Utilizing Next-Generation Sequencing and Third-Generation Sequencing for RNA Quality Attribute Monitoring

by Anna-Marei Böhm
Analytical Development Specialist

In pharmaceutical manufacturing, monitoring Critical Quality Attributes (CQAs) is crucial to ensure the quality and safety of drug substances and drug products. Analytical life cycle management should involve continuous improvement of existing procedures and evaluation of new analytical.

With RNA therapeutics rapidly expanding as a new important class in pharmaceutical manufacturing, there is an urgent requirement for new analytical methods for monitoring CQAs. Next-Generation Sequencing (NGS) and Third-Generation Sequencing (TGS) are valuable tools for assessing multiple CQAs in RNA drug substance and RNA drug product identity, RNA integrity, poly(A)-tail length, RNA mixing ratio, dsRNA and residual DNA, often even combined within one assay setup and most importantly at nucleotide resolution level.

This application note briefly explains NGS and TGS, highlights advantages and limitations of both platforms, and evaluates the applicability of NGS and TGS for monitoring and characterizing specific CQAs in RNA drug substance and drug product.

What is Next-Generation Sequencing and Third-Generation Sequencing?

Next-Generation Sequencing (NGS) refers to high-throughput sequencing technologies that allow the simultaneous sequencing of large numbers of DNA or RNA fragments. NGS has revolutionized genomics by enabling rapid and cost-effective sequencing.

NGS generates short reads of nucleotide sequences ranging from tens to hundreds of base pairs. The most widely used NGS technology is Illumina, and therefore this application note focuses on this NGS technology.

As part of the Illumina sequencing, DNA or RNA is fragmented, followed by ligation of sequencing adapters and, if applicable, barcoding of individual DNA or RNA fragments using Unique Molecular Identifiers (UMIs) for increased sequencing accuracy. Subsequently, reverse-transcription (RT) to cDNA (in the case of RNA sequencing), cluster generation of prepared cDNA or DNA fragments, and cluster amplification are performed. Illumina technology makes use of the Sequencing-by-Synthesis approach, where fluorescently labeled nucleotides are sequentially added to the clusters. Each incorporation event leads to emission of light signals; the color and intensity of the light signals indicate the identity of the incorporated base. Base calling software translates the fluorescence signals into a nucleotide sequence read. The full DNA or RNA sequence is reconstructed from the short reads ('assembled').

Third-Generation Sequencing (TGS) represents a newer generation of sequencing technologies that, in contrast to NGS, generate long sequencing reads, often spanning thousands to tens of thousands of base pairs.

The most frequently used TGS technologies include Nanopore and PacBio sequencing. Nanopore sequencing reads the nucleotide sequence as the (c)DNA or RNA molecule passes through a nanoscale pore. Multiple nanopores are embedded in a membrane with an electric potential. As the nucleotides pass through the pore, they temporarily block the ion flow, causing measurable changes in the electric current: the duration and amplitude of the changes correspond to specific nucleotide identities.

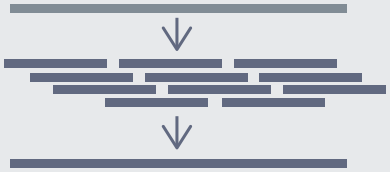
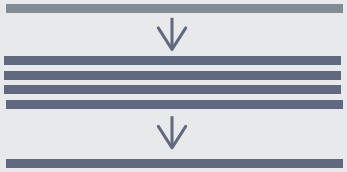
Base calling software converts the electrical signals into a nucleotide sequence. Different RNA sequencing methods are available for Nanopore sequencing, including Direct RNA sequencing, where the RNA molecule is directly sequenced, eliminating the risk of introducing reverse transcription (RT) and PCR errors. However, using Direct RNA sequencing for RNA containing modified nucleotides can lead to compromised accuracy and decreased sequencing yield as modifications might hinder smooth passage through the nanopore. Alternative nanopore methods for RNA are cDNA-PCR sequencing or direct cDNA sequencing (now commonly known as ligation sequencing) which include a reverse transcription step.

PacBio sequencing, another popular TGS technology developed by Pacific Biosciences, utilizes single-molecule, real-time (SMRT) sequencing to determine the sequence of DNA or RNA molecules immobilized on a SMRT cell. PacBio applies the Sequencing-by-Synthesis approach, outlined above for Illumina sequencing, to generate long reads.

Advantages and limitations of NGS/TGS

Each sequencing platform has specific advantages and limitations (Table 1) that need to be considered to select the most suitable sequencing approach for each CQA.

Table 1: Advantages and limitations of NGS and TGS
UMI = Unique Molecular Identifier, RT= reverse transcription, PCR= Polymerase-Chain-Reaction

	Next Generation Sequencing (NGS) Short read sequencing *e.g. Illumina*	**Third Generation Sequencing (TGS)** Long read sequencing *e.g. Nanopore, PacBio*
Platform		
Advantages	▪ High accuracy: 99.9% (>99.99% using UMIs) ▪ High sample throughput ▪ High sequencing speed ▪ Low sequencing costs	▪ Full-length transcript information ▪ No assembly errors ▪ Even read coverage ▪ No risk of RT and PCR biases/errors when using direct sequencing
Limitations	▪ No full-length transcript information ▪ Risk of assembly errors ▪ Risk of uneven read coverage ▪ Risk of RT and PCR biases/errors due to indirect sequencing	▪ Lower accuracy: 99% - 99.9% (depending on the method) ▪ Lower sample throughput (can be increased by multiplexing) ▪ Lower sequencing speed ▪ High sequencing costs ▪ Standard RNA sequencing methods depend on poly(A)-tail (Nanopore)

Advantages of NGS technologies include high accuracy (99.9%) with the potential to exceed 99.99% when utilizing Unique Molecular Identifiers (UMIs), which enable the identification and elimination of false positive variants during post-sequencing analysis. NGS provides both high sample throughput thanks to sample multiplexing, and swift sequencing speed due to massive parallel sequencing of DNA or RNA clusters, allowing quick and cost-effective generation of large datasets.

One of the biggest limitations of short read sequencing is the lack of full-length transcript information, meaning that information on the length of the different transcripts present in the sequenced sample is not retained using NGS. Another downside of short reads is vulnerability to assembly errors, particularly in repetitive or complex genomic regions, which impact the accuracy of full-length sequence reconstruction. Moreover, Reverse transcription (RT) and PCR amplification of sequencing libraries expose NGS to potential RT and PCR biases and errors, compromising sequencing accuracy and sequence coverage.

In contrast to NGS, TGS technologies offer full-length transcript information without the need for sequence assembly, thereby avoiding assembly errors and assuring more even read coverage. To this end, long read sequencing provides a more comprehensive and accurate representation of the entire DNA or RNA sequence than short-read sequencing. With the option of using direct sequencing methods, TGS also eliminates the risk of RT and PCR biases and errors.

However, TGS comes with its own set of challenges. It generally exhibits lower accuracy, ranging from 99% to 99.9%, depending on the specific method. TGS also tends to have lower sequencing speed and higher sequencing costs per base pair compared to NGS, and sample throughput is lower even when using multiplexing. Moreover, a limitation specifically of Nanopore sequencing is that existing methods for RNA sequencing rely on poly(A) tail-binding adapters; this means that adapted sequencing protocols using poly(A) tail-independent adapters or polyadenylation of RNA samples need to be developed for sequencing of transcripts without poly(A)-tail.

How can NGS and TGS be used to monitor and characterize specific CQAs?

By using RNA identity, RNA integrity, poly(A)-tail length and RNA mixing ratio, dsRNA, and residual DNA for monitoring and characterization at nucleotide resolution, a large number of CQAs in drug substance and drug product can benefit from sequencing as analytical platform. However, it is still essential to consider the specific advantages and limitations of the different sequencing platforms to select the most suitable sequencing approach for each CQA (Table 2).

Table 2: Application of NGS and TGS for monitoring and characterization of specific CQAs in RNA drug substance and drug product

UMI = Unique Molecular Identifier, RT= reverse transcription, PCR= Polymerase-Chain-Reaction

		Next Generation Sequencing (NGS) = Short read sequencing		**Third Generation Sequencing (TGS) = Long read sequencing**			
		Illumina		**Nanopore**			**PacBio**
		RNA-seq	*DNA-seq*	*Direct RNA-seq*	*cDNA-PCR-Seq (e.g. VAX-Seq)*	*Direct cDNA-Seq (e.g. Nano3P-Seq)*	*Full-length RNA-Seq*
Drug substance	RNA identity	■		■	■	■	■
	Poly(A)-tail length			■	■	■	■
	RNA integrity				■	■	■
	dsRNA			■			
	Residual DNA		■				
Drug product	RNA ratio	■					
	RNA identity	■		■	■	■	■
	RNA integrity				■	■	■

RNA Identity

Identity testing is performed to distinguish and confirm the correct sequence of the active ingredient and to ensure that there are no changes towards the reference sequence. RNA identity of RNA dug substance and drug product is traditionally assessed by quantitative PCR (qPCR), using one primer pair to amplify a small part of the RNA sequence. qPCR gives confirmation of RNA identity, but does not allow assessment of sequence correctness and identification of potential insertions, deletions, or single nucleotide variants (SNVs).

Multiple NGS and TGS methods can be used for identity confirmation and detection of insertions/deletions and abundant single nucleotide variations (SNVs). However, for detection of single nucleotide variations (SNVs) with lower variant frequency (< 1%), NGS is preferred over TGS due to its lower sequencing error rate. Illumina sequencing, making use of Unique Molecular Identifiers (UMIs) and allowing discrimination of true variants from sequencing errors, offers highest accuracy for SNV detection.

Poly(A) Tail Length

Poly(A) tail length can affect the half-life (by protecting RNA from degradation) and translation efficiency of the mRNA and thereby impact drug performance. Determination of Poly(A)-tail length and distribution in RNA drug substance is conventionally performed using ion-pair reversed-phase high performance liquid chromatography (IP-RP-HPLC) or Liquid Chromatography with tandem mass spectrometry (LC-MS/MS).

Sequencing as orthogonal method requires smaller sample volumes and allows poly(A)-tail length assessment in the same assay setup with other CQAs, while showing comparable accuracy to conventional methods. Since Illumina sequencing encounters challenges with low complexity sequences such as the homopolymeric poly(A)-tail, TGS is more suitable for poly(A)-tail characterization.

Nanopore Direct RNA and cDNA-PCR sequencing in combination with an adapted version of the established poly(A)-tail length analysis software for nanopore, tailfindR, have been successfully used for poly(A)-tail length determination by Gunter et al as part of their developed VAX-Seq protocol1, with slightly better accuracy for cDNA-PCR sequencing. Another nanopore-based method for poly(A)-tail determination that was recently published, Nanopore 3□ end-capture sequencing (Nano3P-Seq), is using an adapted direct cDNA sequencing protocol again in combination with an improved tailfinR algorithm2.

With PacBio Full-length RNA Seq another long read sequencing method was recently introduced as an effective method for poly(A) tail length determination.[3]

RNA Integrity

As part of purity/integrity testing the proportion of intact full length RNA molecules compared to shorter or longer species is assessed. mRNA should be intact from 5′ to 3′ to exert its function, which is the translation of the encoded protein of interest. Unwanted shorter or longer species may impact efficacy as impurities can contribute to an immunological response.

RNA purity/integrity of RNA drug substance and drug product is traditionally assessed through capillary gel electrophoresis (CGE) or IP-RP-HPLC. While these techniques allow detection and size estimation of different RNA fragments and the calculation of % purity, they do not provide information on the sequence of the detected RNA fragments.

Next to poly(A)-tail length determination, the above mentioned Nanopore methods VAX-Seq1 and Nano3P-Seq2 using an adapted, poly(A) tail-independent sequencing protocol as well as Pac-Bio Full length RNA-Seq3 allow for sequencing of all transcripts present in the RNA sample, including 3′ and 5′ truncated and full-length sequences. RNA fragments are thereby not only quantified in number and size, but also their sequence and type of truncation (3′, 5′) is identified, providing a comprehensive overview of the RNA integrity profile.

dsRNA

dsRNA is a process-related impurity that is generated by aberrant polymerase activity or residual DNA impurities. dsRNA can potentially induce or enhance immunogenicity activity and thus affect product quality, efficacy (protein translation), and safety.

Detection and quantification of dsRNA in RNA drug substance is commonly assessed by immunoblotting using slot blot and enzyme-linked immunosorbent assay (ELISA). Both methods rely on a dsRNA standard which, to allow accurate dsRNA quantification, should have the same length and composition with respect to the chemical characteristics of the used nucleotides as the dsRNA present in the analyzed sample.

To this end, length characterization of the dsRNA species in each RNA drug substance is imperative for the development of accurate dsRNA quantification assays. As demonstrated by Gunter et al1, characterization of off-target RNAs, within those antisense RNA species that react with the respective sense RNA to form dsRNA, can be achieved using nanopore direct RNA sequencing.

Sequencing of poly(A)-tail less antisense RNA is enabled by polyadenylation of the complete RNA sample prior to RNA library preparation. Direct RNA sequencing keeps the directional information of sequence reads, thereby allowing the identification of antisense reads and the characterization of the dsRNA species present in the RNA drug substance. To enrich dsRNA in low-dsRNA containing RNA samples, immunoprecipitation using J2-anti-dsRNA antibody or targeted digest of ssRNA using RNAse I could be considered prior to sequencing.

Residual DNA

Residual DNA is a process-related impurity that potentially can induce or enhance immunogenicity activity and thus affect product quality, efficacy (protein translation) and safety. Effective purification of the DNA start material during manufacturing typically is demonstrated by qPCR using primers binding to the backbone of the DNA template. However, this approach does not consider potential truncated residual DNA fragments that might be present in the sample but are lacking the specific primer binding sites due to the truncation.

Characterizing the length and sequence of residual DNA in RNA drug substance would enable the development of more targeted residual DNA detection and quantification methods, e.g. being able to align primer design to the residual DNA profile to ensure that the complete residual DNA population in samples is quantified accurately. Due to low amounts of residual DNA in RNA samples, Illumina DNA sequencing using ultra-low sample input protocol with high sequencing depth and a potential upfront RNA depletion step seems might be a suitable sequencing approach to characterize residual DNA.

RNA ratio

The RNA ratio in drug product consisting of a mixture of multiple RNAs can impact product quality and efficacy. Verification of the target RNA ratio in RNA drug product is currently performed using qPCR. Illumina-sequencing based differential expression analysis of the different RNAs in the drug product could be used as an orthogonal approach for accurate quantification of RNA ratios.

Conclusion

Integration of NGS and TGS, each with their unique strengths, in RNA manufacturing provides a valuable tool for monitoring and characterizing RNA drug substance and drug product CQAs RNA identity, poly(A) tail length, RNA integrity, RNA mixing ratio, dsRNA and residual DNA. Sequencing provides comprehensive understanding of CQAs at nucleotide resolution level, and thus can be beneficial for improvement of existing, traditionally used analytical methods. In addition, it allows multiple CQAs to be addressed within one assay setup (see e.g. VAX-Seq, Nano3P-Seq), rendering it an innovative and versatile analytical platform in RNA therapeutics.

etherna is currently developing pipelines for the implementation of the outlined NGS and TGS approaches for the different CQAs and providing support to clients according to their requirements to development those innovative, sequencing-based analytical approaches.

“Integration of NGS and TGS, each with their unique strengths, in RNA manufacturing provides a valuable tool for monitoring and characterizing RNA drug substance and drug product CQAs RNA identity, poly(A) tail length, RNA integrity, RNA mixing ratio, dsRNA...”

References

1. Gunter, H.M., Idrisoglu, S., Singh, S. et al. mRNA vaccine quality analysis using RNA sequencing. Nat Commun 14, 5663 (2023).
2. Begik, O., Diensthuber, G., Liu, H. et al. Nano3P-seq: transcriptome-wide analysis of gene expression and tail dynamics using end-capture nanopore cDNA sequencing. Nat Methods 20, 75–85 (2023).
3. Tech note Azenta Life Sciences: Full-Length RNA-Seq: A Novel Method to Assess Sequence Integrity for RNA Therapeutics

etherna

For more information feel free to reach out

+32 3 369 17 40
info@etherna.be
Galileilaan 19, 2845 Niel, Belgium

www.etherna.be