on differential expression analyses by RNA-seq. *Sci Rep* **6**: 25533. doi:10.1038/srep25533
- Parker MT, Knop K, Sherwood AV, Schurch NJ, Mackinnon K, Gould PD, Hall AJ, Barton GJ, Simpson GG. 2020. Nanopore direct RNA sequencing maps the complexity of *Arabidopsis* mRNA processing and m6A modification. *eLife* **9**: e49658. doi:10.7554/eLife.49658
- Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**: 417–419. doi:10.1038/nmeth.4197
- Paul L, Kubala P, Horner G, Ante M, Holländer I, Alexander S, Reda T. 2016. SIRVs: spike-in RNA variants as external isoform controls in RNA-sequencing. bioRxiv doi:10.1101/080747
- Pine PS, Munro SA, Parsons JR, McDaniel J, Lucas AB, Lozach J, Myers TG, Su Q, Jacobs-Helber SM, Salit M. 2016. Evaluation of the external RNA controls consortium (ERCC) reference material using a modified Latin square design. *BMC Biotechnol* **16**: 54. doi:10.1186/s12896-016-0281-x
- Pribjelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL, Tilgner HU. 2023. Accurate isoform discovery with IsoQuant using long reads. *Nat Biotechnol* **41**: 915–918. doi:10.1038/s41587-022-01565-y
- Pugh J. 2023. The current state of nanopore sequencing. In *Nanopore sequencing: methods and protocols* (ed. Arakawa K), pp. 3–14, Springer, New York.
- Qin Z, Stoilov P, Zhang X, Xing Y. 2018. SEASTAR: systematic evaluation of alternative transcription start sites in RNA. *Nucleic Acids Res* **46**: e45. doi:10.1093/nar/gky053
- Rausch T, Snajder R, Leger A, Simovic M, Giurgiu M, Villacorta L, Henssen AG, Fröhling S, Stegle O, Birney E, et al. 2023. Long-read sequencing of diagnosis and post-therapy medulloblastoma reveals complex rearrangement patterns and epigenetic signatures. *Cell Genom* **3**: 100281. doi:10.1016/j.xgen.2023.100281
- Reimer KA, Neugebauer KM. 2020. Preparation of mammalian nascent RNA for long read sequencing. *Curr Protoc Mol Biol* **133**: e128. doi:10.1002/cpmb.128
- Reimer KA, Mimoso CA, Adelman K, Neugebauer KM. 2021. Co-transcriptional splicing regulates 3' end cleavage during mammalian erythropoiesis. *Mol Cell* **81**: 998–1012.e7. doi:10.1016/j.molcel.2020.12.018
- Reyes A, Huber W. 2018. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res* **46**: 582–592. doi:10.1093/nar/gkx1165
- Rosa-Mercado NA, Steitz JA. 2022. Who let the DoGs out?: biogenesis of stress-induced readthrough transcripts. *Trends Biochem Sci* **47**: 206–217. doi:10.1016/j.tibs.2021.08.003
- Scarano C, Veneruso I, De Simone RR, Di Bonito G, Secondino A, D'Argenio V. 2024. The third-generation sequencing challenge: novel insights for the omic sciences. *Biomolecules* **14**: 568. doi:10.3390/biom14050568
- Scotto-Lavino E, Du G, Frohman MA. 2006. 5' end cDNA amplification using classic RACE. *Nat Protoc* **1**: 2555–2562. doi:10.1038/nprot.2006.480
- Shabalina SA, Ogurtsov AY, Spiridonov NA, Koonin EV. 2014. Evolution at protein ends: major contribution of alternative transcription initiation and termination to the transcriptome and proteome diversity in mammals. *Nucleic Acids Res* **42**: 7132–7144. doi:10.1093/nar/gku342
- Shagin DA, Lukyanov KA, Vagner LL, Matz MV. 1999. Regulation of average length of complex PCR product. *Nucleic Acids Res* **27**: e23. doi:10.1093/nar/27.18.e23
- Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31**: 1009–1014. doi:10.1038/nbt.2705
- Sheynkman GM, Tuttle KS, Laval F, Tseng E, Underwood JG, Yu L, Dong D, Smith ML, Sebra R, Willems L, et al. 2020. ORF Capture-Seq as a versatile method for targeted identification of full-length isoforms. *Nat Commun* **11**: 2326. doi:10.1038/s41467-020-16174-z
- Shi H, Zhou Y, Jia E, Pan M, Bai Y, Ge Q. 2021. Bias in RNA-seq library preparation: current challenges and solutions. *Biomed Res Int* **2021**: 6647597. doi:10.1155/2021/6647597
- Shiau C-K, Lu L, Kieser R, Fukumura K, Pan T, Lin H-Y, Yang J, Tong EL, Lee G, Yan Y, et al. 2023. High throughput single cell long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors. *Nat Commun* **14**: 4124. doi:10.1038/s41467-023-39813-7
- Soares LMM, Hanscom T, Selby DE, Adjei S, Wang W, Przybylski D, Thompson JF. 2022. DNA read count calibration for single-molecule, long-read sequencing. *Sci Rep* **12**: 17257. doi:10.1038/s41598-022-21606-5
- Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain S. 2019. A comprehensive examination of nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat Commun* **10**: 3359. doi:10.1038/s41467-019-11272-z
- Sousa-Luís R, Dujardin G, Zukher I, Kimura H, Weldon C, Carmo-Fonseca M, Proudfoot NJ, Nojima T. 2021. POINT technology illuminates the processing of polymerase-associated intact nascent transcripts. *Mol Cell* **81**: 1935–1950.e6. doi:10.1016/j.molcel.2021.02.034
- Takahashi H, Kato S, Murata M, Carninci P. 2012a. CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. *Methods Mol Biol* **786**: 181–200. doi:10.1007/978-1-61779-292-1_11
- Takahashi H, Lassmann T, Murata M, Carninci P. 2012b. 5' End-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat Protoc* **7**: 542–561. doi:10.1038/nprot.2012.005
- Taliaferro JM, Vidaki M, Oliveira R, Olson S, Zhan L, Saxena T, Wang ET, Graveley BR, Gertler FB, Swanson MS, et al. 2016. Distal alternative last exons localize mRNAs to neural projections. *Mol Cell* **61**: 821–833. doi:10.1016/j.molcel.2016.01.020
- Tang DTP, Plessy C, Salimullah M, Suzuki AM, Calligaris R, Gustincich S, Carninci P. 2013. Suppression of artifacts and barcode bias in high-throughput transcriptome analyses utilizing template switching. *Nucleic Acids Res* **41**: e44. doi:10.1093/nar/gks1128
- Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN. 2020. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* **11**: 1438. doi:10.1038/s41467-020-15171-6
- Tian B, Manley JL. 2016. Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* **18**: 18–30. doi:10.1038/nrm.2016.116
- Tian L, Jabbari JS, Thijssen R, Gouil Q, Amarasinghe SL, Voogd O, Kariyawasam H, Du MRM, Schuster J, Wang C, et al. 2021. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biol* **22**: 310. doi:10.1186/s13059-021-02525-6
- Tilgner H, Grubert F, Sharon D, Snyder MP. 2014. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc Natl Acad Sci* **111**: 9869–9874. doi:10.1073/pnas.1400447111
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–515. doi:10.1038/nbt.1621
- Ugolini C, Mulroney L, Leger A, Castelli M, Criscuolo E, Williamson MK, Davidson AD, Almuqrin A, Giambruno R, Jain M, et al. 2022. Nanopore ReCapable sequencing maps SARS-CoV-2 5' capping sites and provides new insights into the structure of sgRNAs. *Nucleic Acids Res* **50**: 3475–3489. doi:10.1093/nar/gkac144
- van Dijk EL, Naquin D, Gorrichon K, Jaszczyzyn Y, Ouazahrou R, Thermes C, Hernandez C. 2023. Genomics in the long-read sequencing era. *Trends Genet* **39**: 649–671. doi:10.1016/j.tig.2023.04.006
- Verwilt J, Mestdagh P, Vandesompele J. 2023. Artifacts and biases of the reverse transcription reaction in RNA sequencing. *RNA* **29**: 889–897. doi:10.1261/rna.079623.123
- Vilborg A, Steitz JA. 2017. Readthrough transcription: How are DoGs made and what do they do? *RNA Biol* **14**: 632–636. doi:10.1080/15476286.2016.1149680
- Volden R, Palmer T, Byrne A, Cole C, Schmitz RJ, Green RE, Vollmers C. 2018. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc Natl Acad Sci* **115**: 9726–9731. doi:10.1073/pnas.1806447115
- Wang Z, Gerstein M, Snyder M. 2009. RNA-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**: 57–63. doi:10.1038/nrg2484
- Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, Lu Z, Olson A, Stein JC, Ware D. 2016. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat Commun* **7**: 11708. doi:10.1038/ncomms11708
- Wang R, Nambiar R, Zheng D, Tian B. 2018. Polya_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res* **46**: D315–D319. doi:10.1093/nar/gkx1000
- Wang J, Li B, Marques S, Steinmetz LM, Wei W, Pelechano V. 2020. TIF-Seq2 disentangles overlapping isoforms in complex human transcriptomes. *Nucleic Acids Res* **48**: e104. doi:10.1093/nar/gkaa691

- Wang F, Xu Y, Wang R, Zhang B, Smith N, Notaro A, Gaerlan S, Kutschera E, Kadash-Edmondson KE, Xing Y, et al. 2023. TEQUILA-seq: a versatile and low-cost method for targeted long-read RNA sequencing. *Nat Commun* **14**: 4760. doi:10.1038/s41467-023-40083-6
- White LK, Strugar SM, MacFadden A, Hesselberth JR. 2023. Nanopore sequencing of internal 2'-PO₄ modifications installed by RNA repair. *RNA* **29**: 847–861. doi:10.1261/rna.079290.122
- Wongsurawat T, Jenjaroenpun P, Wanchai V, Nookaew I. 2022. Native RNA or cDNA sequencing for transcriptomic analysis: a case study on *Saccharomyces cerevisiae*. *Front Bioeng Biotechnol* **10**: 842299. doi:10.3389/fbioe.2022.842299
- Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, et al. 2019. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods* **16**: 1297–1305. doi:10.1038/s41592-019-0617-2
- Wright CJ, Smith CWJ, Jiggins CD. 2022a. Alternative splicing as a source of phenotypic diversity. *Nat Rev Genet* **23**: 697–710. doi:10.1038/s41576-022-00514-4
- Wright DJ, Hall NAL, Irish N, Man AL, Glynn W, Mould A, Angeles ADL, Angiolini E, Swarbreck D, Gharbi K, et al. 2022b. Long read sequencing reveals novel isoforms and insights into splicing regulation during cell state changes. *BMC Genomics* **23**: 42. doi:10.1186/s12864-021-08261-2
- Wulf MG, Maguire S, Humbert P, Dai N, Bei Y, Nichols NM, Corrêa IR, Guan S Jr. 2019. 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* **294**: 18220–18231. doi:10.1074/jbc.RA119.010676
- Wulf MG, Maguire S, Dai N, Blondel A, Posfai D, Krishnan K, Sun Z, Guan S, Corrêa IR. 2022. Chemical capping improves template switching and enhances sequencing of small RNAs. *Nucleic Acids Res* **50**: e2. doi:10.1093/nar/gkab861
- Wyman D, Balderrama-Gutierrez G, Reese F, Jiang S, Rahamanian S, Forner S, Mattheos D, Zeng W, Williams B, Trout D, et al. 2020. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRxiv* doi:10.1101/672931
- Zhang S, Aibara S, Vos SM, Agafonov DE, Lührmann R, Cramer P. 2021. Structure of a transcribing RNA polymerase II-U1 snRNP complex. *Science* **371**: 305–309. doi:10.1126/science.abf1870
- Zhang Z, Bae B, Cuddleston WH, Miura P. 2023. Coordination of alternative splicing and alternative polyadenylation revealed by targeted long read sequencing. *Nat Commun* **14**: 5506. doi:10.1038/s41467-023-41207-8
- Zhao C, Liu F, Pyle AM. 2018. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* **24**: 183–195. doi:10.1261/rna.063479.117

Received May 7, 2024; accepted in revised form August 16, 2024.



Challenges in identifying mRNA transcript starts and ends from long-read sequencing data

Ezequiel Calvo-Roitberg, Rachel F. Daniels and Athma A. Pai

Genome Res. 2024 34: 1719-1734

Access the most recent version at doi:[10.1101/gr.279559.124](https://doi.org/10.1101/gr.279559.124)

Supplemental Material <http://genome.cshlp.org/content/suppl/2024/10/30/gr.279559.124.DC1>

References This article cites 153 articles, 28 of which can be accessed free at: <http://genome.cshlp.org/content/34/11/1719.full.html#ref-list-1>

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>