Characterization of *in vitro* transcribed mRNA using NEBNext® Companion Module for Oxford Nanopore Technologies® Direct RNA Sequencing

Evan Janzen Ph.D., Matthew Campbell Ph.D., Chen Song Ph.D., New England Biolabs Inc.

INTRODUCTION

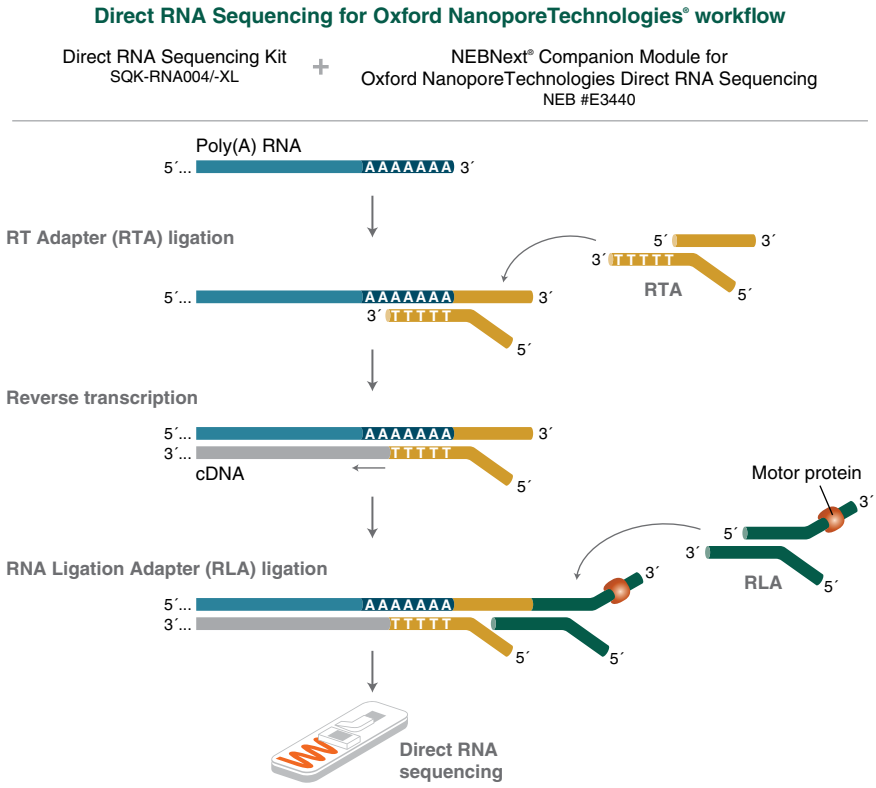
The rapid advancement of *in vitro* transcribed (IVT) mRNA therapeutics has increased the need for analytical methods that ensure product quality, safety, and consistency. IVT mRNA molecules are inherently complex, with multiple critical quality attributes (CQAs), including sequence identity, poly(A) tail characteristics, transcript integrity, capping efficiency, and impurity profiles. Together, these attributes govern biological function and manufacturability. Traditional quality control (QC) approaches rely on a suite of orthogonal assays, each targeting a single attribute or a limited subset of attributes. Although effective, these approaches are often labor-intensive and time-consuming and provide only bulk or indirect measurements, limiting the ability to characterize molecular heterogeneity at the individual transcript level.

As mRNA modalities evolve to include longer constructs, chemically modified nucleotides, and increasingly sophisticated regulatory elements, there is a growing need for integrated, high-resolution analytical strategies capable of interrogating multiple attributes within a single assay. Oxford Nanopore Technologies' direct RNA sequencing addresses this need by enabling multiple CQAs to be assessed from the same sequencing dataset while retaining information about individual RNA molecules. Recent work[1,2] has shown general concordance between sequencing-derived measurements and established orthogonal methods, supporting the feasibility of direct RNA nanopore sequencing as a multi-attribute method (MAM) for IVT mRNA analysis. Table 1 (page 2) summarizes representative mRNA quality attributes, corresponding metrics derived from direct RNA sequencing, and established orthogonal methods commonly used to assess each attribute.

Successful implementation of direct RNA sequencing for IVT mRNA requires more than access to sequencing hardware and analysis tools. Reliable mRNA characterization also depends on robust library preparation, fit-for-purpose workflow optimization, and careful interpretation of complex read-level outputs. The NEBNext Companion Module for Oxford Nanopore Technologies Direct RNA Sequencing provides a streamlined reagent solution for generating high-quality libraries suitable for IVT mRNA characterization. When paired with in-depth data assessment, this workflow supports informed evaluation of key sequencing metrics and a clearer understanding of the capabilities and limitations of direct RNA sequencing for IVT mRNA characterization.

MATERIALS

- FLuc mRNA (NIST RGTM 10202)
- NEBNext® Companion Module for Oxford Nanopore Technologies® Direct RNA Sequencing (NEB #E3440)
- Oxford Nanopore Technologies Direct RNA Sequencing Kit (Oxford Nanopore Technologies, #SQK-RNA004)
- *E. coli* Poly(A) Polymerase (NEB #M0276)



Direct RNA sequencing libraries are prepared following the Oxford Nanopore Technologies Direct RNA Sequencing Kit (SQK-RNA004) protocol, using the NEBNext® Companion Module for Oxford Nanopore Technologies® Direct RNA Sequencing. A poly(T)-containing RT Adapter is first ligated to poly(A)-tailed RNA, enabling selective capture of polyadenylated transcripts and interrogation of poly(A) tail characteristics. Reverse transcription generates an RNA-cDNA hybrid that helps reduce RNA secondary structure and improve sequencing performance. A sequencing adapter carrying the nanopore motor protein (RNA Ligation Adapter) is then ligated to the RNA strand, enabling sequencing to initiate at the 3′ end of the transcript on Oxford Nanopore Technologies flow cells.

In this application note, we use direct RNA sequencing to assess selected CQAs of an *in vitro* transcribed Firefly Luciferase (FLuc) mRNA developed by the National Institute of Standards and Technology (NIST). Designated NIST Research-Grade Test Material 10202 (NIST RGTM 10202), this material was developed to support harmonized CQA assessment for mRNA drug substances and is being evaluated in an interlaboratory study for its suitability as a reference standard or quality control material. We evaluate sequence identity, poly(A) tail length, transcript integrity, and low-abundance antisense transcripts. We also examine how sequence context, poly(A) tail structure, and read-level representation affect interpretation of these sequencing-derived measurements. Together, the sample preparation and data interpretation approaches presented here show how complementary measurements derived from a single direct RNA sequencing dataset can support comprehensive mRNA characterization.

METHODS

NIST RGTM 10202 mRNA Test Material

A unit of RGTM 10202 consists of one 0.5 ml-cryovial containing approximately 50 µl of *in vitro* transcribed (IVT), purified FLuc mRNA (approximately 25 µg) in nuclease-free water, pH 6.4. The FLuc mRNA transcript is 1,937 nucleotides in length and contains canonical, unmodified nucleobases, an enzymatically added Cap1 structure at the 5´end, and a segmented, template-encoded poly(A) tail consisting of two poly(A) tracts separated by a short restriction-site insertion for improved clonal stability (Figure 1).

Direct RNA Sequencing Library Preparation

Direct RNA sequencing libraries were prepared from NIST RGTM 10202 FLuc mRNA following the Oxford Nanopore Technologies Direct RNA sequencing workflow (SQK-RNA004) in conjunction with the NEBNext Companion Module for Oxford Nanopore Technologies Direct RNA Sequencing (workflow, page 1). Three replicate libraries were each prepared from 500 ng input mRNA and sequenced on an Oxford Nanopore GridION® instrument.

To enable capture of transcripts not containing the encoded poly(A) tail, an additional set of libraries was prepared following 3´ polyadenylation with *E. coli* Poly(A) Polymerase, performed according to the Oxford Nanopore Technologies protocol[4].

Data Processing and Analysis

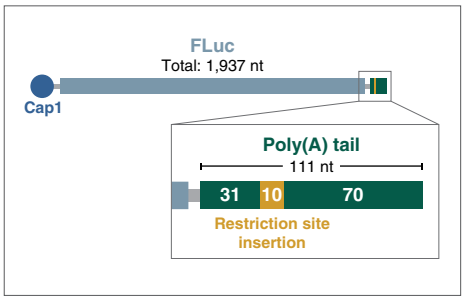
Sequencing data were basecalled using Dorado (v1.1.1) with the RNA004 SUP model (rna004_130bps_sup@v5.2.0) and sequencing datasets were downsampled to one million reads prior to analysis. Reads were aligned to the FLuc reference sequence using minimap2 or BLAST, with additional analyses performed using samtools, Qualimap, and the Dorado poly(A) tail estimation tool. Attribute-specific analytical methods are described in the sections below.

TABLE 1: **Quality attributes measurement with direct RNA sequencing and orthogonal methods**

QUALITY ATTRIBUTE	DIRECT RNA SEQUENCING DERIVED METRIC	ORTHOGONAL MEASURE
Sequence identity	Consensus accuracy, mismatch frequency	NGS / Sanger sequencing
Transcript integrity	Full-length read fraction, read-length distribution	Capillary electrophoresis (CE)
Poly(A) tail length	Tail length estimation from signal dwell time	LC-MS
Truncation /fragmentation	5′–3′ coverage drop-off, read truncation profiles	CE
Sequence variants	Variant calling from aligned reads	NGS
Cap presence	5′-cap derived reads	Cap-specific assays (e.g., LC-MS, enzymatic)
dsRNA impurities	Antisense reads	ELISA (J2 antibody), dot blot
Modified nucleotides	Signal deviation models, basecalling error patterns	LC-MS/MS

Adapted from the U.S. Pharmacopeia draft guidelines, Analytical Procedures for Quality of mRNA Vaccines and Therapeutics[3]

FIGURE 1: **NIST RGTM 10202 FLuc mRNA**



Schematic of the transcript structure.

RESULTS

Sequence Identity

Sequence identity is a fundamental attribute of mRNA characterization, providing confirmation that the expected transcript sequence was successfully generated. To evaluate sequence identity of the FLuc mRNA, direct RNA sequencing reads were aligned to the reference sequence using minimap2 and identity was calculated at each position as the percentage of aligned reads matching the expected base. Sequence identity and error frequencies were determined from aligned read pileups generated by samtools pileup.

Across the 1,937 nt FLuc gene, direct RNA sequencing exhibited excellent agreement with the expected sequence, with an average positional identity of 99.14% (Figure 2A, page 3). Although identity appeared high across most positions, some localized regions exhibited noticeably reduced identity relative to the surrounding sequence that was consistent across replicates.

To better understand these low-identity positions, sequencing errors were categorized as mismatches, insertions, or deletions. Among errors, apparent deletions represented the most abundant class, accounting for 54.2% of all observed errors, followed by mismatches (29.3%) and insertions (16.5%) (Figure 2B, page 3).

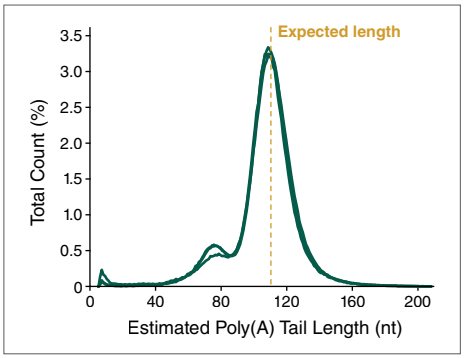
Inspection of the affected positions revealed that the largest decreases in identity occurred within homopolymeric sequence contexts. As homopolymer length increased, insertion and deletion frequencies rose considerably, whereas mismatch frequencies remained comparatively low across homopolymeric contexts (Figure 2C). This trend is consistent with a well-recognized challenge in nanopore sequencing, where determining the precise length of homopolymeric stretches remains more difficult than identifying the underlying nucleotide composition, resulting primarily in insertion and deletion errors.

While these sequence-specific errors remain an important consideration when interpreting direct RNA sequencing data, they were largely confined to a limited number of positions and did not substantially impact overall sequence determination. The high sequence identity observed across the transcript enabled accurate confirmation of the expected FLuc consensus sequence, demonstrating that consensus-level sequence characterization can remain highly accurate despite residual per-read sequencing errors.

Poly(A) Tail Length

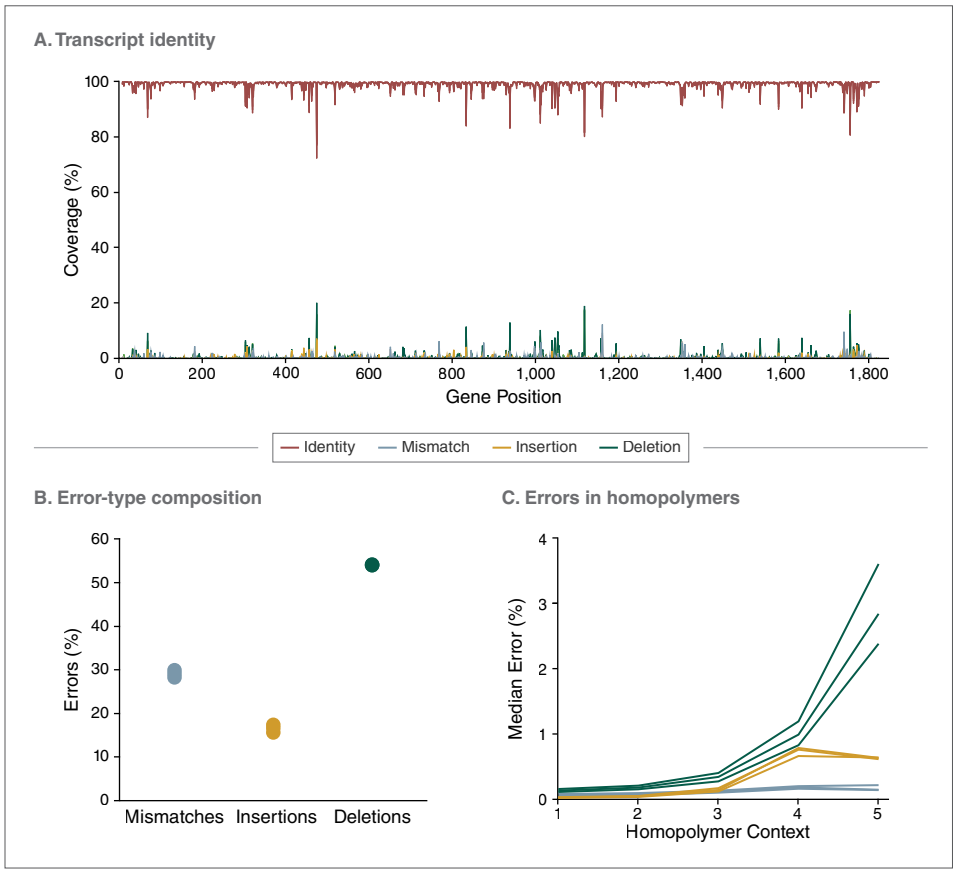
Poly(A) tails play important roles in mRNA stability, translation efficiency, and intracellular fate, making tail length a commonly evaluated attribute during mRNA characterization. One advantage of direct RNA sequencing is the ability to estimate poly(A) tail lengths directly from sequencing data without requiring separate library preparation or tail-specific assays. Notably, poly(A) tail estimation is performed using a dedicated analysis tool within Dorado rather than being derived solely from the basecalled sequence.

FIGURE 3: **Poly(A) tail length estimation**



Poly(A) tail lengths were estimated from direct RNA sequencing data using the poly(A) analysis tool built into Dorado, allowing insertions up to 10 nt. Shown are the percentage of aligned reads of each estimated tail length, with the vertical dashed line indicating the expected tail length of 111 nucleotides from the plasmid template. Most reads exhibited tail lengths near the expected value, with 78.3% of analyzed transcripts falling within 20 nucleotides of the expected length.

FIGURE 2: **Sequence identity analysis**



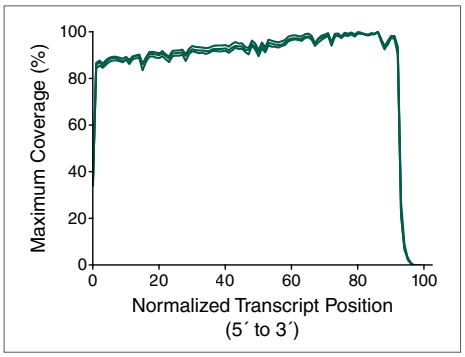
(A) Per-position sequence identity across the FLuc gene, shown as the percentage of aligned reads matching the reference sequence at each position. The percentages of reads containing mismatches, insertions, and deletions at each position are also shown. Sequence identity exceeded 99% across most positions of the transcript. (B) Composition of sequencing errors, presented as the percentage of each error type among all observed errors. Apparent deletions represented the most abundant error class. (C) Error frequencies as a function of homopolymer context. Homopolymer context corresponds to the number of consecutive identical bases beginning at the position of interest, where 1 indicates an isolated base and larger values indicate progressively longer homopolymer stretches. Error rates increased substantially with increasing homopolymer length, particularly for insertion and deletion events.

The FLuc mRNA contains a segmented, template-encoded poly(A) tail consisting of a 31-nt poly(A) tract followed by a 10-nt insertion and a second 70-nt poly(A) tract, resulting in an expected total tail length of 111 nucleotides (workflow, page 1). This type of segmented poly(A) tail design has been observed to improve stability of plasmid templates containing long homopolymeric sequences. To accommodate the intervening sequence, tail lengths were estimated using the Dorado poly(A) estimation tool while allowing a tail interrupt length of up to 10 nucleotides.

Most tail length estimates clustered near the expected value, with median and modal tail length estimates of 109 nucleotides (Figure 3). Among reads with a reported tail-length estimate, 78.3% fell within 20 nucleotides of the expected tail length, indicating that direct RNA sequencing was generally able to recover the overall tail architecture of the transcript. Approximately 9.8% of aligned reads did not receive a tail-length assignment from Dorado.

In addition to the primary population near the expected length, a second peak was observed at approximately 75 nucleotides. Because this secondary population was not observed in analyses of transcripts containing uninterrupted poly(A) tails, it likely reflects the unique segmented structure of the FLuc mRNA, rather than a separate population in the sample. One possible explanation is that the intervening 10-nt insertion disrupts tail-length estimation for a subset of molecules, causing the software to preferentially measure only the terminal poly(A) tract.

FIGURE 4: **Transcript coverage**



Normalized transcript coverage across the FLuc transcript determined using the Qualimap RNA-seq tool. Coverage is plotted as a percentage of the maximum observed coverage as a function of normalized transcript position from the 5′ to 3′ end of the transcript. Coverage remained high across the majority of the transcript, with a modest decrease observed at the 5′ end.

Despite this complexity, direct RNA sequencing successfully estimated the expected poly(A) tail length for the majority of transcripts analyzed. As with sequence identity analysis, the most reliable conclusions were obtained from consensus measurements across the sample rather than individual per-read estimates. These results support the use of direct RNA sequencing as a practical tool for poly(A) tail characterization while highlighting important considerations when interpreting complex engineered tail structures.

Transcript Integrity

Transcript integrity describes the extent to which RNA molecules retain their expected full-length structure. It is therefore an important attribute of mRNA characterization. Truncated, fragmented, or otherwise incomplete transcripts can have significant impacts on biological activity. To assess transcript integrity of the FLuc mRNA, transcript coverage was first examined across the full reference sequence.

Normalized transcript coverage remained high across the majority of the FLuc transcript, with >80% coverage observed near the 5′ end of the molecule (Figure 4). Because direct RNA sequencing initiates from the poly(A) tail at the 3′ end of each transcript, progressive loss of coverage toward the 5′ end is often expected and can provide a useful indication of transcript completeness. However, as transcript coverage reflects the fraction of reads contributing sequence information at each position rather than the fraction of molecules spanning the entire transcript, read length distributions and read alignments were examined to provide a more complete assessment of transcript integrity.

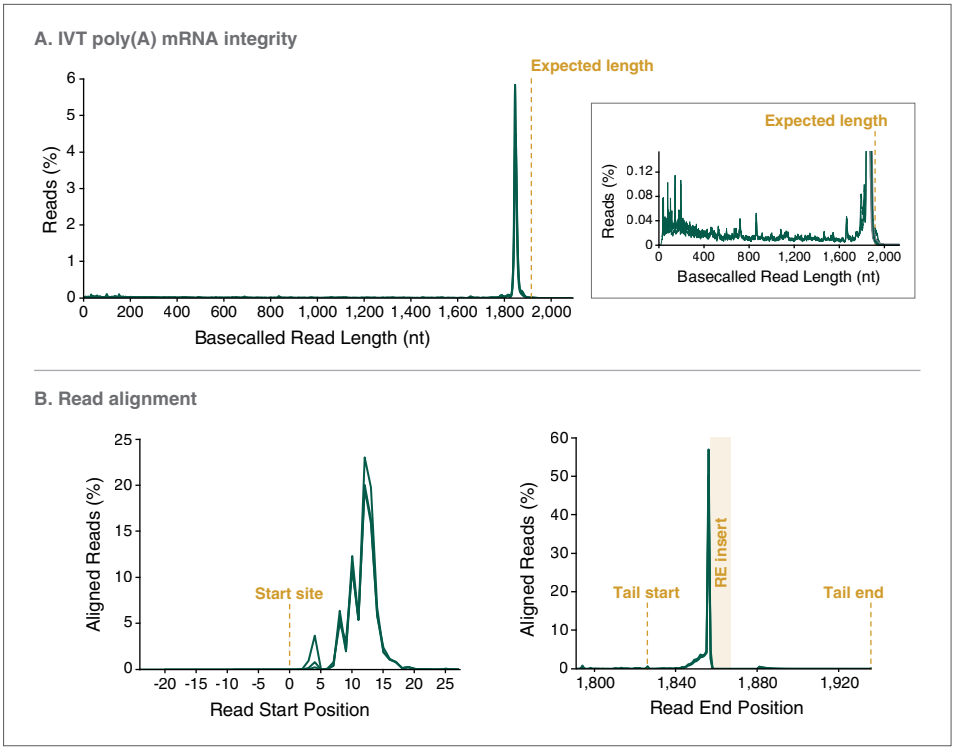
Examination of basecalled read lengths revealed a prominent peak centered near 1,850 nucleotides that comprised more than 70% of sequenced reads (Figure 5A), indicating a highly uniform transcript population. However, this peak was noticeably shorter than the expected transcript length of 1,937 nucleotides.

To better understand this discrepancy, aligned read start and end positions were examined relative to the expected transcript structure (Figure 5B). Although direct RNA sequencing initiates at the poly(A) tail, basecalled reads are reported in the conventional 5′→3′ orientation. Here, read start and end positions refer to the aligned boundaries of the reported sequence. Read start positions clustered near, but downstream of, the transcription start site, consistent with the expected loss of sequence information from the extreme 5′ end of transcripts as the RNA terminus traverses the nanopore without the guidance of the motor protein. Similarly, read end positions were concentrated near the end of the first poly(A) tract rather than the end of the full transcript, indicating that a substantial portion of the poly(A) tail is not represented in the basecalled sequence despite contributing signal used for poly(A) tail estimation. Together, these observations explain the approximately 87-nucleotide difference between the observed and expected transcript lengths.

Examination of the read boundary distributions showed that 76.4% of reads initiated within 20 nucleotides of the expected transcription start site, while 86.3% of reads terminated within a window extending from 20 nucleotides upstream of the poly(A) tail through the end of the reference sequence. To provide a more stringent assessment of transcript integrity, full-length transcripts were defined as reads satisfying both of these boundary criteria. Using this definition, 75.3% of reads were classified as full-length across sequencing replicates.

Notably, the fraction of full-length reads was lower than might be inferred from transcript coverage alone. While the majority of reads terminated within the expected end-position

FIGURE 5: **Transcript integrity analysis**



(A) Distribution of basecalled read lengths obtained from direct RNA sequencing. The vertical dashed line indicates the expected transcript length of 1,937 nucleotides. A magnified view of shorter reads is shown on the right. (B) Distribution of aligned read start positions relative to the expected transcription start site. Distribution of aligned read end positions relative to the template-encoded poly(A) tail. Positions corresponding to the start and end of the poly(A) tail, as well as the intervening 10-nt insertion, are indicated. The majority of reads begin near the transcription start site and terminate near the beginning of the insertion site within the segmented poly(A) tail.

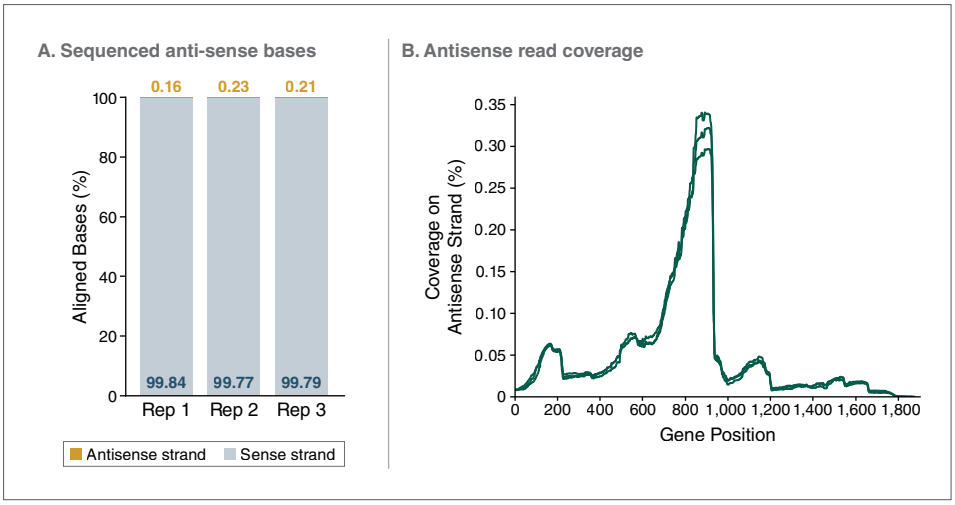
window, a subset of reads exhibited internal termination sites throughout the transcript. These reads contribute sequence information and transcript coverage but do not satisfy the criteria used for classification as full-length transcripts. Further analysis indicated that internally-terminating reads were substantially more likely to lack a reported poly(A) tail estimate than reads ending within the expected tail region, providing additional confidence that the Dorado tail estimation tool is successfully assigning tail lengths to the majority of reads reaching the poly(A) tail. Additionally, if some of these internally-terminating reads arise from transcript fragments captured during library preparation, the resulting full-length estimate may represent a conservative measure of transcript integrity by counting both 5′- and 3′-derived fragments from the same transcript as independent observations.

Taken together, these results indicate that the majority of molecules within the FLuc mRNA test material retain the expected transcript structure. At the same time, they highlight both the complexity of assessing transcript integrity from direct RNA sequencing data and several important considerations that can influence interpretation of the resulting metrics. Despite these challenges, direct RNA sequencing provides valuable insight into transcript structure and integrity at the level of individual molecules.

Antisense Transcripts

Double-stranded RNA (dsRNA) is a common byproduct of *in vitro* transcription and is often monitored because of its potential impact on downstream biological responses. One possible indicator of dsRNA-related impurities is the presence of antisense transcripts within the RNA population. To investigate antisense transcript contaminants using direct RNA sequencing, FLuc mRNA was treated with *E. coli* Poly(A) Polymerase prior to library preparation to facilitate capture of non-polyadenylated transcript species that would otherwise be underrepresented by the standard workflow.

FIGURE 6: **Characterization of antisense transcript contaminants**



FLuc mRNA was treated with *E. coli* Poly(A) Polymerase prior to library preparation to facilitate capture of non-polyadenylated transcript species. Reads were aligned to the FLuc reference sequence and classified according to alignment to the expected transcript strand or the antisense strand. (A) Fraction of sequenced bases mapping to the expected and antisense strands of the FLuc transcript. Antisense bases accounted for approximately 0.20% of the total aligned bases. (B) Positional distribution of antisense read coverage across the FLuc transcript. Although antisense transcripts represented a very small fraction of the overall sample, antisense reads were concentrated within a discrete region of the transcript rather than distributed uniformly across the gene sequence.

Reads were aligned to the FLuc reference sequence and classified according to alignment to either the expected transcript strand or the antisense strand. Across replicate libraries, antisense bases accounted for approximately 0.20% of all aligned bases (Figure 6A), suggesting that antisense transcripts represented only a minor component of the overall RNA population.

Although antisense transcripts were rare, direct RNA sequencing also provided information regarding their potential origin within the transcript. Antisense coverage was not distributed uniformly across the FLuc sequence but was instead concentrated within a discrete region originating near position 900 of the reference gene (Figure 6B), accounting for more than 50% of antisense reads. This localized distribution suggests that antisense transcript formation may be associated with specific features of the template or transcription process and could potentially be explained by cryptic promoter activity, abortive transcript priming, or other sequence-dependent events occurring during *in vitro* transcription.

Taken together, these results demonstrate how direct RNA sequencing can be used to both quantify antisense transcript contaminants and investigate their origin within an mRNA construct. While antisense transcripts represented only a small fraction of the FLuc mRNA sample, the positional information provided by sequencing may help guide construct design and process optimization efforts aimed at reducing dsRNA-related impurities.

CONCLUSIONS

Direct RNA sequencing provides a powerful approach for characterization of *in vitro* transcribed mRNA, enabling assessment of multiple quality attributes from a single instrument platform. Using the NIST RGTM 10202 FLuc mRNA test material, direct RNA sequencing was used to evaluate sequence identity, poly(A) tail length, transcript integrity, and antisense transcript contaminants, demonstrating the breadth of information that can be obtained through integrated analysis of nanopore sequencing data.

Across these analyses, the expected properties of the FLuc transcript were generally recovered, including accurate confirmation of the consensus transcript sequence, poly(A) tail length estimates consistent with the designed transcript architecture, and evidence supporting a largely intact transcript population. In addition, direct RNA sequencing provided insight into low-abundance antisense transcript species and their potential origin within the construct.

At the same time, these studies highlight several important considerations when interpreting direct RNA sequencing data. Sequence-specific errors, complex engineered tail structures, incomplete transcript representations, and other workflow-specific characteristics can influence individual analytical metrics. As demonstrated throughout this application note, the most informative conclusions were often obtained by evaluating multiple sequencing-derived measurements together rather than relying on any single metric in isolation.

As mRNA therapeutics continue to evolve, there is increasing interest in analytical approaches capable of simultaneously characterizing multiple transcript attributes while preserving information at the level of individual RNA molecules. The workflow and analytical approaches described here demonstrate how direct RNA sequencing can contribute to this emerging paradigm, providing a practical and information-rich framework for multi-attribute characterization of IVT mRNA samples.

References

1. Gunter HM, et al., Nature Communications 2023 14 (1), 5663.
2. Chatla K, et al., Analytical Chemistry 2026 98 (10), 7452-7461.
3. U.S. Pharmacopeia. Analytical Procedures for Quality of mRNA Vaccines and Therapeutics (Draft Guidelines: 3rd Edition). Rockville (MD): USP; 2024. Available from:https://www.uspnf.com/notices/analytical-procedures-mrna-vaccines-20240802 [accessed 27 August 2026]
4. Oxford Nanopore Technologies. 3′ polyadenylation of RNA transcripts using *E. coli* poly(A) polymerase (PAP). Oxford (UK): Oxford Nanopore Technologies; [n.d.]. Available from: https://nanoporetech.com/document/extraction-method/3-poly-rna-ecoli-pap [accessed 27 August 2026].

Products and content are covered by one or more patents, trademarks and/or copyrights owned or controlled by New England Biolabs, Inc (NEB). The use of trademark symbols does not necessarily indicate that the name is trademarked in the country where it is being read; it indicates where the content was originally developed. See www.neb.com/trademarks. The use of these products may require you to obtain additional third-party intellectual property rights for certain applications. For more information, please email busdev@neb.com.

Your purchase, acceptance, and/or payment of and for NEB's products is pursuant to NEB's Terms of Sale at www.neb.com/support/terms-of-sale. NEB does not agree to and is not bound by any other terms or conditions, unless those terms and conditions have been expressly agreed to in writing by a duly authorized officer of NEB.

OXFORD NANOPORE TECHNOLOGIES® and GRIDION® are registered trademarks of Oxford Nanopore Technologies Limited.
B CORPORATION® is a registered trademark of B Lab Company.

© Copyright 2026, New England Biolabs, Inc.; all rights reserved.