Università
Ca'Foscari
Venezia

Corso di Dottorato di ricerca in
Scienze e Tecnologie dei Bio e
Nanomateriali
ciclo 38°

Tesi di Ricerca

Implementation of NGS technologies to speed up and improve the rAAV and IVT mRNA production stage for Cell and Gene Therapy

SSD: BIO/11 – Biologia Molecolare

Coordinatore del Dottorato
ch. prof. Flavio Rizzolio

Supervisore
ch. prof. Flavio Rizzolio

Dottoranda
Violina Potlog
Matricola: 956750

Anno Accademico: 2025-2026

Table of contents

Abstract

Cell and gene therapies (CGTs) are among the most transformative fields of modern medicine, yet their development is slowed by a critical bottleneck: the design and sequence verification of recombinant adeno-associated virus (rAAV) vectors and the DNA templates used for in vitro transcribed (IVT) mRNAs. These molecules must be error-free to ensure safety and efficacy, but their structural complexity, including inverted terminal repeats (ITRs) in rAAV plasmids and poly(A) tails in IVT mRNA templates, renders them extremely challenging to validate. Conventional workflows, that are based on Sanger sequencing, are not suited to these complex regions and are not scalable for high-throughput screening. This means that CGT manufacturers must perform repeated rounds of cloning and sequencing, which can delay construct validation for months and increase costs.

This thesis addresses this challenge by implementing next-generation sequencing (NGS) technologies as a robust quality control (QC) framework for rAAV plasmids and IVT mRNA templates. Oxford Nanopore Technologies (ONT) was benchmarked as a high-throughput platform capable of sequencing entire plasmids, including structurally inaccessible ITRs, and of directly sequencing full-length IVT mRNAs containing only standard ribonucleotides. Illumina sequencing was evaluated as a complementary solution, providing base-level accuracy and enabling precise analysis of homopolymeric regions such as poly(A) tails. By integrating these platforms into a dual-technology pipeline, our Biofoundry demonstrated the capacity to verify hundreds of DNA constructs in parallel, confirm the integrity of long and complex plasmids, and extend QC to RNA molecules with unmodified ribonucleotides that had so far been characterized only by indirect methods.

The results show that, thanks to this integrated pipeline, we can reduce the time and costs required to obtain fully sequence-verified constructs. This acceleration allows CGT development companies to move more quickly from design to in vivo testing, enabling faster translation of new therapies into the clinical setting. By introducing an NGS-based Quality Control pipeline specifically designed for rAAV and IVT mRNA synthesis, this thesis contributes to providing a practical solution to one of the most pressing

bottlenecks in CGT development, directly supporting the industrial production of safe and effective advanced therapies.

List of Abbreviations

The following list summarizes all abbreviations used throughout this thesis.
Abbreviations are reported in alphabetical order together with their full names.

AAV: Adeno-Associated Virus

AAP: Assembly-Activating Protein

AMX: Adapter Mix

ATMP: Advanced Therapy Medicinal Products

BCL: Base Call file format (Illumina)

BFU: Biofoundry Unit

DBTL: Design-Build-Test-Learn cycle

BLT: Bead-Linked Transposomes

Cap: Capsid proteins

CGT: Cell and Gene Therapy

cDNA: complementary DNA

DNA: Deoxyribonucleic Acid

dsDNA: Double-Stranded DNA

dNTPs: deoxynucleotide Triphosphates

EB: Elution Buffer

EPM: Enhanced PCR Mix

EPI2ME:Oxford Nanopore Technologies cloud-based analysis software

EXP-CTL001: Control Expansion Kit (ONT)

EXP-WSH004: Flow Cell Wash Kit (ONT)

€: Euro

FAST5: Raw Nanopore signal data format

FASTQ: Sequence and quality data file format

FB: Flush Buffer

FCF: Flow Cell Flush

FLT: Flush Tether

HT1: Hybridization Buffer

IPB: Illumina Purification Beads

ITR: Inverted Terminal Repeat

IVT: In Vitro Transcribed

LB: Loading Beads

LIS: Library Solution (ONT)

LNB: Ligation Buffer

mRNA: messenger Ribonucleic Acid

NGS: Next-Generation Sequencing

OB: Officinae Bio

ONT: Oxford Nanopore Technologies

ORF: Open Reading Frame

PCR: Polymerase Chain Reaction

QC: Quality Control

Q score (Q10, Q20, Q30, Q40): Probability of an incorrect base call

RCS: RNA Control Strand

REB: RNA Elution Buffer

Rep: Replication proteins

RFT: RNA Flush Tether

RNA: Ribonucleic Acid

RIN: RNA Integrity Number

RLA: RNA Ligation Adapter

RSB: Resuspension Buffer

RT: Reverse Transcription

RTA: Reverse Transcription Adapter

S: Storage Buffer

SQB: Sequencing Buffer

SQK-LSK109: Ligation Sequencing Kit (ONT)

SQK-RBK110.96: Rapid Barcoding Kit 96 (ONT)

SQK-RNA004: Direct RNA Sequencing Kit (ONT)

TB1: Tagmentation Buffer 1

TSB: Tagment Stop Buffer

TWB: Tagment Wash Buffer

UDI: Unique Dual Indexes

USD: United States Dollar

UTR: Untranslated Region

VF: Variant Frequency

VP1–3: Viral Capsid Proteins

WMX: Wash Mix

WSB: Wash Buffer

1 Introduction

1.1 Background

In the last decades, synthetic biology has emerged as a transformative discipline, shifting biology from a descriptive science to an engineering practice. Instead of only analyzing existing biological systems, researchers now design and construct new ones, using tools from molecular biology, computational modeling, and automation. This approach has already demonstrated its impact in many areas, including biosensing, sustainable biomanufacturing, biofuels, biomaterials, and the development of innovative pharmaceuticals (Khalil & Collins, 2010).

Biofoundries were established around the world to support these advances. These facilities integrate the design-build-test-learn (DBTL) cycle with robotics, high-throughput screening, and data analysis. Their role is to standardize and accelerate synthetic biology, making workflows more reproducible, scalable and accessible. The importance of this infrastructure has been recognized globally, leading to the creation of the Global Biofoundries Alliance. This coordinates efforts to serve academic, industrial, and translational environments (Hillson et al., 2019).

Cell and gene therapy (CGT) has emerged as one of the most transformative applications in synthetic biology. CGTs has the potential to address the underlying causes of disease by repairing defective genes or by delivering therapeutic proteins directly in vivo. They are therefore different from conventional drugs, which typically target only symptoms. Clinically, CGTs have already demonstrated life-changing effects in patients with genetic disorders, cancer and rare diseases: multiple therapies received regulatory approval in recent years (Naso et al., 2017; Pardi et al., 2018).

CGTs represent a rapidly expanding industrial and economical sector, beyond their clinical relevance. Recent reports estimate the global CGT market at USD 21-25 billion in 2024, with projections of USD 117-167 billion by 2034, depending on the source (Precedence Research, 2024; Market.US, 2024). This rapid growth reflects both the rising number of clinical approvals and the strong investment in advanced therapy medicinal

products (ATMPs). Together, the medical and economic impact of CGT explains why it is acknowledged as one of the most important translational frontiers of synthetic biology.

The development of constructs used for therapeutic rAAV (recombinant Adeno Associated Virus) vectors and IVT (In Vitro Transcribed) mRNAs (messenger Ribonucleic Acid) represents one of the most critical and resource-intensive steps in cell and gene therapy pipelines. The identification and generation of candidates that meet all functional, regulatory, and manufacturability criteria is inherently challenging: sequences must be fully accurate, structurally stable, and free of deleterious features, while simultaneously supporting efficient expression or packaging. In practice, this process is often time-consuming and cost expensive, as multiple iterations may be required before a construct is considered suitable for downstream use. For CGT companies, the timeline for obtaining fully sequence-verified rAAV plasmids or DNA (DeoxyriboNucleic Acid) templates for IVT mRNA can extend from several months to over a year, since repeated cloning and Sanger sequencing rounds are typically necessary to troubleshoot difficult motifs such as ITRs (Inverted Terminal Reapeats), poly(A) (polyAdenosine) tracts, and GC (Guanine-Cytosine)-rich regions. These inefficiencies directly delay preclinical testing and significantly increase development costs. It is within this context that our Biofoundry at Officinae Bio (OB) decided to focus on the CGT field. By specializing in the design, assembly, and validation of DNA and RNA constructs for rAAV vectors and in vitro transcribed (IVT) mRNA, OB aligns its technological capabilities with one of the most demanding and rapidly growing sectors of synthetic biology. This strategic positioning also reflects the industrial setting in which my doctoral project was carried out: while the Biofoundry is responsible for the complete design-to-construct pipeline, my contribution concentrated on the development and integration of sequencing-based Quality Control (QC) methods as the final step, ensuring that the generated candidates are accurate and reliable. Through the integration of next-generation sequencing technologies, Oxford Nanopore Technologies (ONT) for high-throughput screening and Illumina for base-level precision, our pipeline can reduce the time needed to obtain fully verified rAAV and IVT mRNA constructs to approximately one month, directly lowering costs and accelerating

the delivery of robust molecules to CGT companies, ultimately enabling faster development of novel therapies while safeguarding patient safety.

1.1.1 Recombinant adeno-associated virus (rAAV) vectors

Recombinant adeno-associated virus (rAAV) vectors are one of the most widely used delivery systems in gene therapy. Once the rAAV vectors are delivered into target cells, they release their single-stranded DNA genome, which is converted into double-stranded DNA and maintained as episomal concatemers in the nucleus. This supports the long-term transgene expression without integrating into the host genome (Mingozzi & High, 2011; Srivastava, 2016). This episomal persistence, combined with broad tissue tropism and a favorable safety profile, reinforces their widespread application in treating genetic diseases.

Since 2017, several rAAV-based products have been approved, including voretigene neparvovec (AAV2 - inherited retinal dystrophy), onasemnogene abeparvovec (AAV9 - spinal muscular atrophy), and more recently valoctocogene roxaparvovec (AAV5 - hemophilia A, approved in the EU in 2022 and by the FDA in 2023) (Byrne et al., 2025; Chen et al., 2024; Naso et al., 2017).

AAV is a small, non-enveloped parvovirus with a single-stranded DNA genome of approximately 4.7 kb. Its capsid is icosahedral and composed of three structural proteins, VP1–3, which are arranged around a central pore at the fivefold symmetry axis (Figure 1.1 a). The genome is flanked by inverted terminal repeats (ITRs), 145 nucleotides in length, which fold into stable T-shaped hairpins that are essential for viral replication and packaging. Within the genome, two major open reading frames encode the Rep proteins, which mediate replication, and the Cap proteins, which form the viral shell together with the assembly-activating protein (AAP) (*Figure 1.1* b) (Samulski & Muzyczka, 2014).

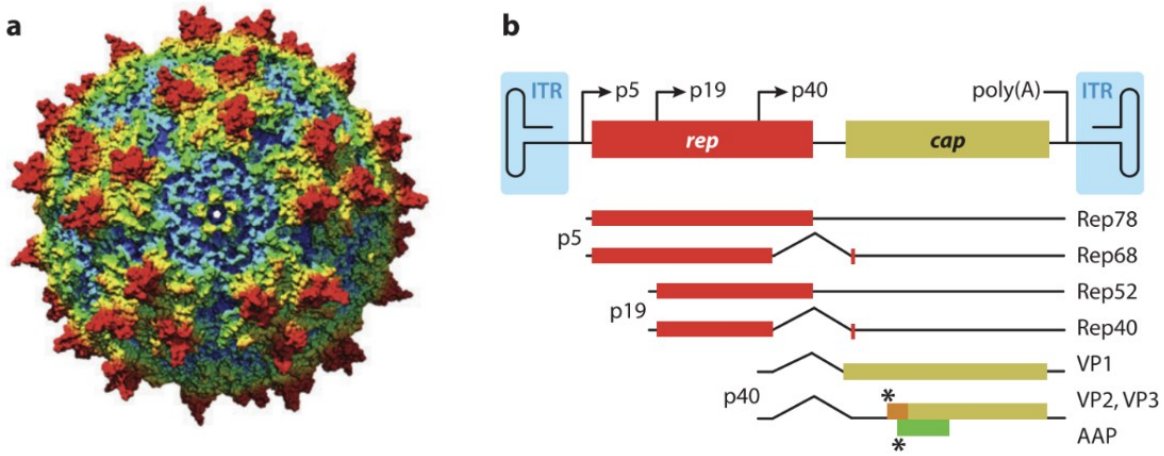

Figure 1.1: Structure and genome organization of AAV. (a) Surface model of the AAV2 capsid, showing the fivefold symmetry axis and central pore. (b) Genetic map of the 4.7 kb AAV genome. The rep open reading frame (red) encodes four Rep proteins (Rep78, Rep68, Rep52, Rep40) required for replication and packaging. The cap open reading frame (yellow) encodes three capsid proteins (VP1, VP2, VP3) in a ratio of 1:1:10, as well as the assembly-activating protein (AAP, green). The genome is flanked by the 145-nt inverted terminal repeats (ITRs, blue), which are essential for replication and encapsidation. *(Samulski & Muzyczka, 2014).*

In rAAV vectors, the original viral genome is re-engineered: the therapeutic cassette, typically consisting of a promoter, the gene of interest, and a polyadenylation signal, is inserted between the ITRs, while the viral rep and cap genes are removed. These genes, which encode the proteins required for genome replication (Rep78/68) and packaging (Rep52/40), as well as the capsid proteins VP1–3, are instead provided during vector production on a helper plasmid. In addition, efficient rAAV production requires a third adenoviral helper plasmid, which supplies a minimal set of adenoviral functions: E2A, a DNA-binding protein that stabilizes replication intermediates; E4 ORF6, which promotes viral mRNA export and supports genome replication; and VA RNAs, which prevent PKR-mediated shutdown of protein synthesis, ensuring efficient expression of viral proteins(Grieger & Samulski, 2005; J. H. Wang et al., 2024). This strategy ensures that only the therapeutic cassette, and not viral replication machinery, is packaged into the capsid (*Figure 1.2*).

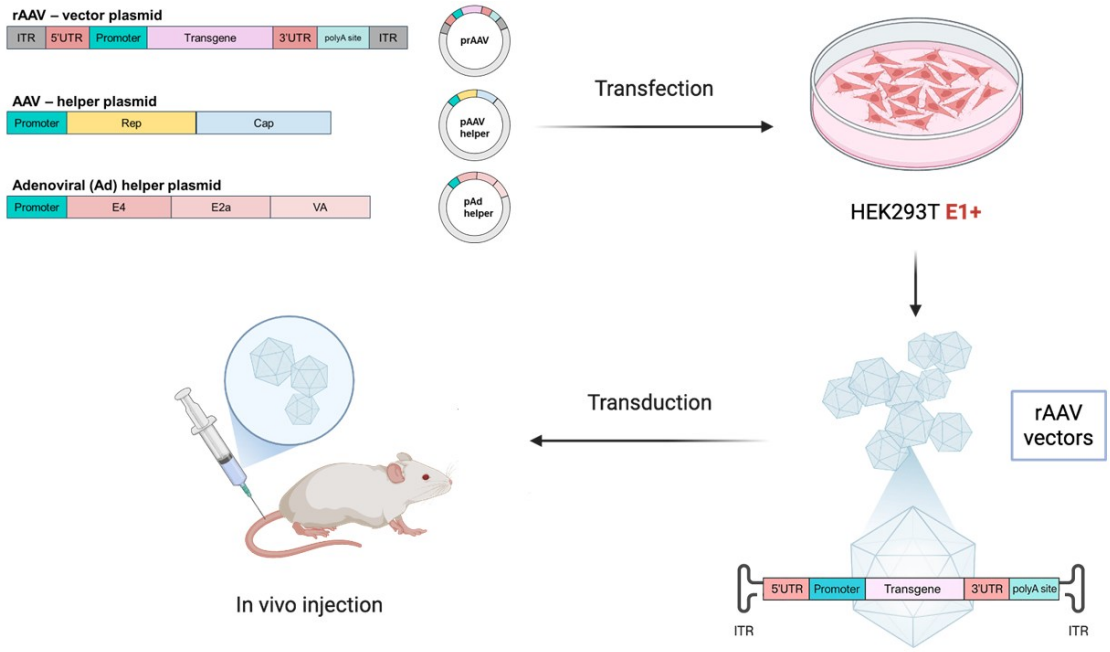

Figure 1.2: Production of recombinant AAV (rAAV) vectors by triple transfection. In this process, three plasmids are introduced into producer cells (HEK293, E1+): (i) the transfer plasmid, carrying the therapeutic cassette (ITRs flanking promoter–transgene–polyA signal); (ii) the AAV helper plasmid, which provides the rep genes (Rep78/68 for genome replication and Rep52/40 for packaging) together with the cap genes encoding capsid proteins VP1–3; and (iii) the adenoviral helper plasmid, which supplies the minimal adenoviral functions required for efficient production, including E2A (stabilizes replication intermediates), E4 ORF6 (facilitates viral mRNA export and genome replication), and VA RNAs (counteract PKR-mediated translational shutoff to maintain protein synthesis). Inside the producer cells, these elements act in concert to assemble rAAV particles, which are then purified for in vivo delivery of the therapeutic gene.

The limited packaging capacity is an important constraint for rAAV vectors. When genomes are larger than 4.7 kb, they are packaged inefficiently and often result in truncated or heterogeneous products that can compromise therapeutic potency (Murlidharan et al., 2014). In therapeutic designs, this limitation forces the use of compact promoters, optimized coding sequences and minimal regulatory elements.

From a production perspective, rAAV manufacturing relies on the simultaneous transfection of multiple plasmids, each contributing an important role in the vector assembly. This multicomponent system introduces several practical challenges: the plasmids are large, with helper plasmids often approaching 10 kb, and their sequences contain complex motives (ITRs and GC-rich regions) that are prone to recombination or instability during bacterial amplification (Grieger & Samulski, 2005; Naso et al., 2017;

Samulski & Muzyczka, 2014). Even minor errors or rearrangements in these constructs can compromise vector yield and quality, underscoring the need for stringent sequence verification. Traditional sequencing approaches such as Sanger sequencing are not well suited for these contexts, as their short reads fail to resolve repetitive elements and long secondary structures. This limitation has motivated the exploration of next-generation sequencing methods, such as Nanopore sequencing, which are capable of spanning entire plasmids and thereby validating both the therapeutic cassette and the larger helper plasmids used in rAAV production.

It has been demonstrated that structural defects, such as truncated, rearranged, or defective vector genomes, especially within ITR regions, reduce both the production and the therapeutic efficacy, compromising the therapeutic outcomes (Ersing et al., 2023; Kontogiannis et al., 2024; J. Zhang et al., 2024). This evidence emphasizes the importance of rigorous quality controls at the sequence level. Sanger sequencing is still the preferred method for high accuracy verification of small standard regions. However, its short read length and inability to resolve repetitive or highly structured regions make it unsuitable for comprehensive validation of rAAV constructs. To overcome these limitations, scientists are increasingly adopting next-generation sequencing, which enable full-length, high-throughput and more accurate assessment of vector genomes.

1.1.2 In vitro transcribed (IVT) mRNA

In vitro transcribed (IVT) mRNA has rapidly emerged as a versatile therapeutic modality, with applications spanning prophylactic vaccines (e.g. the COVID-19 vaccines BNT162b2) (Polack et al., 2020), protein replacement therapies (e.g. mRNA-3927 for propionic acidemia) (Baek et al., 2024), and cancer immunotherapies (e.g. mRNA-4157/V940 in melanoma) (Weber et al., 2025). While rAAV vectors deliver DNA that persists as episomes in the nucleus, IVT mRNA remains confined in the cytoplasm and is expressed transiently. This transient expression is often advantageous in clinical contexts where controlled, time-limited protein production is desired.

A therapeutic mRNA is designed to recapitulate the architecture of endogenous transcripts. It typically comprises a 5' cap, a 5' untranslated region (UTR), an open reading

frame (ORF) encoding the protein of interest, a 3' UTR, and a poly(A) tail (*Figure 1.3*). Each of these elements determines the therapeutic efficacy of the molecule thanks to their contribution to transcription stability, subcellular localisation and translational efficiency.

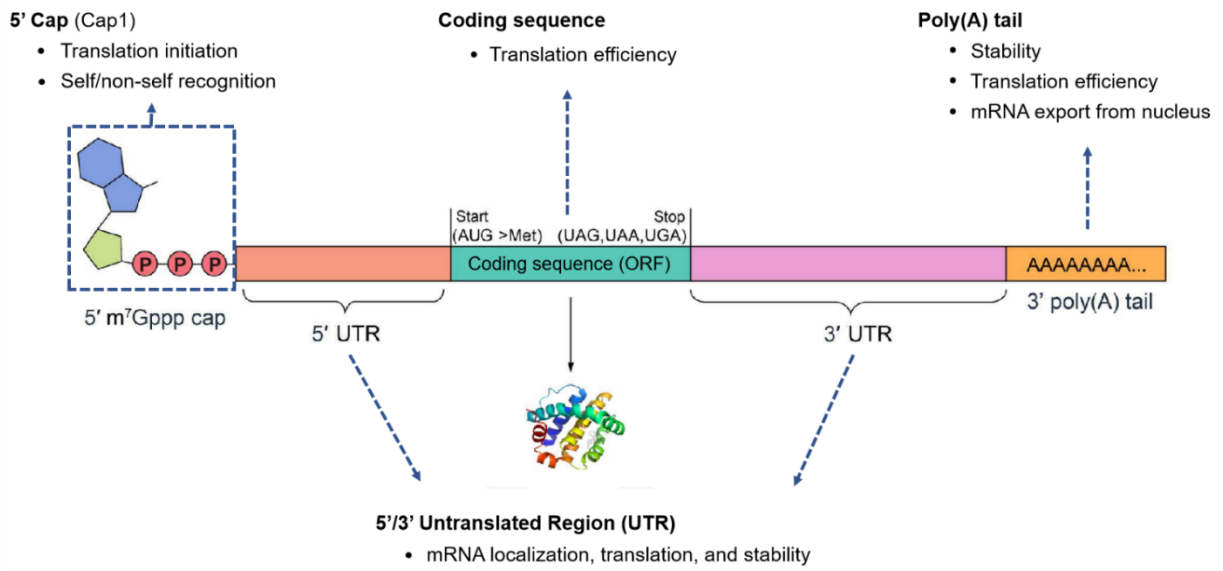

Figure 1.3: Schematic structure of therapeutic IVT mRNA. The IVT mRNA molecule contains a 5' Cap, which enhances stability and promotes ribosome recruitment; a 5' untranslated region (UTR) that contributes to translational regulation; an open reading frame (ORF) encoding the protein of interest; a 3' UTR that influences transcript stability and localization; and a poly(A) tail that protects against degradation and supports translation. Figure adapted from VectorBuilder (https://en.vectorbuilder.com/products-services/service/IVT-RNA/mRNA.html?).

One of the most transformative improvements in IVT mRNA technology has been the use of modified nucleosides, such as pseudouridine (Ψ) or N1-methylpseudouridine. These modifications reduce the innate immune recognition and enhance the translational efficiency (Karikó et al., 2005; Sahin et al., 2014). In particular, the first mRNA-based vaccine that was approved against SARS-CoV-2 contained chemically optimized transcripts incorporating modified nucleosides (N1-methylpseudouridine). This was a milestone that validated mRNA as a clinically viable therapeutic (Hou et al., 2021; Morais et al., 2021).

When it comes to quality control, several aspects of IVT mRNAs present specific challenges. One of the most important is the poly(A) tail, since its length directly affects transcript stability and translation. In many mRNA therapeutics, a tail of about 100

nucleotides is considered optimal. For instance, BioNTech's COVID-19 vaccine includes a 120-adenine tail, which has been shown to improve both stability and translation efficiency compared to shorter tails (Fang et al., 2022; Granados-Riveron & Aquino-Jarquin, 2021). Homopolymeric tracts, however, are notoriously difficult to analyze with standard sequencing methods, which often introduce compression or length biases. Moreover, even though the linear DNA template used for IVT can be completely verified before transcription, errors or heterogeneity can still occur during transcription itself, potentially affecting the quality of the final RNA product (Daniel et al., 2022). Conventional QC methods, such as spectrophotometry and electrophoretic profiling (TapeStation), are useful for checking the RNA concentration, size distribution and integrity. However, they lack the resolution to verify the nucleotidic sequence, a parameter that is critical for ensuring the functionality and safety of therapeutic mRNAs.

These limitations highlight the need of sequencing approaches capable of directly assessing full-length mRNA molecules, including poly(A) tails and chemical modifications. In this context, Oxford Nanopore Technologies (ONT) could be a good candidate for direct RNA sequencing to ensure the integrity and reliability of IVT mRNAs in therapeutic applications.

1.2 Challenges in rAAV and mRNA QC

The clinical and commercial development of cell and gene therapies (CGTs) depends in particular on the accuracy and reproducibility of recombinant DNA constructs and IVT mRNAs. Quality control must therefore ensure that each therapeutic molecule is accurately produced, since even minor deviations can compromise safety, efficacy, and regulatory compliance. Despite their transformative potential, both recombinant adeno-associated virus (rAAV) vectors and IVT mRNAs present specific sequence-level challenges that are not adequately addressed by traditional sequencing methods, such as Sanger Sequencing. These difficulties often translate into repeated design-build-test cycles, delaying development timelines and causing significant manufacturing costs before a final construct suitable for testing or clinical application is obtained.

For rAAV vectors, the primary difficulty arises from the inverted terminal repeats (ITRs). These 145-bp sequences form complex, T-shaped secondary structures rich in GC content and long repeats. They are indispensable for genome replication and packaging, yet notoriously difficult to sequence. Conventional Sanger sequencing typically fails to traverse ITRs, as their palindromic nature allows the duplex to unwind and extrude into cruciform-like secondary structures, which stall the polymerase and lead to incomplete or ambiguous chromatograms (*Figure 1.4*) (Bowater et al., 2022; Mroske et al., 2012). This limitation complicates the verification of constructs used for vector production, where any mutation or truncation within ITRs can directly impair viral assembly and reduce therapeutic efficacy. In addition to ITRs, the high overall complexity and size of rAAV plasmids, with helper and packaging plasmids often reaching 10 kb and containing repetitive elements, further strain standard QC pipelines.

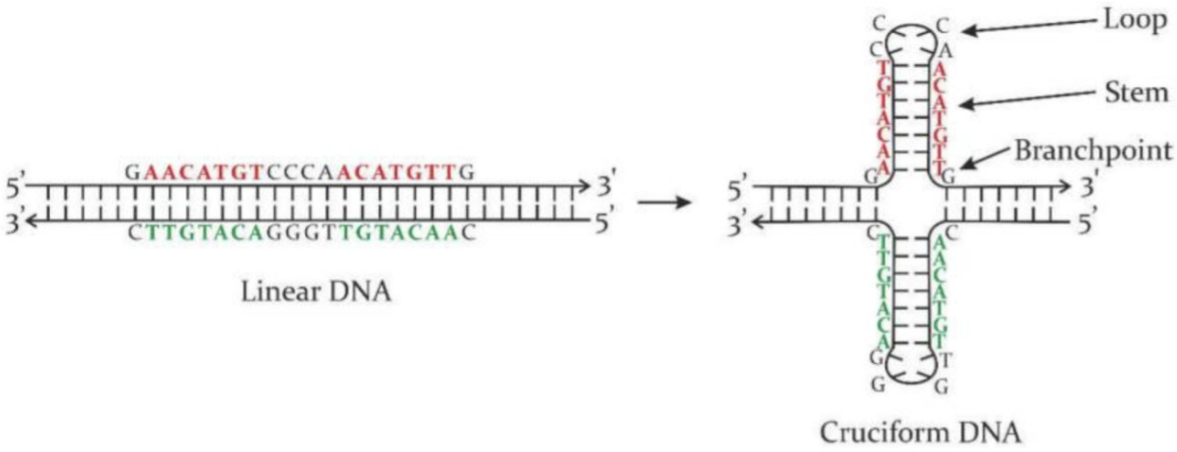

Figure 1.4: Representation of cruciform DNA formation from palindromic sequences. Upon unwinding, AAV inverted terminal repeats (ITRs) first adopt a linear intermediate conformation, before refolding into a cruciform-like secondary structure. These stable conformations block DNA polymerase progression during Sanger sequencing, resulting in incomplete or ambiguous chromatograms (Bowater et al., 2022; Mroske et al., 2012).

For IVT mRNAs, the poly(A) tail represents an equally critical yet challenging element. Polyadenylation directly influences transcript stability and translational efficiency, with tails of 100–120 adenosines conferring optimal performance in many therapeutic contexts (A. Sachs, 1990; Fang et al., 2022). Homopolymeric tracts, such as poly(A) tails, remain difficult to resolve with traditional sequencing methods: in Sanger sequencing, DNA polymerase frequently undergoes slippage at the end of long mononucleotide

stretches, generating a mixture of length variants that appear as overlapping or wave-like signals in the chromatogram, ultimately leading to ambiguous base calls (*Figure 1.5*) (Pereira et al., 2010). Moreover, while plasmid DNA templates can be verified by Sanger sequencing, the poly(A) tails of final IVT mRNA products cannot be directly confirmed with this approach, underscoring the need for sequencing strategies that can access native RNA molecules. Furthermore, the incorporation of modified nucleosides (e.g., pseudouridine) is essential for clinical performance, but complicates sequence verification because conventional basecallers are not trained to interpret modification-induced signal shifts.

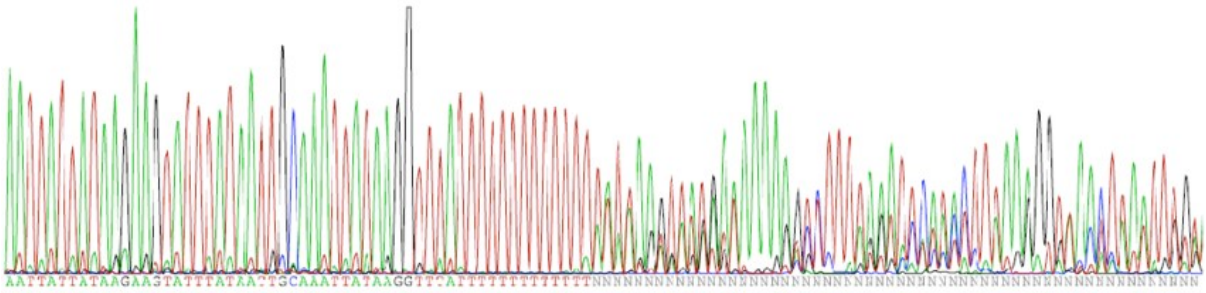

Figure 1.5: Example of sequence slippage occurring in a homopolymeric tract during Sanger sequencing. The DNA polymerase maintains processivity across most of the poly(A) stretch but frequently slips at its end, producing a heterogeneous mixture of products with variable lengths. This results in overlapping or wave-like signals in the chromatogram downstream of the homopolymer, which complicates accurate base calling ((Pereira et al., 2010).

Taken together, these aspects highlight an important limitation: current QC methods are not well indicated to capture the complexity of rAAV constructs and of IVT mRNAs at nucleotide resolution. Conventional analytical tools such as spectrophotometry, capillary electrophoresis, and Sanger sequencing provide valuable information on concentration, integrity, or short regions of sequence, but fail in structurally complex, repetitive, or homopolymeric regions. This creates blind spots that are particularly problematic in clinical development: undetected errors during construct synthesis can result in costly delays, failed batches or compromised safety in patients.

These challenges collectively motivate the adoption of next-generation sequencing approaches for CGT QC. By enabling comprehensive, high-throughput, and sequence-level analysis, platforms such as Oxford Nanopore Technologies and Illumina provide

opportunities to overcome the intrinsic limitations of traditional QC workflows, offering a path toward more reliable and scalable production of advanced therapeutics.

1.3 Sequencing Technologies for QC in CGT

The rigorous quality control of recombinant adeno-associated virus (rAAV) vectors and IVT mRNA is central to the development of safe and effective cell and gene therapies. Sequencing-based verification is fundamental to detect synthesis errors, such as mutations or structural anomalies that could compromise the therapeutic efficacy or safety. Over the past decades, sequencing technologies have evolved from traditional methods to next-generation platforms, each offering distinct advantages and limitations for QC pipelines.

1.3.1 Sanger Sequencing

Sanger sequencing, developed by Frederick Sanger and colleagues in 1977, marked a revolution in molecular biology by enabling the first complete sequencing of DNA molecules (Sanger et al., 1977). The method is based on selective incorporation of chain-terminating dideoxynucleotides (ddNTPs) during DNA synthesis, producing DNA fragments of varying lengths that can be separated by capillary electrophoresis and read with single-nucleotide resolution (Lloyd M. Smith et al., 1986; Maxam & Gilbert, 1977). Over time, the improvements in fluorescent labeling and automation have transformed the original labor-intensive method into a widely accessible and robust platform, establishing it as the “gold standard” for sequence verification in research and diagnostic laboratories. A schematic overview of the principle is illustrated in *Figure 1.6*.

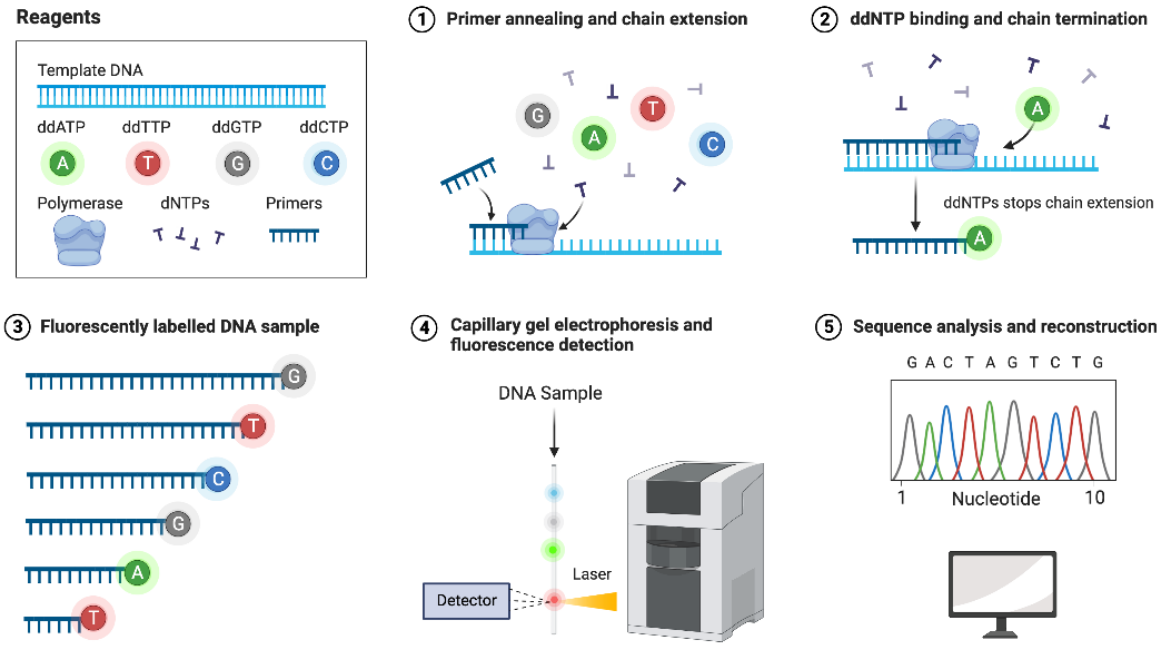

Figure 1.6: Principle of Sanger sequencing. A primer is extended on a single-stranded DNA template by DNA polymerase in the presence of both normal deoxynucleotides (dNTPs) and fluorescently labeled chain-terminating dideoxynucleotides (ddNTPs). Incorporation of a ddNTP halts elongation, producing a set of fragments that differ in length by one nucleotide. These fragments are separated by capillary electrophoresis, and the terminal fluorophores are detected by laser excitation. The output is a four-color chromatogram in which each peak corresponds to a nucleotide, enabling sequence reconstruction with single-base resolution. Figure adapted from BioRender template: https://www.biorender.com/template/sanger-sequencing

The strengths of Sanger sequencing lie in its accuracy and reliability. With reported error rates as low as 0.001% (Q40), it remains one of the most precise methods available for confirming short DNA sequences (Cheng & Xiao, 2022).It is particularly well-suited for applications such as mutation validation, small plasmid verification, or diagnostic assays where a limited number of sequences require precise base-level confirmation. For these reasons, Sanger sequencing has been a cornerstone in both academic and industrial biotechnology for decades.

However, the method presents intrinsic limitations that restrict its application in the context of cell and gene therapy (CGT). Individual reads typically do not exceed 800–1000 base pairs, meaning that longer DNA molecules, such as the helper and packaging plasmids used for recombinant adeno-associated virus (rAAV) production, often reaching 10 kb, require multiple overlapping reactions and custom primers to achieve full coverage

(Jan Kieleczawa, 2006).Moreover, the accuracy of the method decreases sharply in structurally complex regions, such as inverted terminal repeats (ITRs), long homopolymers like polyA tails, or GC-rich motifs, where secondary structures and repetitive content frequently lead to premature termination or unreadable chromatograms (Mroske et al., 2012). These characteristics make Sanger sequencing poorly scalable for high-throughput pipelines and unsuited for the verification of the large, complex DNA constructs and RNA templates required in modern therapeutic applications (Mohammadi & Bavi, 2022).

Sanger sequencing therefore continues to serve as a robust confirmatory method for short and simple constructs, where its high accuracy and interpretability remain unmatched. Yet its intrinsic limitations in read length, scalability, and performance on structurally complex regions raise important questions about its suitability for quality control in the context of advanced cell and gene therapy products. These limitations provide the rationale for investigating next-generation sequencing platforms in this thesis, with the aim of determining whether they can overcome Sanger's shortcomings and be effectively integrated into Biofoundry pipelines.

1.3.2 Oxford Nanopore Technologies (ONT)

Oxford Nanopore Technologies (ONT) introduced a paradigm shift in sequencing by enabling the direct, real-time analysis of nucleic acids as individual molecules translocate through a biological nanopore embedded in a membrane. As an electrical potential is applied, each nucleotide alters the ionic current in a characteristic manner, and these signal deviations are decoded into base sequences by dedicated algorithms (Mohammadi & Bavi, 2022; Y. Wang et al., 2021). Unlike sequencing-by-synthesis approaches, ONT does not rely on amplification or chemical termination, allowing the generation of reads that can span tens to hundreds of kilobases, even across repetitive or structurally complex regions. This single-molecule, long-read capability makes ONT uniquely suited for applications in synthetic biology and cell and gene therapy, where complete verification of large plasmids and direct RNA analysis are essential. An overview of the principle is illustrated in *Figure 1.7*.

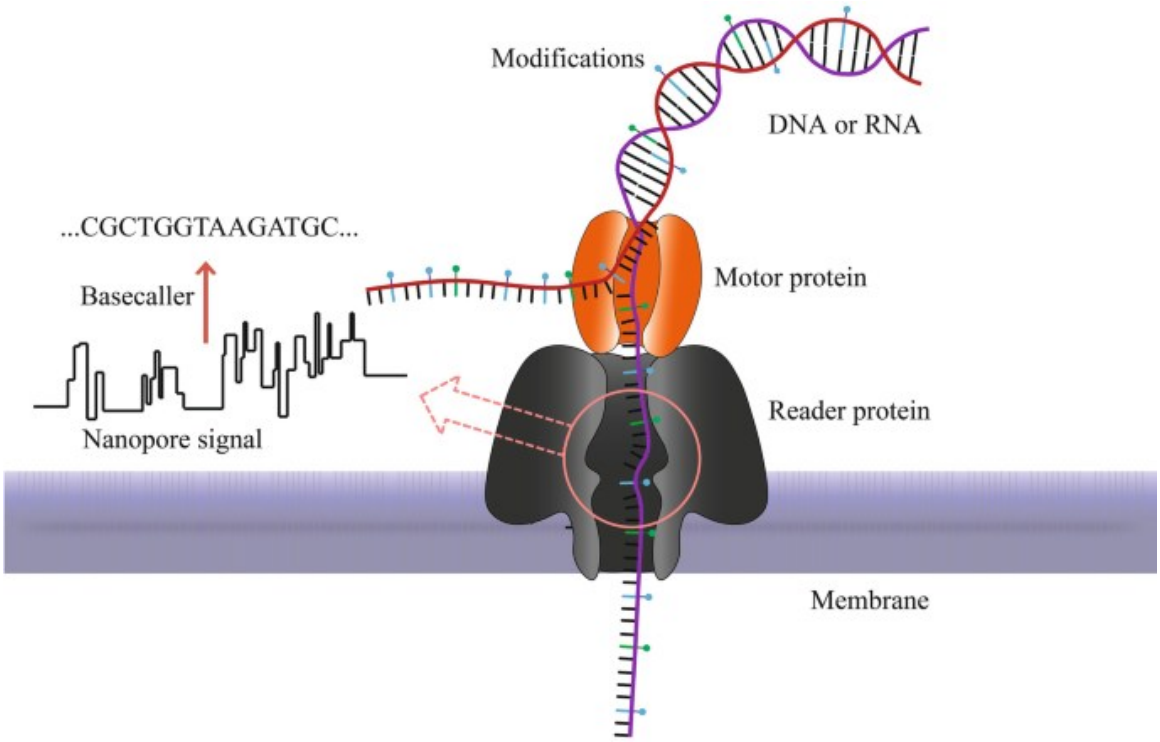

Figure 1.7: Principle of Oxford Nanopore Technologies (ONT) sequencing. A single-stranded DNA or RNA molecule is driven through a biological nanopore embedded in a membrane by a motor protein. As nucleotides pass through the pore, they produce characteristic disruptions in the ionic current. These current modulations are measured in real time and decoded into nucleotide sequences by basecalling algorithms. The method directly analyzes long native molecules without amplification, enabling the sequencing of both DNA and RNA. (Figure from He, M. et al., 2021)

The main strength of ONT lies in its capacity to produce long reads, often exceeding hundreds of kilobases, making it uniquely capable of spanning entire plasmids or transcripts in a single read (Jain et al., 2016). This feature is particularly advantageous in the context of cell and gene therapy (CGT), where recombinant adeno-associated virus (rAAV) vectors are built from plasmids of up to 10 kb containing repetitive elements and inverted terminal repeats (ITRs) that are systematically inaccessible to Sanger sequencing (Mroske et al., 2012). ONT can also sequence RNA molecules directly, enabling verification of IVT mRNAs without the intermediate step of reverse transcription (Lee et al., 2021). Furthermore, the platform supports barcoding strategies that allow multiplexing of up to 96 DNA samples per run, substantially increasing throughput and reducing costs to a few euros per sample when runs are performed at full capacity (Rang et al., 2018).

Despite these advantages, ONT is still limited by lower base accuracy compared to other sequencing technologies. Raw read quality typically ranges from Q10 to Q20,

corresponding to error rates of approximately 1–10% (Rang et al., 2018; T. Zhang et al., 2024). While consensus calling from high coverage can substantially improve accuracy, systematic errors such as insertions and deletions remain more frequent than in sequencing-by-synthesis platforms (Delahaye & Nicolas, 2021). In addition, although throughput is enhanced by sample barcoding for DNA applications, the absence of standardized multiplexing solutions for RNA currently restricts its scalability for transcript-level quality control (Heba H. Mostafa, 2024). Continuous improvements in pore chemistry, sequencing kits, and base-calling algorithms are gradually addressing these challenges, but at present they are holding back the use of ONT as a stand-alone solution for comprehensive quality control in cell and gene therapy.

Overall, ONT sequencing offers the ability to access and sequence structurally complex DNA regions and support rapid, scalable screening of large constructs, while also enabling direct analysis of RNA molecules. At the same time, its low accuracy and some unresolved technical constraints suggest that ONT alone may not be sufficient for comprehensive quality control in cell and gene therapy. These factors provide the rationale for the experimental work detailed in this thesis, which aims to evaluate the actual performance of ONT in these contexts and explore how it can be integrated with complementary sequencing technologies to ensure a robust and cost-effective quality control pipeline.

1.3.3 Illumina Sequencing

After the acquisition of Solexa, Illumina sequencing was commercialized in 2006 and become the most widely adopted NGS platform worldwide (Bentley et al., 2008). Its core technology is based on sequencing by synthesis (SBS): DNA fragments are ligated to adapters, immobilized on a flow cell, and clonally amplified into clusters by bridge amplification. During sequencing, fluorescently labeled reversible terminator nucleotides are incorporated one base at a time by a DNA polymerase. After each incorporation, the flow cell is imaged to detect the fluorescent signal, the terminator is chemically removed and the cycle is repeated (Bentley et al., 2008). This cyclic massively parallel process enables millions of short reads, typically 50-300 bp, that are generated with high accuracy. An overview of this workflow is illustrated in *Figure 1.8*.

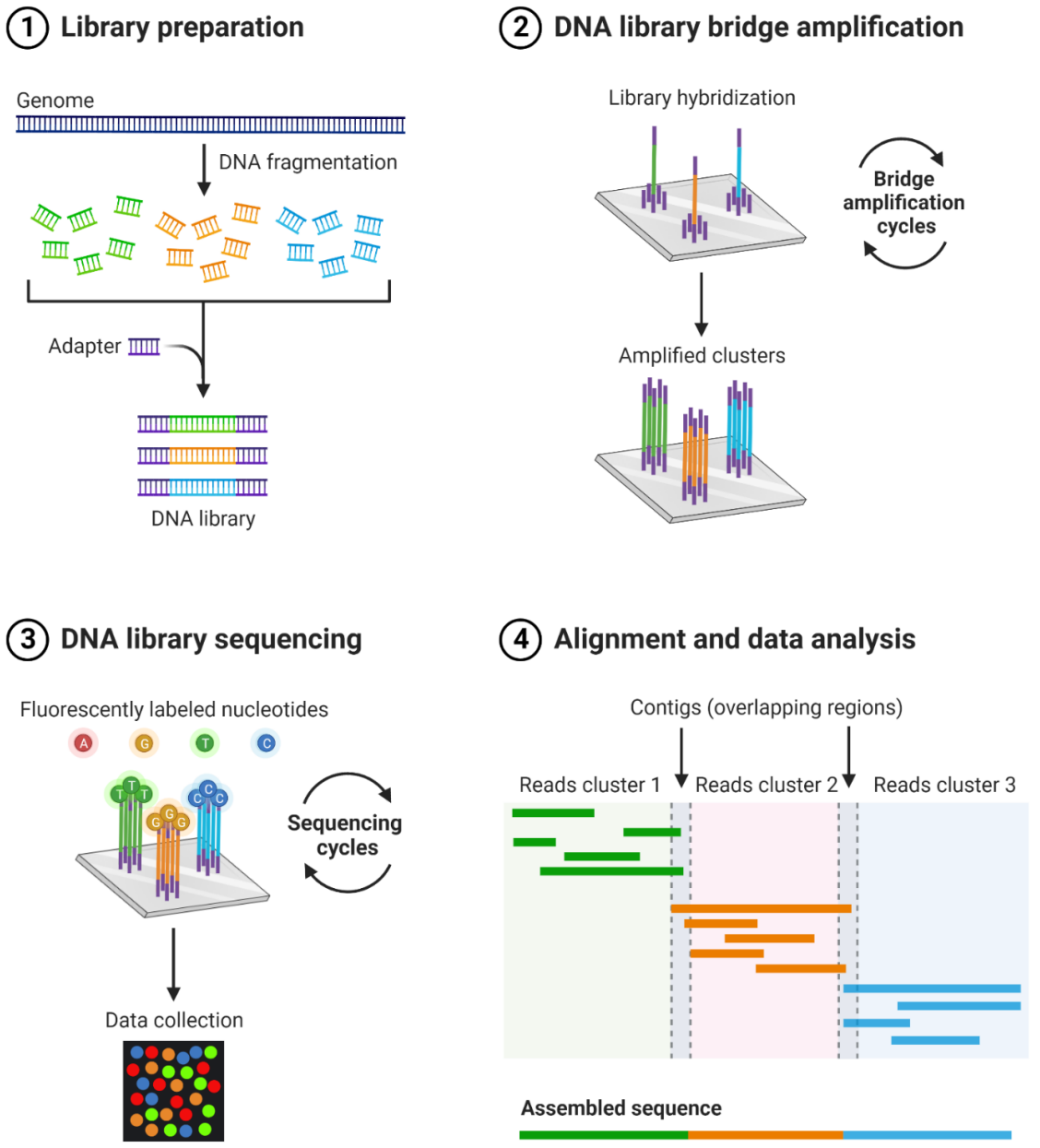

Figure 1.8: Principle of Illumina sequencing by synthesis. (1) DNA is fragmented and ligated to adapters to generate libraries. (2) Fragments are immobilized on a flow cell and clonally amplified through bridge amplification to form clusters. (3) Sequencing occurs by cyclic incorporation of fluorescently labeled reversible terminator nucleotides, imaged after each cycle to identify the added base. (4) Millions of short reads are aligned computationally, and overlapping regions are assembled to reconstruct the original sequence. Figure adapted from BioRender template: https://www.biorender.com/template/next-generation-sequencing-illumina.

One of the main strengths of Illumina sequencing is its base-level precision. With reported accuracies ≥Q30 (error rate ≤0.1%), it has become the benchmark for reliable variant detection and high-confidence sequence verification (Ross et al., 2013). Because of this, it has been considered for applications that require accurate reconstruction of

nucleotide stretches that are often challenging for other technologies, such as homopolymeric tracts or fragment termini. In addition, Illumina supports the use of up to 384 unique indexes, allowing the simultaneous sequencing of up to 384 samples in a single run, which makes it particularly suitable for high-throughput applications in Biofoundry workflows.

Despite these advantages, Illumina sequencing also presents some practical constraints. Sequencing runs are relatively long, often requiring up to 18 hours for paired-end 2x150 bp protocols, in contrast to Oxford Nanopore sequencing, which can generate usable data within minutes and be stopped once sufficient coverage is reached (Barton E. Slatko et al., 2018). Moreover, sequencing only a few samples makes the per-sample cost disproportionately high, limiting cost-efficiency for smaller experiments, while multiplexing allows the fixed cost of a flow cell to be distributed across many samples.

Taken together, Illumina sequencing emerges as a highly accurate short-read platform with potential to address some of the challenges posed by rAAV plasmids. Yet uncertainties remain about how well its advantages can be balanced against its longer runtime and higher costs in the context of routine Biofoundry applications. These questions form part of the investigations in this thesis, where Illumina is evaluated alongside Oxford Nanopore and Sanger sequencing to assess the extent to which complementary strategies may provide a robust solution for quality control in cell and gene therapy.

Although PacBio long-read sequencing platforms, particularly the Revio System, offer exceptional read accuracy through HiFi (High-Fidelity) sequencing chemistry, they were not selected for this work due to several practical limitations relative to the project's requirements. PacBio SMRT workflows typically demand microgram-scale quantities of high-molecular-weight DNA to generate high-quality long-read libraries (PacBio, 2023a), a factor that can limit throughput when working with multiple small plasmid samples. Moreover, the Revio System, while capable of producing up to 360 Gb of HiFi data per run, is associated with substantial costs: the instrument itself lists for approximately US $ 779,000, and each run costs roughly US $ 1,500–2,000 (PacBio, 2023b; 2024a). Even though the platform supports multiplexing up to 96 or more barcoded libraries per SMRT Cell, higher multiplexing inevitably decreases per-sample coverage, since the sequencing

runtime is fixed and cannot be extended (PacBio, 2023c). Additionally, the turnaround time represents a major operational constraint: PacBio library preparation typically requires around 6 hours, and each sequencing run occupies 24 hours of continuous instrument time, meaning that even a single batch may take 2 to 3 business days from DNA to data (PacBio, 2023d; University of Minnesota Genomics Core, 2024). Although PacBio's HiFi reads achieve accuracies up to Q50 (≥ 99.999 %), the combination of high DNA input, elevated operational costs, fixed sequencing duration, and long turnaround renders the system less suitable for the rapid, high-throughput plasmid sequencing workflows prioritized in this thesis. Instead, Oxford Nanopore Technologies (ONT) and Illumina platforms were selected as complementary solutions, combining scalability, lower per-sample cost, and faster processing times while maintaining adequate accuracy for plasmid verification and quality-control applications.

1.3.4 Cross-platform complementarity

Taken together, the three sequencing technologies discussed above illustrate the trade-offs inherent in applying sequencing to quality control of recombinant DNA and RNA molecules. Sanger sequencing remains the gold standard for high-accuracy analysis of short constructs and target regions, but it cannot resolve structurally complex plasmids. In fact, ITRs and other repetitive or GC-rich elements form stable secondary structures that make these regions inaccessible to chain termination chemistry, limiting the applicability of Sanger sequencing in construct synthesis pipelines for cell and gene therapy.

On the other hand, Oxford Nanopore sequencing can generate reads of virtually unlimited length, allowing entire plasmids to be spanned in a single molecule and making it possible to sequence regions that are not accessible to Sanger. In addition, the platform allows barcoding strategies that enable multiplexing of up to 96 DNA samples per run, reducing significantly per-sample cost when flow cells are used at full capacity. Its ability to directly analyze RNA without reverse transcription further increases its potential usage for quality control of therapeutic molecules. However, because of its lower per-base accuracy, it is not yet reliable enough to consider it as a stand-alone verification method.

By contrast, Illumina sequencing produces short but highly accurate reads, with typical error rates below 0.1%. While individual reads cover only a few hundred bases,

computational assembly makes possible to reconstruct full-length plasmids, so that there is no inherent limitation on construct size. As Nanopore, Illumina also supports multiplexing, allowing up to 384 indexed samples to be sequenced in parallel on a single flow cell. These features make it a powerful technology for sequence verification on large-scale. Although, relatively long run times and higher per-sample costs, when flow cells are not fully used, remain important drawbacks.

These characteristics highlight that there is not a single technology that can fully address all the needs of the quality control in cell and gene therapy. Instead, their complementarity suggests an integrated strategy in which each platform contributes with distinct strengths: Sanger as a confirmatory method for short, simple constructs; Oxford Nanopore for long-read coverage of entire plasmids, rapid screening, and cost-efficient multiplexing; and Illumina for highly accurate verification of assembled sequences across hundreds of barcoded/indexed samples. Establishing the feasibility and added value of such a combined approach provides the rationale for the investigations presented in this thesis, which aim to develop a robust and scalable sequencing pipeline for the quality control of rAAV plasmids and IVT mRNAs.

1.4 Research Aims and Objectives

The rapid development of cell and gene therapy (CGT) requires nucleic acid constructs that are not only accurately designed but also reliably verified before they are used for preclinical or clinical applications. However, the time and cost needed to obtain fully sequence-verified DNA plasmids and IVT mRNAs often represent a major bottleneck in the discovery pipeline. Conventional approaches, such as Sanger sequencing, remain accurate for short regions but cannot face the structural complexity and scale of the constructs typically used in CGT. Establishing robust, high-throughput, and cost-efficient sequencing strategies is therefore essential to accelerate the production of verified molecules, reducing delays and expenses for CGT companies and enabling faster translation of candidate therapies from design to in vivo testing and ultimately to the clinic.

The central aim of this project is to establish a next-generation sequencing based quality control framework to support the discovery and development of therapeutics, built on

recombinant adeno-associated virus (rAAV) vectors and IVT mRNAs. To achieve this, the project sets out the following objectives:

1. Evaluate Oxford Nanopore Technologies (ONT) sequencing as a rapid, high-throughput platform for the screening of large DNA constructs, with a focus on structurally complex motifs such as ITRs and GC-rich regions.
2. Assess the use of Illumina sequencing for DNA constructs as a complementary approach, offering high-accuracy for confirmatory sequence verification and allowing reliable detection of homopolymeric regions, such as poly(A) tails.
3. Extend the sequencing-based QC to RNA molecules, by applying direct RNA sequencing with ONT and testing its capacity to verify full-length IVT mRNAs.
4. Integrate the strengths of ONT and Illumina into a dual-platform pipeline, demonstrating the possibility to be routinely used in Biofoundry settings, where cost-efficiency, throughput and accuracy need to be carefully balanced.

This thesis thus establishes one of the first integrated NGS pipelines specifically designed to support the needs of the CGT industry. By systematically benchmarking Oxford Nanopore Technologies (ONT) and Illumina across challenging sequence contexts, including ITRs, GC-rich regions, poly(A) tails, and modified transcripts, the project defines their complementary roles in Biofoundry workflows. ONT enables rapid, high-throughput full-construct screening, while Illumina provides the base-level precision required for confirmatory analysis. Extending QC to the final IVT mRNA molecules, including pseudouridine-modified transcripts, this work demonstrates a scalable and cost-effective sequencing strategy aimed at reducing the time and cost of obtaining verified DNA and RNA molecules. Ultimately, the goal of this thesis is to accelerate the development of new rAAV- and mRNA-based therapeutics, bridging academic innovation with industrial translation in cell and gene therapy.

2 Materials and Methods

2.1 Lambda Control Experiment (SQK-LSK109)

Lambda DNA Library Preparation for Nanopore Sequencing

Lambda DNA (LMD, Oxford Nanopore Technologies) was used as the input material for library preparation, following the protocol provided with the Ligation Sequencing Kit (Oxford Nanopore Technologies, SQK-LSK109,) and the Control Expansion kit (Oxford Nanopore Technologies, EXP-CTL001,) with minor modifications. All procedures were performed in a nuclease-free environment using low-binding tubes and calibrated pipettes.

DNA Repair and End-Prep

A total of 20 µL of Lambda DNA (50 µg/mL) was combined with 1 µL of DNA Control Standard (DNA CS), 27 µL of nuclease-free water, 3.5 µL of NEBNext FFPE DNA Repair Buffer (New England Biolabs, Cat. No. M6630), 2 µL of NEBNext FFPE DNA Repair Mix (New England Biolabs, Cat. No. M6630S), 3.5 µL of Ultra II End-prep reaction buffer (New England Biolabs, Cat. No. E7647A), and 3 µL of Ultra II End-prep enzyme mix (New England Biolabs, Cat. No. E7646A), in a 0.2 mL PCR tube. The reaction (final volume: 60 µL) was mixed gently by flicking, briefly spun down, and incubated in a thermal cycler at 20 °C for 5 min followed by 65 °C for 5 min.

AMPure XP Bead Purification

The reaction was transferred to a 1.5 mL DNA LoBind tube and mixed with 60 µL of thoroughly resuspended AMPure XP beads (Beckman Coulter, Cat. No. A63881). The tube was incubated on a Hula mixer at room temperature for 5 min. After magnetic separation, the supernatant was discarded. The beads were washed twice with 200 µL of freshly prepared 70% ethanol (in nuclease-free water), then air-dried for 2 mins, avoiding over-drying.

Beads were resuspended in 61 µL of nuclease-free water and incubated for 2 min at room temperature. After magnetic separation, 61 µL of eluate was collected and quantified using Nanodrop (ThermoFisher Scientific).

Adapter Ligation

The eluted DNA was mixed with 25 µL of Ligation Buffer (LNB), 10 µL of NEBNext Quick T4 DNA Ligase (New England Biolabs, Cat.No. M2200S), and 5 µL of Adapter Mix (AMX), for a total volume of 100 µL. The reaction was incubated at room temperature for 10 min.

Post-Ligation Clean-Up

To the ligation reaction, 40 µL of resuspended AMPure XP beads was added. After a 5 min incubation on a Hula mixer, the sample was placed on a magnetic rack. The beads were washed twice with 250 µL of Long Fragment Buffer (LFB). The beads were briefly air-dried (2 mins), then resuspended in 15 µL of Elution Buffer (EB), incubated at room temperature for 10 min, and the eluate was recovered.

Flow Cell Priming and Library Loading

Before library loading, the active pores number of the Flow Cell (R9.4.1 FLO-MIN106, Oxford Nanopore Technologies) was evaluated using MinKNOW software. Only flow cells with at least 800 active pores were included in the experiments. Once verified the usability of the flow cell, this was primed, through the Priming Port, with 800 µL of the priming mix previously made of 1,170 µL of Flush Buffer (FB) supplemented with 30 µL of Flush Tether (FLT). During this step, care was taken to avoid introducing air bubbles during loading. While waiting 5 minutes of incubation at room temperature to allow equilibration, the sequencing library was prepared by mixing 12 µL of the eluted DNA with 37.5 µL of Sequencing Buffer (SQB) and 25.5 µL of thoroughly mixed Loading Beads (LB). At the end of the 5 minutes incubation, an additional 200 µL of the same priming mix was loaded via the priming port, with the Spot On open. Then, a total of 75 µL of the final library mixture was loaded dropwise into the SpotON sample port, ensuring that each drop was fully absorbed before adding the next.

Sequencing Data Collection and Analysis

Sequencing was initiated on a MinION Mk1B device using MinKNOW software (Oxford Nanopore Technologies). Raw signal data (FAST5 format) were collected continuously and converted into real-time basecalled reads (FASTQ format) using the integrated MinKNOW basecaller. Basecalling was performed with the Dorado neural-network engine, which is the default model integrated into MinKNOW for R9.4.1 and R10.4.1 flow

cells. In this configuration, MinKNOW automatically selected the appropriate basecalling model (typically from the dna_r9.4.1_e8 or dna_r10.4.1_e8.2_400 bps families) according to the flow cell chemistry and sequencing kit in use.

Following the completion of Nanopore sequencing, the generated FASTQ files were imported directly into Geneious Prime for alignment and assembly. The reads were mapped to the corresponding reference sequence using the Geneious mapper “Map to Reference” with Medium Sensitivity/Fast settings, corresponding to a mismatch tolerance of approximately 3–5 SNPs per 100 nt, reflecting the typical raw read error rate (95–97% accuracy) reported for Oxford Nanopore sequencing (Delahaye & Nicolas, 2021). The alignment process was configured to iterate up to five times to improve local mapping accuracy, and sequences were not trimmed prior to mapping. The coverage statistics were generated automatically within Geneious, and consensus sequences were assembled from the mapped reads using majority-rule calling.

2.2 Flow Cell Wash Protocol

To enable the reuse of MinION flow cells between sequencing runs, the Flow Cell Wash Kit (Oxford Nanopore Technologies, EXP-WSH004) was used following the manufacturer's instructions. After pausing the sequencing run in MinKNOW, a wash mix consisting of 2 µL of Wash Mix (WMX, containing DNase I) and 398 µL of Wash Diluent (DIL) was prepared, gently mixed by pipetting, and kept on ice. This mix (400 µL) was loaded into the flow cell via the priming port, taking care to avoid introducing air bubbles. The priming port was then closed, and the flow cell was incubated at room temperature for 60 minutes. Following the wash, the waste channel was cleared using a P1000 pipette. The flow cell was then either reloaded with a new library or stored at 4 °C in 500 µL of Storage Buffer (S) provided in the kit.

2.3 Nanopore sequencing using Rapid Barcoding Kit 96 (SQK-RBK110.96)

Genomic DNA Extraction and Quantification

Plasmid DNA was extracted from Escherichia coli NEB Stable competent cells (New England Biolabs, Cat. No. C3040H) previously transformed with recombinant plasmids. Following overnight culture in LB medium supplemented with the appropriate antibiotic, cells were harvested by centrifugation and subjected to plasmid purification using a commercial miniprep kit, E.Z.N.A.® Plasmid DNA Mini Kit (Omega Bio-Tek, Cat. No. D6942) following the manufacturer's protocol.

The plasmid DNA was eluted in nuclease-free water and quantified using NanoDrop (ThermoFisher Scientific). With the same measurement, the DNA purity was assessed, and only samples exhibiting $OD_{260/280} \geq 1.8$ and $OD_{260/230} \geq 2.0$ were used for sequencing. Approximately 50 ng of plasmid DNA per sample was used as input for library preparation.

Rapid Barcoding and Transposase-Based Fragmentation

Library preparation was performed using the Rapid Barcoding Kit 96 (Oxford Nanopore Technologies , Cat. No. SQK-RBK110.96), which utilizes a transposase-mediated fragmentation strategy that simultaneously attaches barcodes to DNA molecules. Each plasmid DNA sample (9 µl, 50 ng) was mixed with 1 µl of the appropriate Rapid Barcode (RB), and the mixture was incubated at 30°C for 2 minutes, followed by heat inactivation at 80°C for 2 minute.

Purification of Barcoded DNA

After barcoding, the samples were pooled together in equimolar amounts (up to 96 samples, depending on the experiment). The pooled barcoded DNA underwent a purification step using.

AMPure XP beads (Beckman Coulter, Cat. No. A63881) were resuspended by vortexing and added to the pooled sample at a 1:1 volumetric ratio. The bead–DNA mixture was incubated for 5 minutes at room temperature on a gentle rotator (Hula mixer) to allow binding. The beads attached to the barcoded DNA were pelleted using a magnetic rack, and the supernatant was carefully removed. The pellet was washed twice with 1.5 mL of freshly-prepared 80% ethanol, ensuring not to disturb the beads. After the second wash,

residual ethanol was removed, and the beads were briefly air-dried on the magnet (2-5 minutes).

The dried pellet was then resuspended in 15 µL of Elution Buffer (EB) and incubated at room temperature for 10 minutes. Finally, the eluate (which contains the purified library) was transferred to a new LoBind tube.

Adapter Attachment

Following DNA purification, 11 µL of barcoded DNA was transferred into a new 1.5 mL DNA LoBind tube. To this, 1 µL of Rapid Adapter F (RAP F) was added. The mixture was gently mixed by flicking the tube and briefly centrifuged to collect the liquid at the bottom. Adapter attachment was carried out by incubating the mixture at room temperature for 5 minutes.

Flow Cell Priming and Library Loading

Before the library loading, the available pores of the Flow Cell (Oxford Nanopore Technologies, R9.4.1 FLO-MIN106) was checked using the MinKNOW software (Oxford Nanopore Technologies). Specifically, the number of available pores was assessed using the Flow Cell Check function to ensure sufficient sequencing capacity. Only flow cells with ≥800 active pores were used in the experiment.

While the adapter ligation reaction proceeded, the MinION flow cell was primed. The priming solution was prepared by mixing 1.17 mL of Flush Buffer (FB) with 30 µL of Flush Tether (FLT) and vortexing thoroughly. A total of 800 µL of the priming mix was loaded into the priming port of the flow cell, taking care to avoid introducing air bubbles. The flow cell was left to equilibrate for 5 minutes before proceeding.

In parallel, the final library loading mix was prepared by combining: 37.5 µL of Sequencing Buffer II (SBII), 25.5 µL of Loading Beads II (LBII), freshly mixed immediately prior to use and 12 µL of the barcoded DNA library. This resulted in a final volume of 75 µL, which was mixed gently by pipetting up and down.

Once the initial priming incubation was completed, both the SpotON sample port and the priming port were kept open, and an additional 200 µL of the priming mix was carefully loaded into the priming port.

Finally, the prepared library (75 µL) was loaded into the SpotON sample port of the flow cell in a dropwise manner. Care was taken to ensure that each drop was fully absorbed before the next was applied, minimizing the risk of introducing air bubbles into the flow cell.

Sequencing Data Collection and Analysis

After the loading of the library, the sequencing was initiated on a MinION Mk1B device using the ONT proprietary MinKNOW software. Raw signal data (FAST5 format) were collected continuously and converted into real-time basecalled reads (FASTQ format) by using the integrated MinKNOW basecaller.

Following the completion of Nanopore sequencing, the generated FASTQ files were imported directly into Geneious Prime for alignment and assembly. The reads were mapped to the corresponding reference sequence using the Geneious mapper "Map to Reference" with Medium Sensitivity/Fast settings, corresponding to a mismatch tolerance of approximately 3–5 SNPs per 100 nt, reflecting the typical raw read error rate (95–97% accuracy) reported for Oxford Nanopore sequencing (Delahaye & Nicolas, 2021). The alignment process was configured to iterate up to five times to improve local mapping accuracy, and sequences were not trimmed prior to mapping. The coverage statistics were generated automatically within Geneious, and consensus sequences were assembled from the mapped reads using majority-rule calling.

Procedural adjustments for different Rapid Barcoding Kit and flow cell combinations

Over the course of the study, sequencing libraries were prepared using multiple versions of the Oxford Nanopore Rapid Barcoding Kit, in combination with either R9.4.1 or R10.4.1 flow cells. Although the overall workflow remained the same, minor adjustments were necessary to be implemented for differences in kit formats (12/24/96-sample configurations) and chemistry version. These adjustments involved incubation times, reagent composition, and loading steps, and are summarised in

*Table **2.1***.

Step	R9.4.1 - SQK-RBK004	R9.4.1 – SQK-RBK110.96	R10.4.1 – SQK- RBK114.24	R10.4.1 – SQK-RBK114.96
Barcode capacity	12	96	24	96
DNA Input amount (per sample)	400 ng (7.5 µL)	50 ng (9 µL)	200 ng (9 µL)	200 ng (9 µL)
Barcodes volume	2.5 µL [1]RB01–12 per sample	1 µL [1]RB01–96 per sample	1.5 µL [1]RB per sample	1.5 µL [1]RB per sample
Tagmentation incubation	30 °C × 1 min → 80 °C × 1 min	30 °C × 2 min → 80 °C × 2 min	30 °C × 2 min → 80 °C × 2 min	30 °C × 2 min → 80 °C × 2 min
Elution volume after clean-up	10 µL [2]EB	15 µL [2]EB	15 µL [2]EB (≤24 barcodes)	60 µL [2]EB (96-barcode pool)
Adapter attachment input	10 µL barcoded DNA + 1 µL [3]RA, 5 min RT	11 µL barcoded DNA + 1 µL [4]RAP-F, 5 min RT	11 µL barcoded DNA + 1 µL Diluted RA (prepared by mixing 1.5 µL [3]RA with 3.5 µL [5]RAB), 5 min RT	11 µL barcoded DNA + 1 µL Diluted RA (prepared by mixing 1.5 µL [3]RA with 3.5 µL [5]RAB), 5 min RT
Priming mix preparation	1.17 mL [6]FB + 30 µL [7]FLT	1.17 mL [6]FB + 30 µL [7]FLT	1.17 mL [8]FCF + 30 µL [9]FCT + 5 µL [10]BSA (50 mg/mL)	1.17 mL [8]FCF + 30 µL [9]FCT + 5 µL [10]BSA (50 mg/mL)
Sequencing buffer	[11]SB	[11]SBII	[11]SB (Kit 14)	[11]SB (Kit 14)
Loading beads	[12]LB	[12]LBII	[12]LIB (Kit 14)	[12]LIB (Kit 14)

Table 2.1: Procedural adjustments among different Oxford Nanopore Rapid Barcoding Kit and flow cell combinations.Comparison of key parameters for library preparation using R9.4.1 and R10.4.1 flow cells with various Rapid Barcoding Kit formats (SQK-RBK004, SQK-RBK110.96, SQK-RBK114.24, SQK-RBK114.96).*[1]RB: Rapid Barcoding; [2]EB: Elution Buffer; RA: [3]Rapid Adapter; [4]RAP-F: Rapid Adapter-F; [5]Rapid Adapter Buffer; [6]FB: Flush Buffer; [7]FLT: Flush Tether; [8]FCF: Flow Cell Flush Buffer; [9]FCT: Flow Cell Tether; [10]BSA: Bovine Serum Albumin (ThermoFisher Scientific, Cat. No. AM2616); [11]SB/SBII: Sequencing Buffer (legacy, version II, version 14); [12]LB/LBII/LIB: Loading Beads (legacy, version II, version 14).*

2.4 Nanopore Sequencing using Direct RNA Sequencing (SQK-RNA004)

RNA Input and Quality Control

A total of 300 ng of poly(A)-tailed RNA was used per reaction, adjusted to 8 µl with nuclease-free water. RNA quality was assessed prior to library preparation, ensuring appropriate integrity, length, and purity, as poor quality or contaminated samples can impair the sequencing outcome. Quality control was conducted using a NanoDrop spectrophotometer (ThermoFisher Scientific) to assess RNA purity and concentration, while RNA integrity and fragment length were evaluated using a TapeStation system (Agilent).

Preparation of Reagents

All reagents, including the NEBNext Quick Ligation Reaction Buffer (New England Biolabs, Cat. No. B6058S), T4 DNA Ligase (New England Biolabs, Cat. No. M0202S), RT Adapter (RTA), RNA Control Strand (RCS), RNA Ligation Adapter (RLA), Wash Buffer (WSB), and RNA Elution Buffer (REB), were thawed, vortexed, and briefly spun down prior to use. T4 DNA Ligase was not vortexed, as per manufacturer recommendations. RCS was used for the Control Experiment that was performed to test the protocol; its concentration was considered based on the kit batch.

RT Adapter Ligation

The RNA sample (8 µl, corresponding to 300 ng of poly(A)-tailed RNA) was transferred into a 0.2 ml thin-walled PCR tube. To this, 3 µl of NEBNext® Quick Ligation Reaction Buffer (New England Biolabs, Cat. No. B6058S), 0.5 µl of nuclease-free water, 1 µl of Murine RNase Inhibitor (New England Biolabs, Cat. No. M0214S), 1 µl of RT Adapter (RTA), and 1.5 µl of T4 DNA Ligase (2M U/ml) (New England Biolabs, Cat. No. M0202S) were added, for a total volume of 15 µl. The reaction was mixed gently by pipetting, briefly centrifuged, and incubated for 10 minutes at room temperature to enable ligation of the RT adapter to the RNA.

Reverse Transcription

A reverse transcription master mix was prepared in a clean 1.5 ml DNA LoBind tube by combining 13 µl of nuclease-free water, 2 µl of 10 mM dNTPs (ThermoFisher Scientific, Cat. No. R0192), and 8 µl of 5X Induro® RT Reaction Buffer (New England Biolabs, Cat. No.

M0681S) resulting in a total of 23 µl. This mix was transferred to the 15 µl RT adapter-ligated RNA reaction, followed by the addition of 2 µl of Induro® Reverse Transcriptase (New England Biolabs, Cat. No. M0681S). The final 40 µl reaction was gently mixed by pipetting.

The tube was incubated in a thermal cycler at 60 °C for 30 minutes, followed by 70 °C for 10 minutes to inactivate the enzyme. The sample was then cooled to 4 °C. After brief centrifugation, the full volume of each reaction was transferred to a clean 1.5 ml DNA LoBind tube.

First Purification with RNAClean XP Beads

The RNAClean XP beads (Beckman Coulter, Cat. No. A63987) were resuspended by vortexing and 72 µl of beads were added to the reverse transcription reaction (bead-to-sample ratio of 1.8X). The mixtures were pipette-mixed and incubated on a Hula mixer for 5 minutes at room temperature. The sample was then placed on a magnetic rack for separation, and once the solution cleared, the supernatant was carefully removed.

The beads were washed twice with 150 µl of freshly prepared 70% ethanol in nuclease-free water. For each wash, ethanol was added, the tubes were rotated 180° to allow complete bead movement, and the supernatant was removed after the beads had fully pelleted. After the final wash, residual ethanol was eliminated, and beads were allowed to air-dry briefly..

The pellet was resuspended in 23 µl of nuclease-free water and incubated at room temperature for 5 minutes. The tubes were placed back on the magnetic rack, and once the eluate cleared, 23 µl of supernatant was recovered and transferred to a clean 1.5 ml DNA LoBind tube. At this stage, the RT-RNA sample could be stored at -80 °C for later use. This represented the only permissible stopping point in the protocol.

RNA Adapter Ligation

To the 23 µl of purified RT-RNA, 8 µl of NEBNext® Quick Ligation Reaction Buffer (New England Biolabs, Cat. No. B6058S), 6 µl of RNA Ligation Adapter (RLA), and 3 µl of T4 DNA Ligase (New England Biolabs, Cat. No. M0202S) were added, reaching a total volume of 40 µl. The mixture was pipette-mixed and incubated for 10 minutes at room temperature to ligate the sequencing adapters.

Second Purification with RNAClean XP Beads

The RNAClean XP beads (Beckman Coulter, Cat. No. A63987) were again vortexed to resuspend. For each ligation reaction, 16 µl of beads were added, and the mixture was pipette-mixed and incubated for 5 minutes on a Hula mixer at room temperature. The sample was centrifuged briefly and placed on the magnetic rack for 5 minutes. Once the solution was clear, the supernatant was removed.

The pellet was washed twice using 150 µl of Wash Buffer (WSB). The tubes were removed from the magnet, the beads were resuspended by flicking and then returned to the magnetic rack for pelleting. The supernatant was removed carefully after each wash. Residual Wash Buffer was eliminated after a final brief spin and magnetic separation.

The bead pellet was then resuspended in 13 µl of RNA Elution Buffer (REB) and incubated for 10 minutes at room temperature. The samples were placed back on the magnetic rack until the eluate became clear. The final 13 µl of RNA library was recovered and transferred into a clean DNA LoBind tube.

Quantification

1 µl of the of the final RNA library was quantified using the NanoDrop dsDNA assay (ThermoFisher Scientific). A minimum quantity of 30 ng was typically targeted. The remaining 12 µl of eluted RNA library was carried forward into the flow cell loading step to maximize sequencing output.

Flow Cell Priming

Prior to priming, pore availability on the MinION/GridION Flow Cell RNA (FLO-MIN004RA) was assessed using MinKNOW software to verify adequate sequencing capacity. Flow cells with <800 active pores were considered unsuitable for direct RNA sequencing and excluded from further use.

The priming solution was prepared by mixing 1.17 mL of Flush Buffer (FB) with 30 µL of Flush Tether (FLT) and vortexing thoroughly. A total of 800 µL of the priming mix was loaded into the priming port of the flow cell, taking care to avoid introducing air bubbles. The flow cell was left to equilibrate for 5 minutes before proceeding.

Library Preparation and Loading

During the incubation period, the library was prepared for loading by combining 37.5 µl of Sequencing Buffer (SB), 25.5 µl of Library Solution (LIS), and 12 µl of the final RNA library (from the adapter ligation step), yielding a final volume of 75 µl. The mixture was gently pipette-mixed just prior to loading.

Once the initial priming incubation was completed, both the SpotON sample port and the priming port were kept open, and an additional 200 µL of the priming mix was carefully loaded into the priming port.

The prepared 75 µl library was loaded dropwise through the SpotON port. Once the entire volume had been loaded, both the SpotON port and the priming port were closed. To protect the flow cell from ambient light during sequencing, the MinION Flow Cell Light Shield was installed. The shield was gently positioned around the SpotON port area.

Sequencing Data Collection and Analysis

Sequencing was initiated on a MinION Mk1B device using MinKNOW software (Oxford Nanopore Technologies). Raw signal data (FAST5 format) were collected continuously and converted into real-time basecalled reads (FASTQ format) using the integrated MinKNOW basecaller.

Following the completion of Nanopore sequencing, the generated FASTQ files were imported directly into Geneious Prime for alignment and assembly. The reads were mapped to the corresponding reference sequence using the Geneious mapper “Map to Reference” with Medium Sensitivity/Fast settings, corresponding to a mismatch tolerance of approximately 3–5 SNPs per 100 nt, reflecting the typical raw read error rate (95–97% accuracy) reported for Oxford Nanopore sequencing (Delahaye & Nicolas, 2021). The alignment process was configured to iterate up to five times to improve local mapping accuracy, and sequences were not trimmed prior to mapping. The coverage statistics were generated automatically within Geneious, and consensus sequences were assembled from the mapped reads using majority-rule calling.

2.5 Illumina Sequencing using Illumina DNA Prep kit

DNA Input and Sample Preparation

Twelve samples of synthetic DNA, consisting of both plasmid and linear double-stranded DNA (dsDNA), were processed using the Illumina DNA Prep protocol (Illumina, Cat. No. 20060060). The input amount for each sample was within the range of 100–500 ng. As the DNA concentration was within this range, no prior normalization of input DNA was required before initiating the library preparation workflow. All samples were diluted, when necessary, in nuclease-free water to a final input volume of 30 µL and transferred into individual wells of a skirted 96-well PCR plate for subsequent processing.

Tagmentation of synthetic dsDNA

Tagmentation was performed using the bead-linked transposomes (BLT), which simultaneously fragment the DNA and incorporate the adapter sequences. For each sample, 20 µL of tagmentation master mix, prepared by combining BLT and Tagmentation Buffer 1 (TB1) in equal volumes, was added to the DNA. Samples were mixed thoroughly by pipetting and incubated in a pre-programmed thermal cycler at 55 °C for 15 minutes, followed by a hold at 10°C.

Post-Tagmentation Cleanup

After the tagmentation step, residual transposase activity was stopped by adding 10 µL of Tagment Stop Buffer (TSB) to each sample and mixing thoroughly. Samples were incubated at 37 °C for 15 minutes, then placed on a magnetic stand to pellet the beads. The supernatant was removed, and the beads were washed twice with 100 µL of Tagment Wash Buffer (TWB) to remove excess reagents. For the final wash, TWB was left in the wells to prevent the beads from overdrying before the PCR step.

Amplification of Tagmented DNA

A limited-cycle PCR amplification was performed to add unique dual indexes (UDI) (i7 and i5 index adapters), as well as sequences required for cluster generation. For the 100–500 ng input range, five PCR cycles were applied. The PCR master mix was prepared by combining Enhanced PCR Mix (EPM) with nuclease-free water, and 40 µL of this mix was added directly to the beads. Pre-paired dual index adapters were then added to each sample. The samples were mixed, sealed, centrifuged briefly, and amplified using the thermal cycler program: 68 °C for 3 minutes, 98 °C for 3 minutes, five cycles of 98 °C for

45 seconds, 62 °C for 30 seconds, and 68 °C for 2 minutes, followed by a final 68 °C extension for 1 minute.

Library Cleanup and Pooling

The amplified libraries were purified using Illumina Purification Beads (IPB) in order to remove small fragments and excess reagents. For the first bead binding step, each 45 µL aliquot of PCR product was diluted with 40 µL nuclease-free water and combined with 45 µL IPB (corresponding to a 0.45X bead-to-sample ratio). After mixing by pipetting and incubation at room temperature for 5 minutes, the samples were placed on a magnetic rack and the supernatant was transferred to fresh 1.5 mL LoBind tubes. In the second bead binding step, 15 µL undiluted IPB was added to each tube, followed by the transfer of 125 µL supernatant from the first binding step (equivalent to a 0.12X ratio). After mixing and a 5-minute incubation, the beads were pelleted magnetically, washed twice with 200 µL freshly prepared 80 % ethanol, and air-dried for 5 minutes. Libraries were then eluted in 32 µL Resuspension Buffer (RSB), and 30 µL of the eluate was transferred to new 1.5 mL LoBind tubes for subsequent pooling.

Equal volumes (5 µL) from each of the 12 libraries were combined into a single pool. The pooled library was quantified using a NanoDrop (Thermo Fisher Scientific).

The pooled library was checked for fragment size distribution using the D1000 ScreenTape Assay on a TapeStation system (Agilent, Cat. No. 5067-5582 and 5067-5583), with an expected average fragment size of approximately 600 bp.

Library Denaturation, Dilution, and Loading

The pooled library was prepared for sequencing on the Illumina iSeq 100 platform. The quantified pool was diluted in Resuspension Buffer (RSB) to the recommended starting concentration of 2 nM. Denaturation was carried out by mixing the library with an equal volume of 0.2 M sodium hydroxide and incubating for 5 minutes at room temperature. The denatured library was then diluted with pre-chilled hybridization buffer (HT1) to a final loading concentration of 200 pM. A 20 µL aliquot of the prepared library was loaded into the designated reservoir of the iSeq 100 reagent cartridge. The cartridge and flow cell were then inserted into the iSeq 100 instrument, and sequencing was performed using a paired-end configuration of 2x150 bp.

Read Processing and Bioinformatic Analysis in Geneious

Following completion of the sequencing run, raw basecall files were automatically converted into demultiplexed FASTQ files using the onboard iSeq 100 analysis software. This integrated pipeline performs the following steps: basecalling, adapter removal, and quality trimming according to Illumina's default parameters, ensuring that only high-quality reads are exported. The resulting processed FASTQ files for each sample were then imported into Geneious Prime for quality inspection and downstream analysis. Quality assessment included visual inspection of per-base Phred quality scores (Q) and read-length distributions to confirm data integrity.

The obtained reads were mapped to the corresponding reference sequence using the Geneious mapper "Map to Reference" with Medium Sensitivity/Fast settings, corresponding to a mismatch tolerance of approximately 3–5 SNPs per 100 nt, reflecting the typical raw read error rate (95–97 % accuracy). The alignment process was configured to iterate up to five times to improve local mapping accuracy, and sequences were not trimmed prior to mapping. Coverage statistics were generated automatically within Geneious, and so was done for the consensus sequences which were automatically assembled by the software from the mapped reads using majority-rule calling.

3 Results and Discussion

The quality control of recombinant Adeno-Associated Virus (rAAV) vectors and IVT mRNA is a critical step in the development of safe and effective cell and gene therapies (CGT) (Naso et al., 2017). In rAAV-based products, the inverted terminal repeats (ITRs) are essential for vector replication and packaging but are notoriously difficult to sequence due to their high GC content, repetitive structure, and stable secondary conformations (Samulski & Muzyczka, 2014). Similarly, in IVT mRNA, the length and integrity of the polyA tail directly influence transcript stability and translational efficiency (A. Sachs, 1990; Passmore & Coller, 2022), making accurate sequencing of homopolymeric stretches essential for therapeutic efficacy. Traditional Sanger sequencing, while widely used, is often unsuitable for these challenging motifs and does not scale efficiently for high-throughput screening (Mohammadi & Bavi, 2022).

In our Biofoundry, the synthesis pipeline for double-stranded DNA (dsDNA) begins with the in silico design, then proceeds with the chemical synthesis of DNA fragments, followed by the cloning into plasmid backbones, colony screening and final construct quality check. The output may be a circular plasmid or a linear dsDNA fragment, depending on the intended application. Since undetected errors can compromise downstream processes such as viral packaging or mRNA transcription, both forms must be sequence-verified before proceeding further.

For IVT mRNA production, the coding sequence is first cloned into a plasmid backbone specifically engineered to contain a 120-nucleotide polyA tract, ensuring that this critical feature is encoded at the DNA level and faithfully transcribed into the RNA product. The plasmid is then linearized to generate the DNA template which will be used for in vitro transcription. IVT is carried out with ribonucleotide triphosphates, optionally incorporating modified nucleotides such as pseudouridine (ΨTP) to enhance stability and reduce immunogenicity. The resulting transcripts are capped, producing mature mRNAs intended for in vivo applications. Standard assays, including UV spectrophotometry, electrophoretic profiling, and RNA integrity analysis, provide information on purity and size distribution, but they cannot reveal deviations at the ribonucleotidic sequences.

An overview of the synthesis workflows implemented in our Biofoundry for both DNA constructs and IVT mRNA is presented in *Figure 3.1*. The scheme illustrates the upstream steps required for construct generation, from in silico design to plasmid DNA (pDNA) extraction, and the downstream conversion of sequence-verified templates into IVT mRNA. Two quality control checkpoints are highlighted: the 'Sequence verification' step (in green), which in the current pipeline relies on Sanger sequencing but is insufficient for the complex constructs needed in rAAV vectors and IVT mRNA templates. In this thesis, Nanopore and Illumina are introduced as complementary platforms to address this gap. The second quality check is done on the final IVT mRNA (in yellow), which currently depends on spectrophotometric and electrophoretic profiling. Here, we evaluate the potential of incorporating Nanopore sequencing to provide direct sequence-level verification of the final transcript.

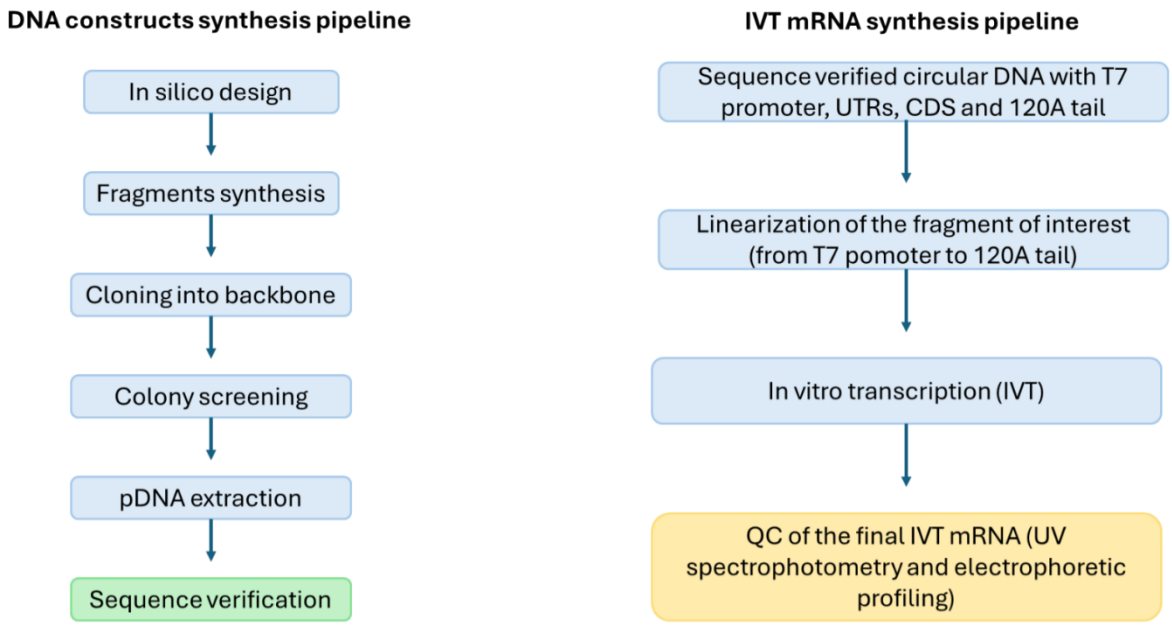

Figure 3.1: Schematic overview of the Biofoundry pipelines for DNA construct synthesis and IVT mRNA production. The left panel illustrates the DNA workflow, from in silico design to plasmid extraction and sequence verification. The right panel shows the IVT mRNA workflow, from sequence-verified plasmid DNA through linearization and transcription to final QC. Two checkpoints are highlighted: DNA sequence verification (green), currently performed with Sanger sequencing but insufficient for the complex constructs used in rAAV and IVT mRNA templates, where this thesis introduces Nanopore and Illumina; and final IVT mRNA QC (yellow), presently limited to spectrophotometric and electrophoretic profiling, where we investigate the use of Nanopore sequencing for direct sequence-level verification.

This PhD project focused on integrating next-generation sequencing (NGS) platforms, Oxford Nanopore Technologies (ONT) and Illumina, into these synthesis pipelines to overcome the limitations of Sanger sequencing. Their adoption enables earlier detection of errors, scalable screening across large numbers of samples, and reliable characterisation of challenging motifs such as ITRs and polyA tails. The following sections present the experimental results obtained over three years, addressing three main challenges in the quality control workflow: (i) high-throughput screening of complex DNA motifs, (ii) precise characterisation of homopolymeric regions, and (iii) sequence verification of full-length mRNA molecules.

3.1 Validation of Nanopore Sequencing: control experiment on λ phage

As an initial step toward establishing an NGS-based quality control (QC) pipeline, I validated Nanopore sequencing on a well-characterised control sample. This preliminary test served to gain hands-on familiarity with the MinION instrument and the associated MinKNOW software and to ensure that all the steps of the workflow could be performed reliably in our laboratory setting. The DNA of *Escherichia coli* bacteriophage λ was chosen for this purpose due to its stable, well-documented reference genome (F. Sanger et al, 1982), making it an ideal benchmark for method validation.

The experiment was performed using the Control Expansion Kit (EXP-CTL001) together with the Ligation Sequencing Kit (SQK-LSK109) and an R9.4.1 flow cell, following ONT's standard guidelines for control runs. Following to the kit protocol, the library preparation was carried out and loaded onto the MinION device for sequencing under standard control-run conditions. The run lasted nearly 18 hours and generated 371,300 reads, corresponding to 6.15 Gb of sequence data. Of this output, 4.84 Gb passed the default Q-score threshold (≥8), while 1.13 Gb were filtered as low quality. The read length distribution was optimal, with an N50 of 35 kb (meaning that half of the total bases generated are contained in reads of 35 kb or longer). This demonstrated the platform's ability to produce long reads from intact λ DNA molecules. Data were analysed using the EPI2ME Control Experiment workflow, which automatically aligns the reads to the λ reference genome and reports run-level quality metrics, including sequence identity and coverage distribution.

The control run completed without technical issues, confirming the integrity of the library preparation, sequencing, and basecalling steps. EPI2ME analysis reported a mean alignment identity of 98.1% to the λ reference genome, a value in line with expected performance for the R9.4.1 chemistry. This result confirmed that the instrument, flow cell, and analysis pipeline were functioning under optimal conditions. The graphical summary of the alignment results, generated by EPI2ME, is shown in *Figure 3.2*. While this identity value serves as a useful performance benchmark for the MinION platform, it should be noted that this metric was obtained only for the control dataset. In subsequent experiments with synthesised plasmids, a different analytical approach was required: because EPI2ME did not yet support the function of aligning each barcoded dataset to its specific reference, all downstream analyses were performed in Geneious to enable targeted assessment of construct-specific sequence integrity.

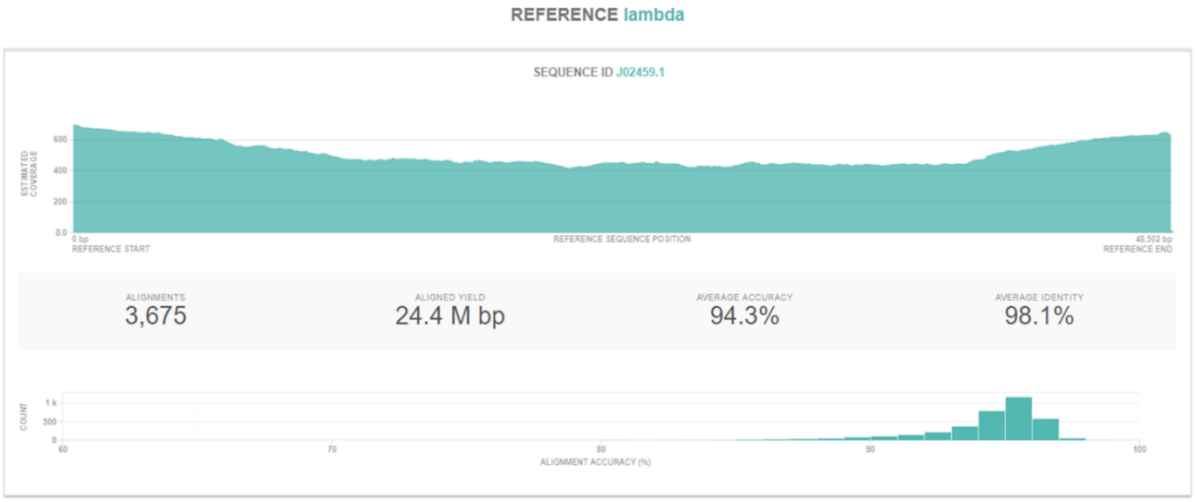

***Figure 3.2*: Validation of MinION sequencing using λ phage DNA.** The output from the EPI2ME Control Experiment workflow shows the read alignment of the sequenced reads to the reference genome (Escherichia coli bacteriophage λ DNA). The analysis reported a mean sequence identity of 98.1%, consistent with expected performance for R9.4.1

Establishing this λ DNA baseline was an essential de-risking step before attempting more challenging sequencing tasks. It provided the confidence to move forward with high-throughput barcoded screening of up to 96 plasmid samples per run and the analysis of technically difficult motifs, such as inverted terminal repeats (ITRs) of AAV genomes and long homopolymeric tracts, both of which are known to pose significant challenges to Sanger sequencing.

3.2 Nanopore Sequencing for dsDNA

3.2.1 Implementation and scalability of Nanopore sequencing

Nanopore sequencing was introduced into our Biofoundry with the goal of establishing a scalable and robust quality control pipeline for DNA constructs destined for recombinant AAV (rAAV) production. Unlike traditional methods such as Sanger sequencing, which is limited to short read lengths and typically focuses only on newly synthesized inserts, Nanopore sequencing has the capacity to span entire plasmids, thereby ensuring complete verification of both insert regions and vector backbones. This distinction is particularly important in cell and gene therapy applications, as undetected mutations in the backbone could compromise downstream processes such as viral packaging or lead to unexpected off-target effects. In the context of rAAV production, this requirement becomes especially relevant since the plasmids involved, including helper and packaging plasmids, typically reach sizes of up to 10 kb (Naso et al., 2017; Samulski & Muzyczka, 2014). Sequencing plasmids of this length is therefore essential to guarantee the integrity of the constructs used in viral production.

In order to meet different experimental needs, we tested several ONT library preparation kits. The Rapid Barcoding Kit 96 (SQK-RBK110.96 and SQK-RBK114.96) was adopted as the primary approach, enabling the simultaneous sequencing of up to 96 plasmids in a single run on MinION R9.4.1 and R10 flow cells, respectively. For smaller-scale experiments, we evaluated the Rapid Barcoding Kit 12 (SQK-RBK110.12) with R9.4.1 chemistry, and the Rapid Barcoding Kit 24 (SQK-RBK114.24) introduced with R10 chemistry, which provide intermediate multiplexing options. This range of kits allowed us to adapt sequencing runs to the number of samples, optimizing both cost-efficiency and turnaround time.

During the 3 years of this project, around 5000 DNA samples were sequenced across different experimental contexts, making this one of the largest systematic applications of ONT sequencing for construct QC in a Biofoundry environment. The majority of these samples consisted of plasmids, but the workflow also included linear double-stranded DNA (dsDNA) obtained either by PCR amplification or by restriction enzyme digestion. The ability to sequence both circular plasmids and linear fragments highlighted the versatility

of Nanopore sequencing in handling different DNA topologies, an important feature for synthetic biology applications where DNA is frequently manipulated in multiple formats.

In the first experimental phase, a cohort of 569 plasmids was analyzed, categorized into three size categories: small (2-5 kb, n = 201), medium (5-15 kb, n = 249), and large (15-40 kb, n = 119). Sequencing runs produced reads that spanned full plasmids, regardless of their size. That highlighted that plasmid length was not a limiting factor for ONT-based analysis. Similarly, linear dsDNA fragments of different sizes were sequenced with comparable efficiency. This confirmed that ONT can be applied to a wide range of DNA constructs without major adjustments to the protocol.

The scalability of ONT sequencing has markedly improved the throughput of our QC pipeline. By comparison, Sanger sequencing typically produces reads of only 1000 bp, requiring multiple reactions and custom primers to verify longer constructs. Consequently, Sanger costs increase proportionally with plasmid size. ONT sequencing, on the other hand, generates continuous reads that span entire plasmids in a single run. When combined with multiplexing of up to 96 plasmids in parallel, this reduces the turnaround time for quality checks from several days to just a few hours, representing a major advantage for high-throughput DNA synthesis workflows.

3.2.2 Case study: sequencing a 30 kb plasmid with ITRs and repetitive regions

One of the most demanding challenges encountered during this project was the verification of highly complex constructs. Among these, a particularly illustrative case was a 30 kb plasmid characterized by multiple inverted terminal repeats (ITRs), complex repetitive regions, and local regions of elevated GC content (>70%). These intrinsic features made the construct not only difficult to synthesize and propagate correctly but also extremely challenging to validate with traditional sequencing approaches. In molecular biology, the successful assembly and stable propagation of a plasmid of this size and complexity is in itself a rare outcome, which required screening a large number of colonies to identify a positive clone.

To enable this screening, we employed Oxford Nanopore sequencing using the Rapid Barcoding Kit 96 (SQK-RBK110.96) on a MinION Mk1B device. The sequencing run, performed on a FLO-MIN106 flow cell, lasted 8 hours and generated 493,480 reads, corresponding to approximately 1.02 Gb of total output, of which 749 Mb passed quality filtering. The dataset exhibited an N50 read length of approximately 2.8 kb and was basecalled with the Fast basecalling model (Guppy 5.0.11).

In this context, Nanopore sequencing proved to be unrivalled: thanks to the Rapid Barcoding Kit 96, it was possible to screen 96 colonies in a single run, even though the synthesised construct was exceptionally large (30 kb) and structurally complex. This combination of high colony number and plasmid length would have been virtually impossible to manage with traditional sequencing methods, such as Sanger Sequencing. Crucially, the entire run cost approximately € 300, corresponding to only € 3.15 per sample, making large-scale screening both technically and economically feasible. This represents a dramatic improvement over conventional approaches, where verifying such a large number of colonies for a construct of this size would have been prohibitively expensive and time-consuming.

On the same level of importance, Nanopore sequencing did not only identify the single positive clone out of the 96 colonies screened, but it also enabled full-length sequence validation, including regions systematically inaccessible to Sanger sequencing, as showed in the direct comparison in *Figure 3.3*. In more detail, in panel *a* the Sanger sequencing alignment of the ITR and Repeat 11 region shows a sharp decline in read quality, with incomplete coverage and ambiguous base calls that left large stretches unresolved despite repeated primer design attempts. In contrast, panel *b* shows the Nanopore sequencing alignment of the same region, obtained with 3000X coverage. Reads spanned the entire plasmid length, including the ITRs and repetitive elements: this produced a clean and reliable consensus sequence despite minor mismatches in individual reads.

This case study illustrates the dual strengths of Nanopore sequencing. First, its high-throughput capacity allows the parallel screening of many colonies, making it possible to isolate rare positive clones of highly challenging constructs. Second, its long-read

capability enables complete plasmid verification, resolving difficult regions such as ITRs, long repeats, and GC-rich domains that cannot be fully captured by Sanger sequencing. Together, these features make Nanopore sequencing particularly well suited for reducing risk in the production of large, complex DNA constructs within Biofoundry workflows.

Figure 3.3: Comparison of Sanger (a) and Nanopore (b) sequencing alignments of a 30 kb plasmid containing ITRs and repetitive regions. Panel a shows the Sanger alignment of the ITR and Repeat 11 region using both forward and reverse primers. Within the repetitive and GC-rich sequences, read quality drops drastically, leading to incomplete coverage, ambiguous base calls, and large unresolved stretches. Panel b shows the Nanopore alignment of the same region from a 30 kb plasmid screened across 96 colonies. In this case, reads spanned the full plasmid, with a 3000X coverage that enabled the complete consensus reconstruction, including ITRs and repetitive regions. While minor mismatches were present in individual reads, the consensus was clean and accurate.

3.2.3 Sequencing of linear dsDNA fragments

Beyond plasmids, Nanopore sequencing was also applied to linear double-stranded DNA (dsDNA) constructs, which represent an important class of molecules in synthetic biology and cell and gene therapy pipelines. Such fragments can be generated ex novo by chemical synthesis, or obtained through PCR amplification or restriction digestion, and

they serve multiple purposes within molecular workflows. On the one hand, they represent the starting bricks for the assembly of circular constructs via methods such as Gibson or Golden Gate cloning; on the other hand, they can act as templates for in vitro transcription (IVT) in the production of therapeutic mRNA. Ensuring the sequence integrity of these fragments is therefore critical, yet their structural features often pose significant challenges for conventional sequencing methods. By applying Nanopore sequencing to different types of linear dsDNA, we sought to evaluate its capacity to provide complete and accurate sequence information under a variety of contexts.

Nanopore sequencing was used to profile a diverse set of linear dsDNA constructs, allowing us to assess whether the platform could deliver complete and high-fidelity sequence information across fragments of varying complexity. The sequencing run was performed on a MinION Mk1B instrument using an FLO-MIN114 flow cell and the SQK-RBK114-96 kit. The run lasted for approximately 70 minutes, generating 189,884 total reads, of which 163,540 passed the quality filter. The dataset exhibited an estimated N50 read length of 5,954 bp and a modal Q score of approximately Q12, consistent with the performance expected from the Dorado fast basecalling model v4.3.0 at 400 bps. These metrics collectively confirmed that the run provided ample depth and read length to support high-confidence consensus calling across the tested fragments.

Three representative examples are presented in *Figure 3.4*. In panel a, it is showed the alignment of a standard PCR amplicon with no structural complexity. In this case, Nanopore sequencing readily produced reads spanning the entire fragment, yielding a clean and unambiguous consensus sequence. This demonstrates the robustness of the method for routine quality control of simple linear DNA fragments. The second example (panel b) illustrates a more structurally challenging fragment, a PCR product containing both direct and palindromic repeats. These features frequently impair Sanger sequencing, leading to premature termination or ambiguous chromatograms. In contrast, Nanopore sequencing provided continuous coverage across the repetitive motifs, and the high read depth ensured that occasional mismatches in individual reads did not propagate into the consensus sequence. Finally, the third example (panel c) shows the sequencing of a synthetic fragment with extremely high GC content (>75%). GC-rich regions are well known for forming stable secondary structures that reduce sequencing

efficiency in conventional methods (Jan Kieleczawa, 2006). Here, Nanopore sequencing successfully produced full-length reads spanning the fragment, enabling accurate reconstruction of the sequence despite its intrinsic complexity.

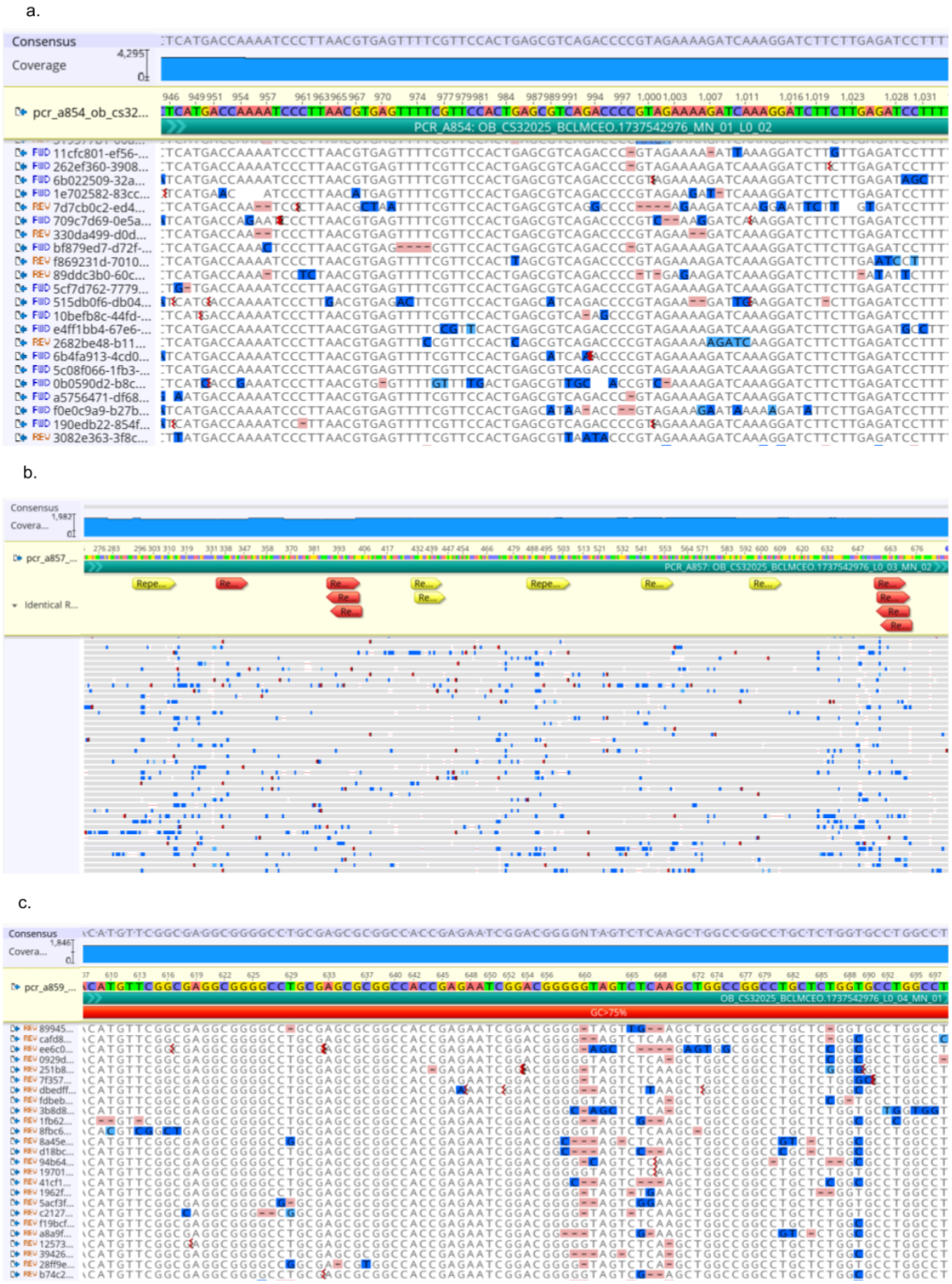

Figure 3.4: ***Nanopore sequencing of representative linear dsDNA fragments.*** *The figure shows three examples of linear double-stranded DNA analyzed with Nanopore sequencing: a) a PCR amplicon with standard sequence composition, b) a PCR product containing direct and palindromic repeats, and c) a synthetic fragment with extremely high GC content (>75%). In all cases, reads spanned the entire fragment length, enabling complete consensus reconstruction. While conventional sequencing methods such as Sanger often fail in repetitive or GC-rich regions, Nanopore sequencing successfully resolved these fragments.*

However, one consistent limitation encountered across experiments is the reduced fidelity of sequencing at the fragment termini. Specifically, the first and last 12–14 nucleotides are often under-represented or exhibit low base-calling confidence despite robust internal coverage (*Figure 3.5*). This is consistent with observations in nanopore performance studies: for instance, Mihovilović et al. demonstrated that DNA strands experience biased capture and translocation dynamics near the nanopore entry and exit, leading to diminished signal resolution in those boundary regions (Mihovilovic et al., 2012). While these limitations do not undermine the validation of internal regions, they become critical in workflows where precise end verification is essential, for example, confirming cloning overhangs required for seamless DNA assembly, and most importantly, ensuring the correct length and integrity of polyA tails (e.g., 120A) in IVT templates, which directly impacts the stability and translational efficiency of therapeutic mRNAs.

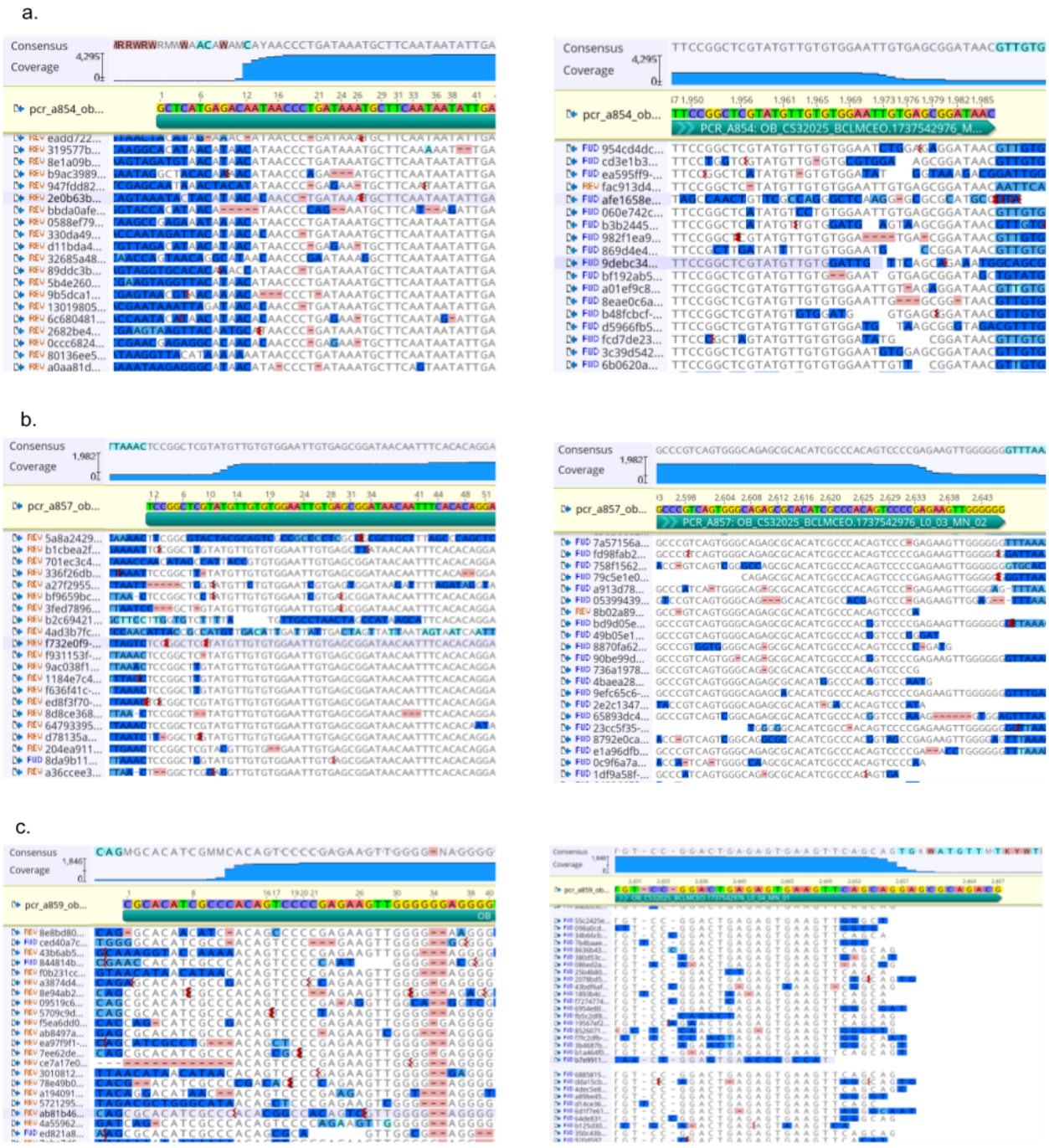

Figure 3.5: Examples of incomplete coverage at the termini of linear dsDNA fragments sequenced with Nanopore. The samples shown correspond to those described in *Figure 3.4*. The left panel illustrates the 5' end of the fragment, while the right panel shows the corresponding 3' end. In both cases, the first and last 12–14 bp are poorly resolved, highlighting a recurrent limitation of Nanopore sequencing at fragment extremities. Short non-aligning stretches visible before and after the termini most likely originate from adapter or alignment artefacts and do not affect the reconstruction of true fragment boundaries.

Together, these results confirm the use of Nanopore sequencing is not restricted to circular plasmids but can also be effectively applied to linear dsDNA molecules of diverse origin and structural composition. This versatility is particularly valuable in synthetic

biology and therapeutic applications, where linear DNA frequently serves as assembly bricks for plasmid construction, as IVT templates for mRNA synthesis, or as standalone functional fragments. Importantly, Nanopore sequencing demonstrates robustness in resolving both simple and structurally complex regions, including repeats and high-GC motifs that often hinder conventional approaches. At the same time, the recurrent loss of sequence fidelity at fragment extremities underscores an inherent limitation that must be considered, especially in workflows where precise end verification, such as the confirmation of cloning overhangs or the accurate validation of polyA tails, is essential. Taken together, these strengths and weaknesses highlight Nanopore sequencing as a powerful, though not exhaustive, quality control solution for linear dsDNA, whose integration into Biofoundry pipelines can substantially enhance throughput and versatility while still requiring complementary strategies for critical end-sequence validation.

3.2.4 Cost analysis of sequencing strategies

An important factor in evaluating the introduction of new sequencing technologies into our Biofoundry workflows is not only their accuracy and scalability, but also their economic sustainability. To assess how the choice of sequencing technology affects plasmid validation, we compared the costs of Sanger and Nanopore sequencing across different experimental scenarios. The analyses considered both sample throughput (number of colonies screened in parallel) and plasmid length, with a reference construct size of 10 kb, reflecting the upper end of typical plasmid sizes used in rAAV production (Naso et al., 2017; Samulski & Muzyczka, 2014).

Considering that one of the strategic aims of ONT is the development of a cost-effective sequencing method, it is relevant to contextualise the expenses associated with the MinION platform. The MinION Mk1B is the most affordable sequencing devices currently available, with an acquisition cost of approximately € 1200. Consumables also contribute to its overall cost-effectiveness: R9.4.1 and R10.4.1 flow cells typically cost around € 800 each, and the Rapid Barcoding Kit 96 (SQK-RBK110.96) is priced around € 1200. Importantly, these consumables are not disposable, as they can be washed and reused for multiple sequencing runs provided sufficient active pores remain. When amortised across their reuse potential and high multiplexing capacity, these consumables reduce

the per-sample cost dramatically, making Nanopore a highly competitive option relative to Sanger sequencing, particularly for medium-to-high throughput colony screening.

Cost scaling with sample throughput

As shown in *Figure 3.6*, when increasing the number of 10 kb DNA constructs to be analyzed, the cost of Sanger sequencing rises strictly linear, as each construct requires a proportional number of sequencing reactions and primers. By contrast, Nanopore sequencing is based on a fixed run cost of about € 300, which does not change whether one sample or 96 samples are processed in the same run. This flat profile highlights the scalability of ONT technology, as the total cost remains constant even as sample throughput increases.

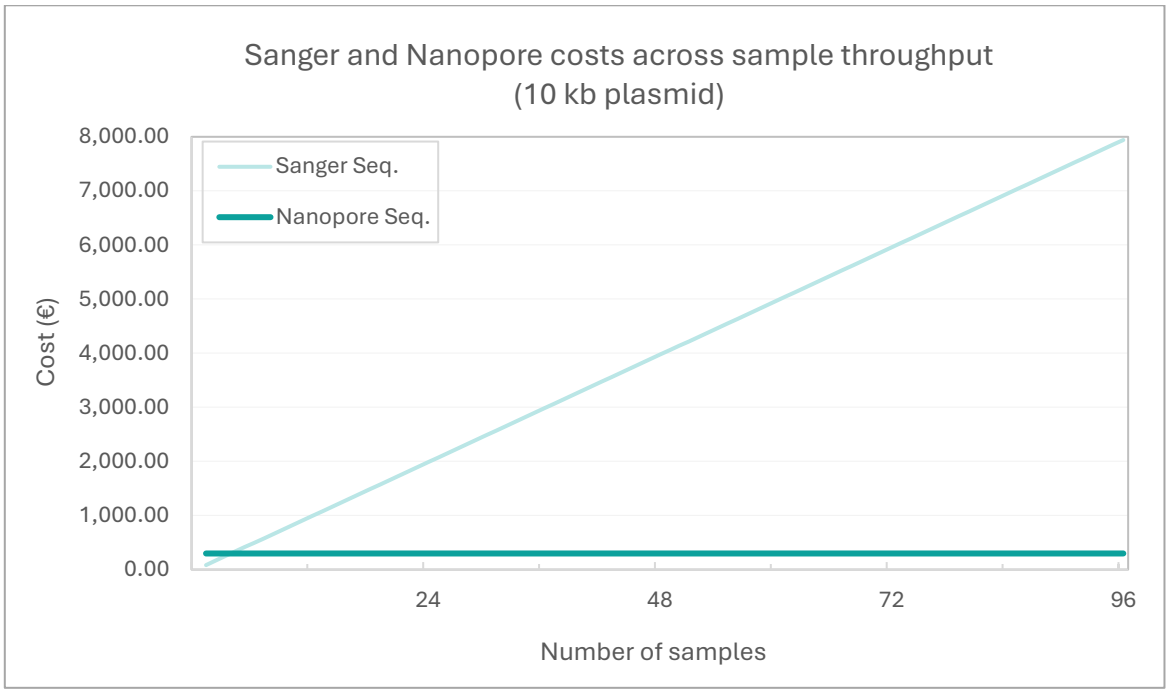

Figure 3.6: Cost scaling with sample throughput. Sequencing costs for Sanger increase linearly with the number of 10 kb dsDNA samples, while Nanopore sequencing maintains a constant total run cost of around € 300 regardless of throughput.

Cost scaling with plasmid length

A similar trend emerges when considering construct size: with Sanger sequencing the cost grows proportionally to the DNA length because each read covers only 800 - 1000 bp, requiring multiple reactions for larger plasmids. For instance, sequencing a 10 kb

construct would necessitate around 10 separate Sanger reactions, with a consequent proportional increase in costs. On the other hand, Nanopore sequencing is independent of plasmid length, since a single run generates reads that can span the entire construct, whether it is 1 kb or 10 kb (*Figure 3.7)*. Because Nanopore sequencing is not limited by plasmid size, it offers a clear advantage for verifying the large helper and packaging plasmids typically used in rAAV production.

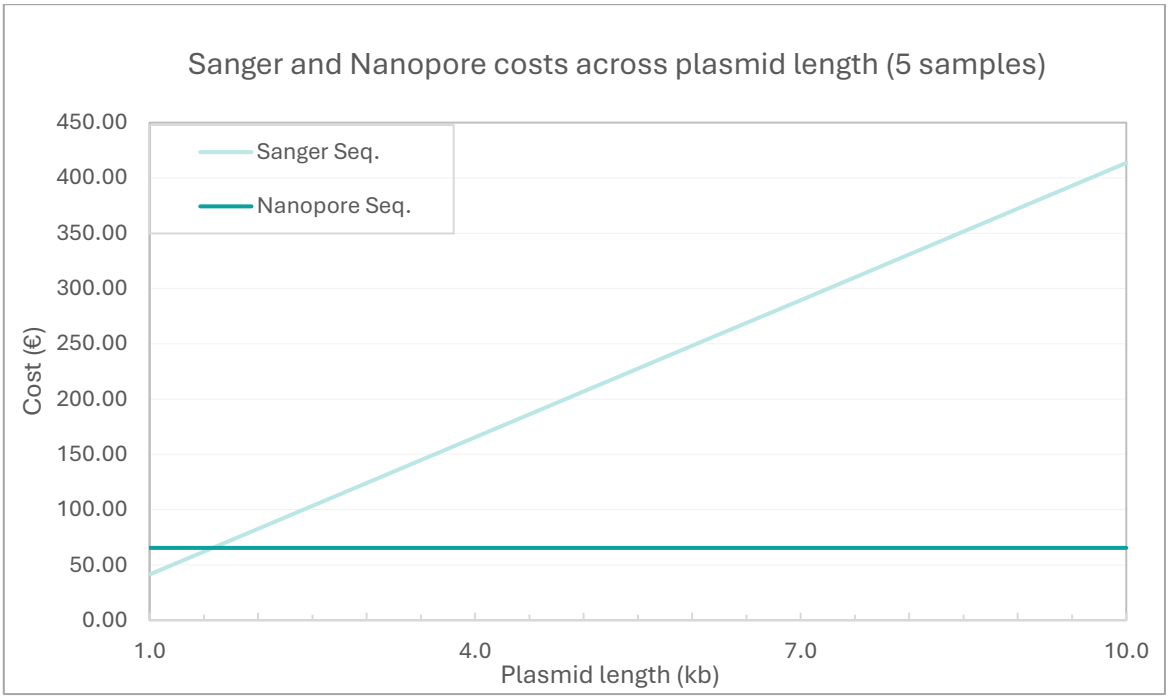

Figure 3.7: Cost scaling with plasmid length. Sanger sequencing costs rise proportionally with plasmid length, whereas Nanopore sequencing remains independent of construct size, since long reads span entire plasmids in a single run.

Nanopore cost per sample

Because Nanopore sequencing relies on a fixed run cost, the per-sample expense depends on how many barcodes are used. Sequencing only a few samples results in relatively high per-sample costs (€ 300.0 – € 60.0 for 1–5 samples), but the value drops rapidly as throughput increases. At full capacity, with 96 samples multiplexed in a single run, the cost stabilizes at around € 3.15 per sample. This non-linear reduction (*Figure 3.8*) demonstrates the clear economic advantage of ONT for Biofoundry applications, where high numbers of constructs are routinely processed in parallel.

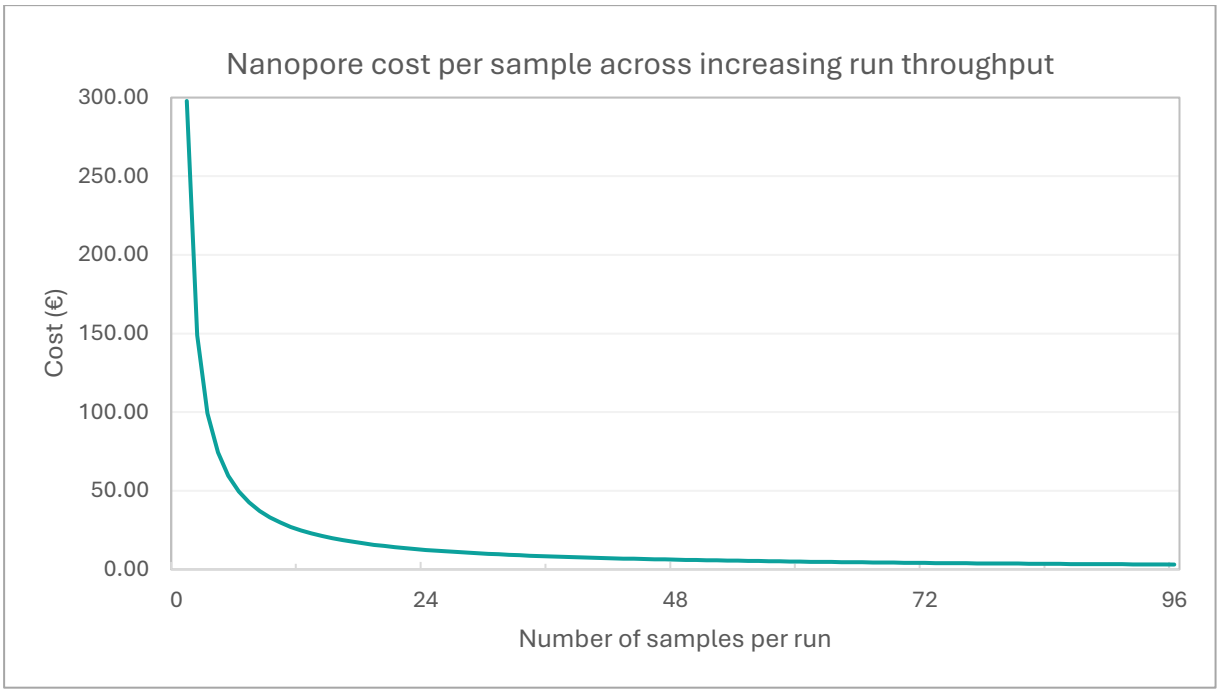

Figure 3.8: Nanopore cost per sample. The per-sample cost of Nanopore sequencing decreases non-linearly with increasing throughput, reaching less than € 4 per sample when 96 barcodes are used in a single run.

In summary, the cost analysis indicates that Sanger sequencing costs scale directly with construct size and sample number, limiting its usefulness for large-scale applications. In contrast, Nanopore sequencing follows a fixed-cost model that is independent of DNA length, with per-sample costs decreasing substantially as throughput increases. This scalability and cost-efficiency make ONT a particularly attractive option for our Biofoundry, where hundreds of plasmids and linear DNA fragments must be validated quickly and reliably.

Operational Cost and Personnel Requirements

In addition to instrument and consumable expenses, a realistic cost assessment must also account for personnel effort and required expertise. In our Biofoundry, Sanger sequencing is fully externalised: samples are shipped to a sequencing facility, and results are returned within 24–48 hours. This minimises hands-on time but makes turnaround dependent on the provider.

Nanopore sequencing is performed entirely in-house and redistributes the workload. Library preparation is relatively simple: requires only a few hours of bench work, with the majority of the protocol being incubation or instrument-driven, and can be carried out by

an undergraduate-level operator after minimal training. In contrast, data analysis represents the most time-consuming step: for a full 96-sample run, the manual alignment and analysis of the barcoded dataset, confirming plasmid identity, coverage, and critical features, can require 3–4 hours, and in some cases demands interpretative judgement from more experienced personnel. This introduces an analytical bottleneck that partially offsets the advantages of ONT's high throughput. However, this limitation is expected to diminish as automated analysis solutions are introduced. A dedicated pipeline developed by our IT team would standardise processing, reduce manual inspection time, and further improve the overall scalability and cost-effectiveness of Nanopore sequencing within the Biofoundry.

3.2.5 Technical challenges and improvements

Despite its successes, several technical limitations were encountered during the implementation of Nanopore sequencing.

Library preparation variability

One limitation encountered during the project was the variability introduced at the library preparation stage, particularly during DNA input normalization prior to barcoding. For each experiment, all DNA samples had to be adjusted to 50 ng, a step performed manually. This procedure was both time-consuming and error-prone, as even small pipetting inaccuracies translated into significant differences in sequencing yield across samples.

This variation is showed in *Figure 3.9*, which reports the distribution of read counts from a 24 barcodes Nanopore run. Instead of a uniform distribution, read counts ranged from as few as 245 reads to as many as 3,064 reads, with a difference of more than a 12-fold. The mean number of reads per barcode was 1,861, with a coefficient of variation of around 39%. Some samples were therefore heavily over-represented, while others received only a fraction of the expected coverage, making downstream analysis less robust. While only one run is shown here, the same variability trend appeared consistently in other experiments, reflecting a general limitation of the manual workflow.

These findings highlight the need of automating the normalization and barcoding steps. The implementation of an automated liquid handling system would help us to minimize pipetting errors, standardize the input amounts, and improve the reproducibility across runs. This implementation would strengthen the robustness of Nanopore sequencing for high-throughput applications within our Biofoundry workflows.

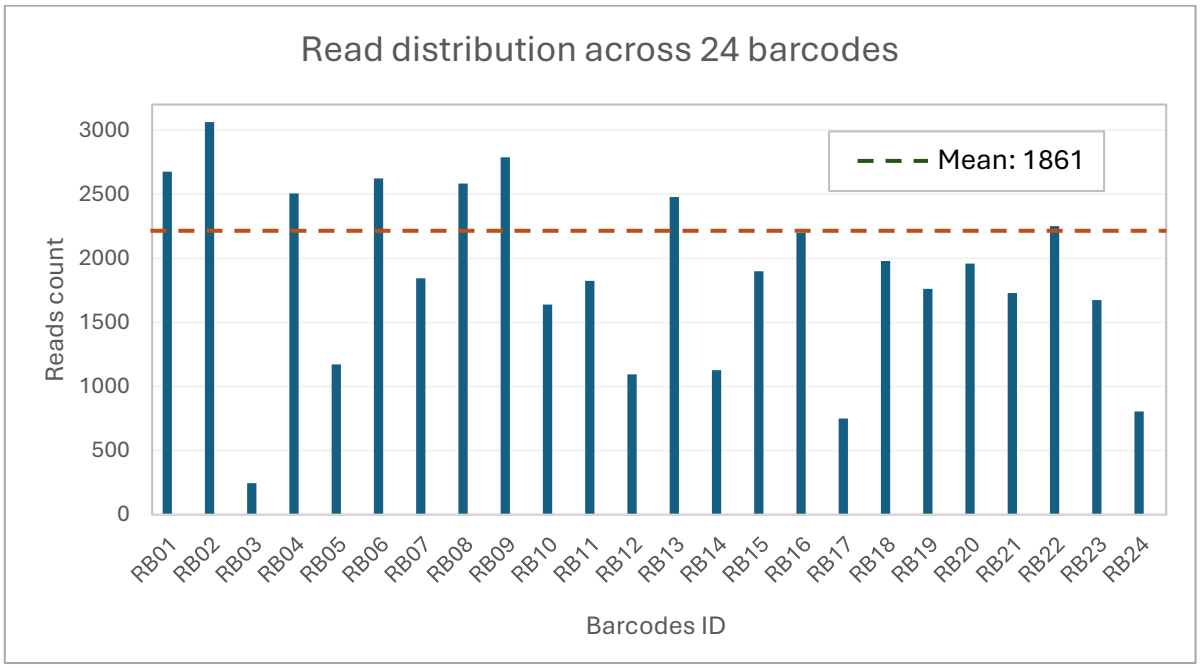

Figure 3.9: Variability in read yield per barcode after manual normalization. The graph above shows the distribution of read counts across 24 barcodes from a single Nanopore run, even if the sample normalization was performed as first step of the library preparation. The uneven representation of barcodes reflects inconsistencies introduced during manual DNA input normalization prior to library preparation.

Sequencing chemistry limitations and transition to R10

With the R9.4.1 flow cell, systematic basecalling inaccuracies were consistently observed. Homopolymeric tracts of 5 or more nucleotides were frequently miscalled, typically appearing as truncations compared to the true sequence, with error rates increasing proportionally as homopolymeric tract length increased. These errors are linked to the underlying chemistry of the technology, which infers nucleotide identity from changes in ionic current as the DNA strand passes through the nanopore: extended homopolymeric regions often fail to generate sufficiently distinct current shifts, leading to uncertainty in basecalling (Y. Wang et al., 2021). These observations are in line with prior reports, which have shown that Nanopore sequencing systematically

underestimates the length of homopolymeric regions, with cytosine- and guanine-rich tracts being particularly error-prone relative to adenine- and thymine-rich stretches (Delahaye & Nicolas, 2021). In addition, we detected recurrent errors at specific trinucleotide motifs, most specifically at CCT and CCA, which were incorrectly interpreted as sequence variants. In fact, the following validation by Sanger sequencing of these candidate clones confirmed these to be false negatives rather than genuine mutations.

By the third year of the project, Oxford Nanopore Technologies had discontinued the R9.4.1 flow cells and library kits, so all following experiments were carried out with the R10 chemistry. This version features a longer barrel and a dual-reader head, improving signal resolution and helping to reduce systematic errors (*Figure 3.10*). In practice, this upgrade led to a clear improvement in data quality. In particular, the recurrent miscalls observed with R9.4.1 at CCT and CCA trinucleotide motifs were resolved, leading to much higher sequence accuracy in these contexts.

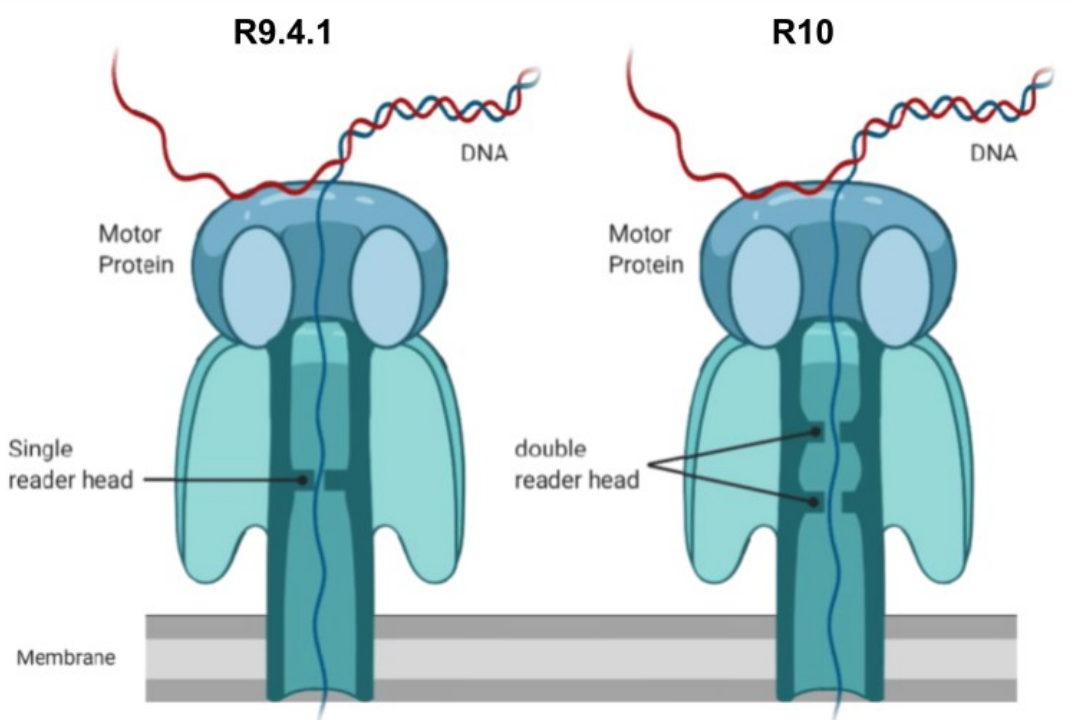

Figure 3.10: Schematic comparison of Oxford Nanopore flow cell chemistries R9.4.1 (left) and R10 (right). The R9.4.1 pore features a single reader head, while the R10 design incorporates an extended barrel with a dual-reader head, enabling improved signal resolution and accuracy.

Another key benefit of the R10 chemistry was the improved handling of homopolymeric regions. Whereas with R9.4.1 even short runs of five or more identical bases were frequently miscalled, R10 allowed reliable detection of shorter tracts, up to approximately five to six nucleotides in length. This improvement is consistent with the mechanistic design of the dual-reader head, which provides a second electrical signal as each k-mer translocates through the pore, thereby increasing the probability of distinguishing repetitive bases. Nevertheless, despite this progress, errors persisted in longer homopolymeric stretches, particularly those extending beyond 8–10 nucleotides, where indel miscalls remained common.

These observations are supported by recent studies: Heba H. Mostafa, (2024) reviewed the progressive advances in ONT chemistries for clinical microbiology, noting that the transition to R10 substantially improved accuracy in homopolymer-rich regions compared to earlier versions. Similarly, Kovaka et al., 2025 demonstrated that the R10.4.1 dual-reader head significantly reduced indel rates in extended homopolymeric runs (≥9 bases), confirming both the strengths and residual weaknesses of the platform.

Together, these findings indicate that, while R10 has mitigated some of the systematic limitations of earlier R9 flow cells (short homopolymers and specific trinucleotide motifs), the accurate resolution of long homopolymeric sequences remains an open issue for nanopore sequencing.

3.2.6 Comparative accuracy and throughput

When benchmarked against Sanger sequencing, ONT showed complementary rather than directly comparable performance. Sanger sequencing remains the most accurate method for base-level verification, with reported error rates of 0.001% (Q40) and very low systematic biases (Cheng & Xiao, 2022). This makes it a gold standard for confirming short DNA regions or standard constructs, such as small plasmid inserts or linear dsDNA fragments. However, its use is fundamentally constrained: individual reads rarely exceed 800–1000 bp, and sequencing quality declines sharply in the presence of structural complexity such as inverted terminal repeats (ITRs), palindromes, long repeats, or regions with high GC content (Jan Kieleczawa, 2006). As a consequence, Sanger sequencing

cannot resolve entire plasmids or highly complex constructs, which limits its utility in advanced synthetic biology and cell and gene therapy pipelines.

By contrast, Oxford Nanopore sequencing provides a broad and comprehensive overview of whole DNA constructs. Although its raw read accuracy is lower, typically ranging from Q10 to Q20 (90–99%) for R9.4 and R10.4 flow cells (Rang et al., 2018; T. Zhang et al., 2024), ONT generates long reads capable of spanning entire plasmids as well as linear DNA fragments of varying sizes. More importantly, Nanopore sequencing can traverse sequence motifs that are systematically inaccessible to Sanger sequencing, such as ITRs, repetitive domains, and regions with very high GC content. This capacity ensures that both the newly synthesized inserts and the conserved plasmid backbones are fully validated within a single run, delivering information that cannot be obtained with Sanger sequencing.

Taken together, the comparison between Sanger and ONT highlights both strengths and persistent gaps. Sanger sequencing is best suited for targeted, high-accuracy confirmation of short and simple constructs, but its limited read length and failure in repetitive or structurally complex regions make it inadequate for the comprehensive verification required in cell and gene therapy pipelines. Nanopore sequencing overcomes these structural limitations by enabling scalable, whole-construct analysis, yet its lower per-base accuracy and systematic challenges in homopolymeric regions remain unresolved.

As a result, the combination of Sanger and ONT does not provide a fully reliable quality control strategy. These limitations highlighted the need for a complementary sequencing technology that could combine high per-base accuracy with reliable performance in homopolymer regions. For this reason, Illumina sequencing was chosen for a subsequent evaluation.

3.2.7 Implications for Biofoundry workflows and CGT applications

The introduction of Nanopore sequencing into our Biofoundry workflow significantly changed how DNA constructs were validated. Its rapid, high-throughput, and cost-effective full-plasmid sequencing, the systematic screening of large numbers of plasmids was made possible. This included also those plasmids with ITRs, repetitive sequences,

and high GC content, features that Sanger sequencing could not reliably address. Importantly, ONT sequencing allowed both the verification of newly synthesized inserts and the backbone regions of plasmids, ensuring complete construct integrity. By reducing the chance of undetected errors in conserved regions, which are not always verified with Sanger Sequencing, this comprehensive approach improves biosafety and helps prevent issues that could affect downstream rAAV packaging or therapeutic use.

Another key advantage demonstrated in this project was the ability of ONT sequencing to analyze linear dsDNA fragments, such as PCR products and linear fragments generated by enzymatic digestion. This versatility expands its role beyond plasmid validation and makes it directly applicable to the quality control of DNA templates intended for in vitro transcription (IVT). In the context of cell and gene therapy, ensuring that linear DNA templates are error-free is crucial for producing high-quality mRNA molecules. Therefore, the capacity of Nanopore sequencing to verify both plasmid and linear DNA within the same pipeline represents a significant advantage for companies operating at the interface of rAAV and mRNA-based therapeutics.

From an operational perspective, ONT sequencing offers both economic and temporal advantages within our Biofoundry workflow. Unlike Sanger sequencing, where costs and processing time linearly increase with fragment length and number of samples, with Nanopore sequencing, the primary expense lies in the flow cell and library preparation reagents, which represent a fixed cost per run. This cost is then distributed across the number of barcoded samples loaded, and therefore the effective per-sample price decreases with increasing multiplexing. This scalability, combined with turnaround times of only a few hours, allowed us to process thousands of constructs more efficiently during the PhD project, reducing synthesis risks and ensuring timely delivery of high-quality DNA products. For the Cell and Gene Therapy companies we collaborate with, this translated into shorter development cycles and improved compliance with pre-clinical and GMP quality requirements.

At the same time, this work revealed some limitations that suggest areas for further improvement and optimization. An important source of variability is the manual library preparation. This emphasizes the urge of automated workflows implementation in order to ensure consistent read distribution across barcodes. Despite the improvements

introduced with the R10 chemistry, which eliminated systematic errors in CCT/CCA motifs and reliably resolved short homopolymer tracts (up to 5–6 bases), longer homopolymeric regions, such as polyA tails, remained challenging and continued to contribute to residual indel errors.

The results obtained in this PhD clearly demonstrate that ONT sequencing already provides a robust and transformative platform for the verification of plasmid and linear DNA constructs within our Biofoundry. In the field of Cell and Gene Therapy, the unique combination of scalability, flexibility, cost-effectiveness, and molecular completeness firmly establishes ONT sequencing as a key enabling technology to accelerate the development of rAAV- and mRNA-based therapies.

3.3 DNA sequencing with Illumina

To complement the Oxford Nanopore sequencing platform previously established in the Biofoundry, I carried out a systematic benchmarking experiment with Illumina sequencing during my second PhD year. The aim was to determine whether Illumina could compensate for intrinsic limitations of Nanopore, including its lower per-base accuracy (Q10), its persistent difficulties with homopolymeric stretches such as long polyA tails, and the recurrent loss of fidelity at fragment extremities. At the same time, we sought to evaluate whether Illumina could serve as a scalable alternative to routine Sanger sequencing, which increases linearly in cost with sample number and construct length and fails in structurally complex regions such as ITRs, essential components of rAAV vectors. To address these questions, we designed a benchmarking experiment on 12 representative samples sequenced on the Illumina iSeq100 platform. The panel included constructs with ITRs, high GC content, long polyA stretches, and linear fragments requiring precise end validation, thereby enabling us to assess Illumina's performance across the main challenges encountered in our quality control pipeline. As shown in the following section, Illumina sequencing successfully addressed many of these bottlenecks, providing both high accuracy and reliable coverage in contexts where Nanopore or Sanger sequencing showed limitations.

3.3.1 Experimental design and run metrics

The benchmarking experiment was performed on 12 DNA samples prepared with the Illumina DNA Prep (Tagmentation) kit and sequenced on the iSeq100 platform. The run generated around 1.7 Gb of paired-end 2×150 bp reads, with more than 88% of bases reaching Q30 quality. This confirmed the expected high accuracy of Illumina sequencing. The panel was deliberately assembled to span the principal challenges in our QC pipeline, ITRs (2–6 copies), high-GC and repetitive regions, long polyA tracts (42–120A), and a designed variant library, so as to benchmark Illumina performance across the full problem space relevant to adoption. One sample (497 bp) was a designed variant library, included solely to benchmark variant-detection sensitivity and not central to the rAAV/mRNA focus. *Table 3.1* summarizes the sample set and their defining features.

Sample ID	Type	Size (bp)	Sample property
#1	Plasmid	23701	23 kb large construct
#2	Plasmid	4013	398 bp long repeats, 6 ITRs
#3	Linearized	1801	Linearized 398 bp long repeats, 6 ITRs
#4	Plasmid	4364	PolyA 42 bp
#5	Plasmid	3847	PolyA 80 bp
#6	Plasmid	3509	PolyA 120 bp
#7	Plasmid	7412	PolyA 120 bp
#8	Plasmid	11467	High GC content region, several repeats
#9	Linearized	589	Short Linearized 120 bp polyA
#10	Linearized	23628	Long fragment
#11	Plasmid	4906	2 ITRs
#12	Linearized Library	497	Library

Table 3.1: Summary of the 12 DNA constructs included in the Illumina benchmarking experiment. The dataset comprised plasmids and linearized fragments ranging from 497 bp to 23 kb. Each construct was selected to test a specific sequencing challenge, including inverted terminal repeats (ITRs), high GC content, long polyA stretches, and designed variant libraries.

3.3.2 Performance across challenging sequence contexts

Large construct

Sample 1 is a long plasmid (23,701 bp) included to test Illumina performance on large constructs. Sequencing on the iSeq100 (2x150 bp) yielded full-length breadth of coverage with a homogeneous depth profile across the entire molecule, without drop-outs at junctions or repetitive segments (*Figure 3.11*). The Geneious alignment shows a flat coverage trace and a clean consensus over the full 23 kb. The assembled consensus matched the expected reference with 100% identity, consistent with the run's high base quality (≥Q30 for ≥88% of bases). These findings indicate that Illumina sequencing ensures even representation after tagmentation and supports high-confidence, base-level verification suitable for QC of whole plasmids. Importantly, the absence of local coverage biases in this construct suggests that sequence length alone does not compromise Illumina performance, and that the platform can be reliably applied to the verification of large plasmids provided that they do not contain high-GC or repetitive motifs. This aligns with previous reports that Illumina sequencing performs consistently across long templates when the base composition is balanced (Aird et al., 2011; Benjamini & Speed, 2012).

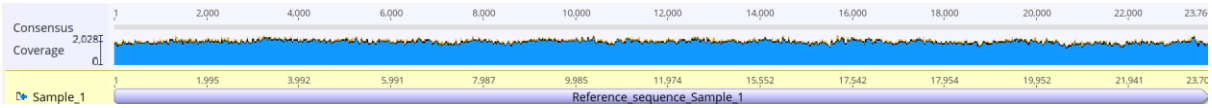

Figure 3.11: Allignment of Sample 1, a 23.7 kb plasmid. Illumina iSeq100 (2×150 bp) reads have been aligned to the reference sequence in Geneious. Here it's showed an even per-base coverage across the entire construct, with no evidence of drop-outs or edge effects. The resulting consensus sequence perfectly matched the reference, confirming 100% identity.

PolyA homopolymers (42–120A)

To probe Illumina's ability to resolve homopolymeric tracts, we included Sample 4 (plasmid, 4,364 bp; 42A polyA) and Sample 6 (plasmid, 3,509 bp; 120A polyA). Geneious alignments showed homogeneous coverage through both tracts and exact base-level reconstruction at the homopolymer–flank junctions, with no compression/expansion or indel laddering. In both cases, the consensus matched the reference sequence. These findings demonstrate that Illumina sequencing can accurately recover both mid-length

(42A) and long (120A) polyA stretches (*Figure 3.12*). This performance contrasts with the well-known difficulties of Nanopore sequencing in long homopolymer runs, where systematic deletion or length variability often occurs due to limitations in signal resolution (Delahaye & Nicolas, 2021). Thus, Illumina emerges as the preferred modality for verifying polyA length and integrity in constructs relevant to IVT mRNA workflows, ensuring reliable assessment of a key feature for mRNA stability and translation efficiency (A. Sachs, 1990; Passmore & Coller, 2022).

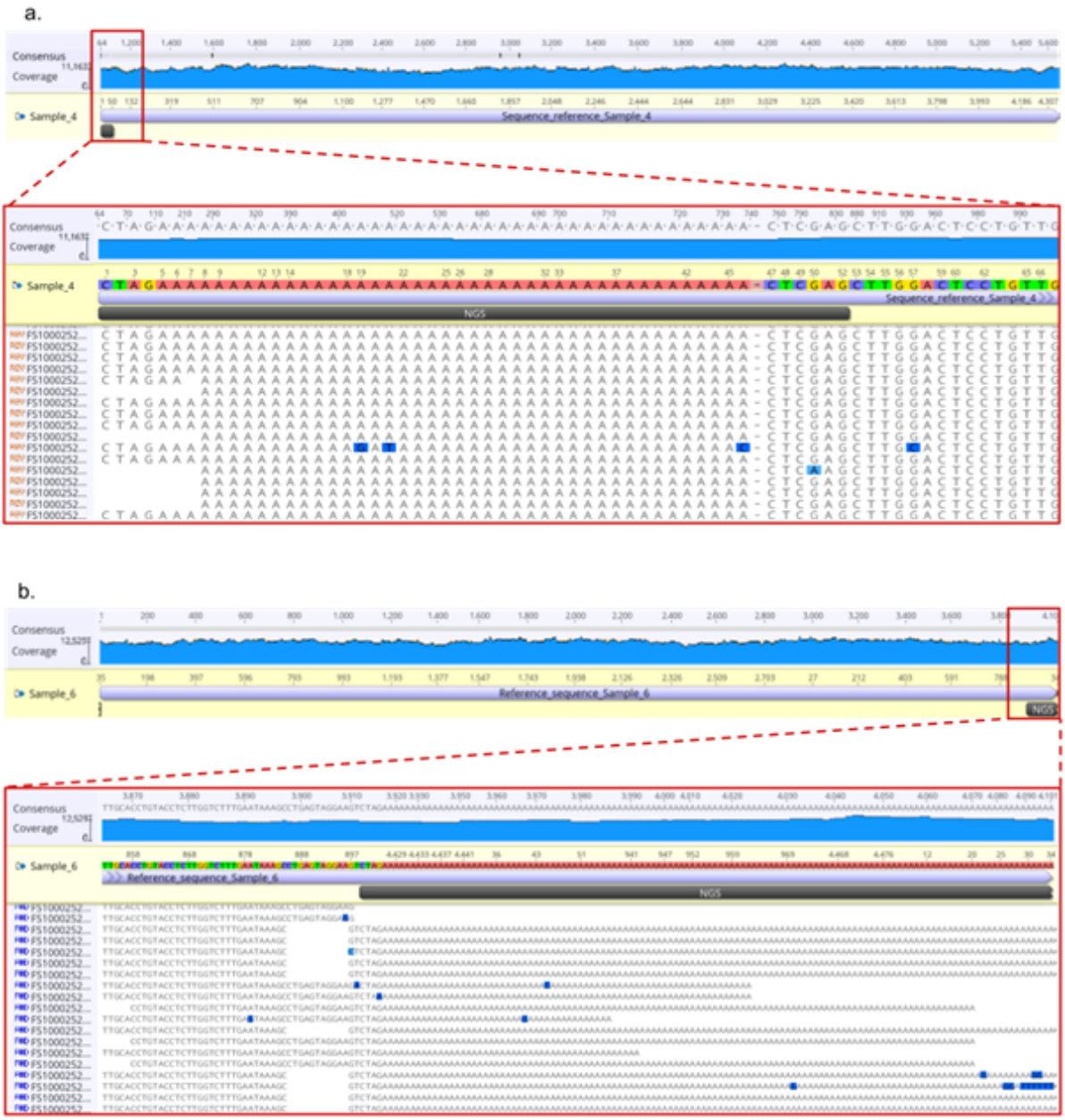

Figure 3.12: Illumina resolution of homopolymeric polyA tracts. *(a) Sample 4 (4,364 bp; 42A): Geneious alignment of iSeq100 2×150 bp reads showing uniform coverage and exact base-level reconstruction of the polyA tract. (b) Sample 6 (3,509 bp; 120A): complete, indel-free reconstruction of a long polyA tail with homogeneous coverage across the region. Alignments are against the expected references.*

ITR-containing constructs (up to six ITRs)

Sample 2 (4,013 bp) was synthesised for a customer and included in the Illumina benchmarking experiment because it contains six inverted terminal repeats (ITRs; 398 bp each), a sequence context directly relevant to rAAV quality control. Sequencing on the iSeq100 (2x150 bp) yielded full-length coverage and a base-accurate consensus across all six ITR units in the Geneious alignment, with clear repeat boundaries and no systematic indels. This result demonstrates that Illumina sequencing can resolve multi-ITR arrays at base level, despite their repetitive and structurally constrained nature (*Figure 3.13*). Importantly, this class of sequences cannot be reliably verified by Sanger sequencing, as the strong secondary structure of ITRs leads to premature termination and unreadable chromatograms (Mroske et al., 2012). By contrast, Illumina enables per-base confirmation across all repeats, providing a robust QC solution for one of the most critical sequence elements in rAAV plasmids. These findings position Illumina as a powerful complement to Nanopore, which can span long constructs including ITRs but typically with lower base-level accuracy.

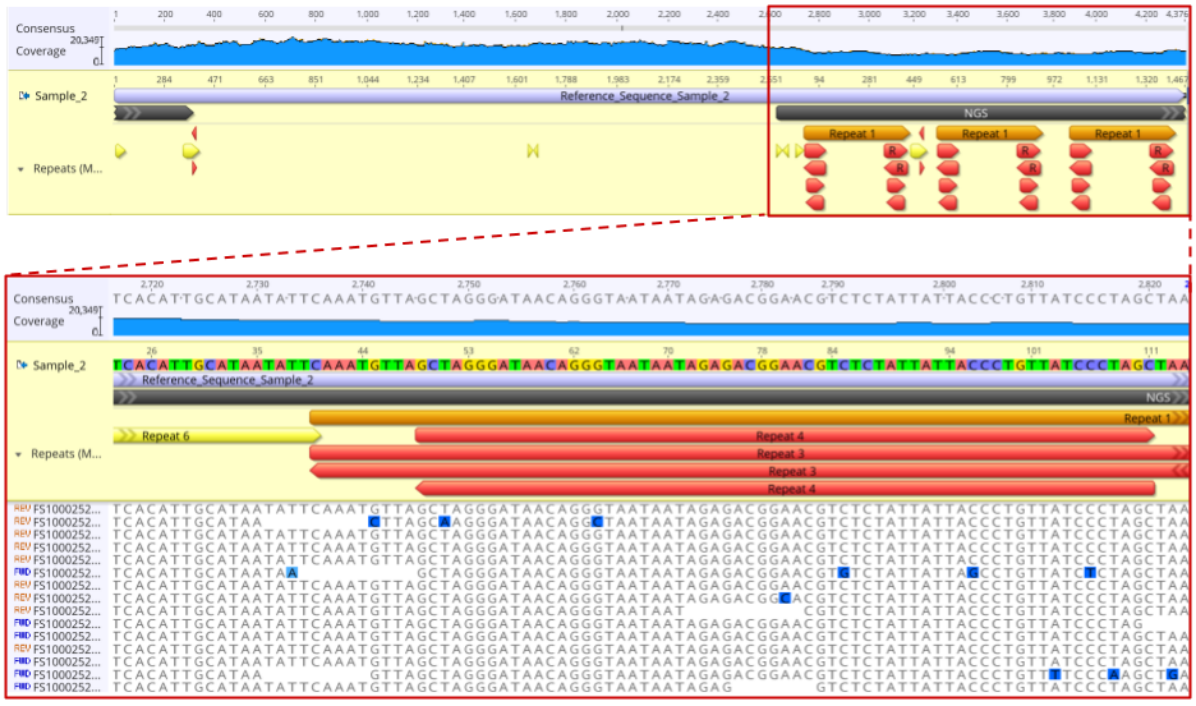

Figure 3.13: Sequencing alignment of Sample 2, with the resolution of 6 ITRs with Illumina sequencing. The reads of iSeq100 for Sample 2 have been aligned to the reference sequence with Geneious software. The top panel displays a full-length coverage across the 4.0 kb plasmid, while the bottom panel provides a zoomed view of the six annotated ITR repeat units with the corresponding read pileup. In both cases, the consensus sequence matched the reference exactly.

GC-rich and repetitive regions

Illumina sequencing was evaluated for its capacity to resolve structurally complex motifs, such as inverted terminal repeats (ITRs) and GC-rich regions. While Sanger sequencing fails to span these sequences, Illumina produced full-length alignments with base-level accuracy, demonstrating its suitability for quality control of multi-ITR arrays. Nonetheless, reductions in read coverage were consistently observed in regions with elevated GC content, particularly when it exceeded 60%. *Figure 3.14* illustrates these findings: in *panel a* it's showed the accurate reconstruction of 6 ITR repeats of Sample 2 (around 50% GC); *panel b* highlights a marked coverage drop in a highly GC-rich region of Sample 8 (64–83% GC); *panel c* depicts a similar reduction across the ITRs region of Sample 11 (61–73% GC). Taken together, these results suggest that although Illumina sequencing is accurate, it is systematically biased by local sequence composition, leading to the underrepresentation of GC-rich segments. The observed GC-rich coverage bias in Illumina data is likely rooted in library preparation, specifically, during PCR amplification and possibly exacerbated by transposase-based tagmentation, where GC-rich DNA may be less accessible or inefficiently processed, leading to uneven representation in the sequencing output. This hypothesis aligns with prior findings that PCR during Illumina library prep is a principal source of GC-dependent bias (Aird et al., 2011), and that enzymatic fragmentation methods, including tagmentation, may introduce additional insertional bias in GC-rich contexts (Ribarska et al., 2022).

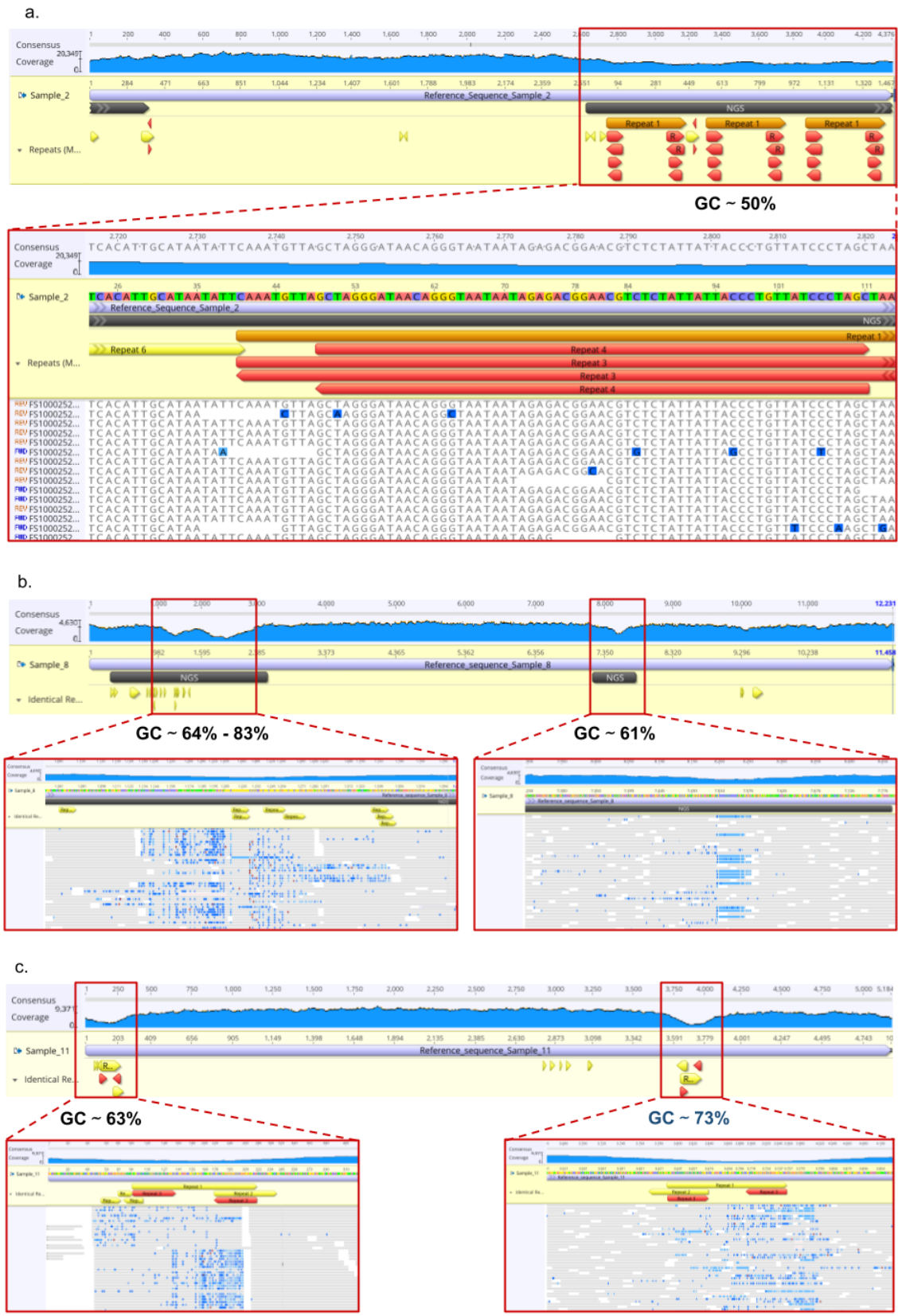

Figure 3.14: Illumina sequencing of constructs containing ITRs and GC-rich regions. (a) Sample 2 (50% GC): full-length reconstruction of six inverted terminal repeats (ITRs), demonstrating Illumina's ability to resolve repetitive motifs. (b) Sample 8 (64–83% GC): pronounced coverage drop and noisier read alignments across a highly GC-rich region, despite accurate consensus calls. (c) Sample 11 (61–73% GC): reduced coverage and increased noise within ITRs, highlighting the systematic underrepresentation of GC-rich sequences.

Cross-platform concordance and complementarity

To further dissect the effect of GC-rich content on sequencing performance, Sample 8 was selected as a case study to directly compare Illumina, Oxford Nanopore, and Sanger sequencing. Illumina sequencing (*Figure 3.15*, panel a) produced an accurate consensus across the construct but exhibited two pronounced drops in coverage: one in the highly repetitive region with 64–83% GC, and another in the 61% GC region that did not contain repeats. These behaviors may stem from library preparation, where GC-rich DNA is less efficiently processed during tagmentation, with the effect being particularly severe when repeats are present. Oxford Nanopore sequencing (*Figure 3.15*, panel b) achieved a more uniform coverage profile across both regions, confirming its ability to span GC-extreme motifs independently of sequence repetitiveness. However, Nanopore reads showed difficulties in homopolymeric tracts, especially in the 16-bp polyG stretch embedded within the repetitive region, consistent with the known limitations of this technology. With Sanger sequencing the reads initiated in the GC-rich regions but were stalled when encountering stretches of consecutive cytosines, resulting in incomplete coverage in both highlighted windows and underscoring its inability to resolve sequences with elevated GC content, regardless of the presence of repeats *(Figure 3.15*, panel c).

Overall, these findings indicate that Illumina and Nanopore form the most effective complementary couple among the technologies evaluated. Illumina delivers base-level precision but is affected by GC-related coverage bias, whereas Nanopore provides a broader accessibility across structurally complex (ITRs) and GC-extreme regions, despite systematic errors in homopolymers. Together, the two platforms enable comprehensive construct verification and high-confidence base calling for challenging and complex DNA constructs, while Sanger sequencing remains unsuitable for these sequence contexts.

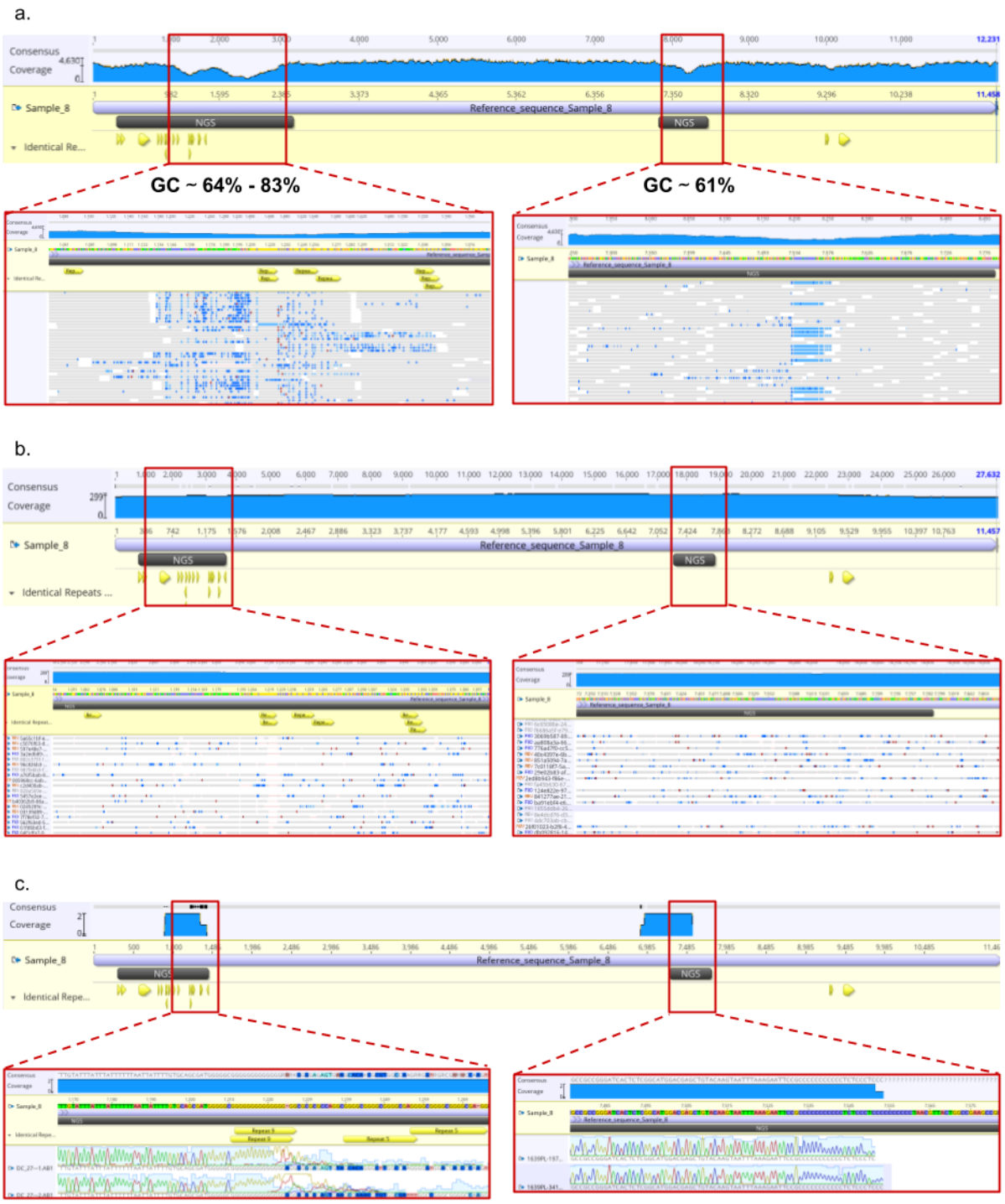

Figure 3.15: Cross-platform sequencing comparison of Sample 8 in two GC-rich regions. (a) Illumina generated an accurate consensus but showed coverage drops and noisier read alignments in the GC-rich segments (64–83% GC with repeats; 61% GC without repeats) compared to the rest of the construct. (b) Nanopore provided more uniform coverage across these regions, though systematic errors appeared in homopolymers, particularly in the 16-bp polyG tract. (c) Sanger produced a progressively noisy signal in the GC-rich repetitive region and terminated at the cytosine homopolymer stretch, leaving both windows only partially covered.

3.3.3 Sequencing of linear templates

Beyond plasmids, we also evaluated the performance of Illumina sequencing on linear double-stranded DNA (dsDNA) constructs, with particular attention to the extremities where Nanopore sequencing showed recurrent limitations. These fragments represent a critical molecular class in synthetic biology and therapeutic applications, as they are commonly used as assembly bricks for plasmid construction, as templates for in vitro transcription (IVT), or as standalone functional elements. Ensuring their complete and accurate sequence verification, including the terminal regions, is therefore essential for downstream processes.

In our benchmarking experiment, some linear fragments were sequenced on the Illumina iSeq100 platform as representative samples of linear dsDNA. Unlike Nanopore, which consistently failed to resolve the first and the last 12–14 nucleotides, Illumina sequencing achieved uniform coverage across the entire molecule, including both 5' and 3' extremities (*Figure 3.16*). Its high per-base accuracy (Q30) allowed full consensus reconstruction without local uncertainty at fragment ends. This capability is particularly important in two contexts: the validation of cloning overhangs required for seamless DNA assemblies, and the precisely reconstruction of polyA tails boundaries, such as the 120A template, which directly affects the stability and translational efficiency of IVT-derived mRNAs.

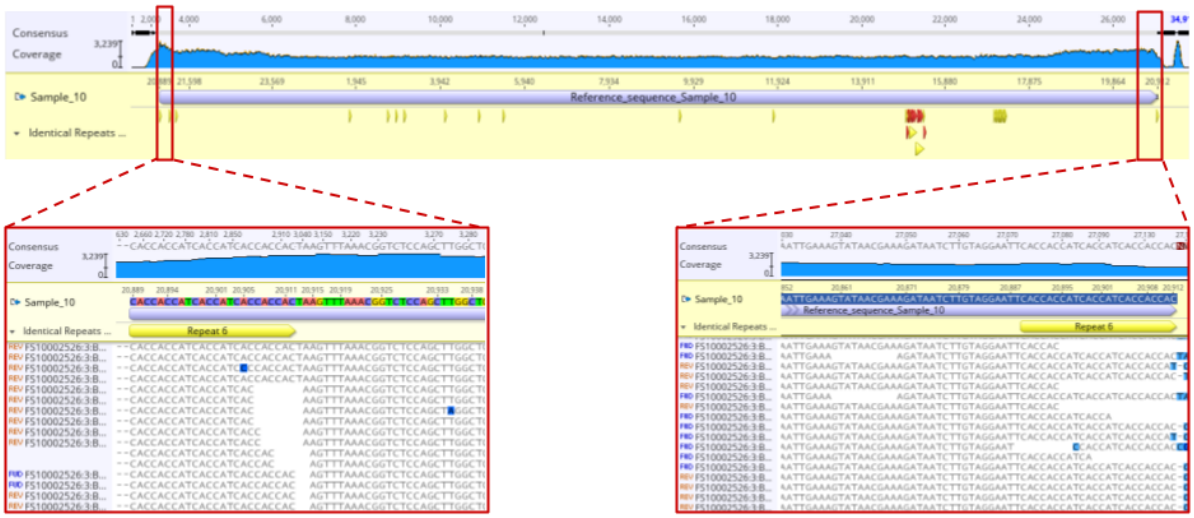

Figure 3.16: Illumina sequencing resolves the extremities of linear dsDNA fragments with no coverage drop. *For sample 10, the upper panel shows uniform coverage across the full fragment, while the lower panels zoom into the 5' (left) and 3' (right) termini. Unlike Nanopore, which missed the first and last 12–14 bp, Illumina achieved complete end-to-end coverage, enabling accurate validation of cloning overhangs and polyA boundaries. Occasional short non-aligning stretches before or after the termini, also seen in Nanopore datasets, likely reflect library or alignment artefacts without affecting sequence reconstruction.*

Together, these results highlight Illumina's strength in complementing Nanopore sequencing for linear DNA constructs. While Nanopore offers versatility and the ability to span structurally complex or GC-rich regions, Illumina ensures that critical boundary sequences are not lost, thereby enabling complete and accurate end-to-end validation of linear dsDNA fragments. This duality reinforces the rationale for integrating both platforms within our Biofoundry workflow, maximizing robustness in quality control across diverse molecular contexts.

3.3.4 Cost analysis

The cost comparison of sequencing platforms for a 10 kb plasmid, representative of rAAV constructs, highlights clear differences in their economic models (*Figure 3.17*). For Illumina, the iSeq100 device, which is the most compact and affordable instrument in the Illumina portfolio, has an acquisition cost of approximately € 20,000, reflecting the higher technological complexity of short-read optical sequencing platforms.

For routine use, the iSeq100 run cost is around € 4,800, independent of the number of samples. The estimated cost of € 4,800 per iSeq100 run reflects the combined cost of the sequencing cartridge (€ 700), library-preparation reagents (€ 2,900 for 96 samples), and consumables (€ 300), including standard operational overheads. As a result, Illumina

sequencing is characterized by a constant cost profile, where efficiency is maximized only when the flow cell is fully utilized (96 constructs). In contrast, Sanger sequencing scales linearly with sample number, since each plasmid must be sequenced individually. For a 10 kb plasmid, the crossover point between Sanger and Illumina occurs at 60 constructs: below this threshold Sanger is more economical, while above it Illumina becomes more cost-effective. Nanopore sequencing shows a similar constant-cost profile to Illumina, but with a lower overall run cost, making it advantageous when analyzing full 96-sample batches of plasmids in the 10 kb range. Taken together, this analysis illustrates how the three platforms occupy distinct niches: Sanger remains suitable for small-scale projects, Illumina is optimal for high-throughput confirmatory sequencing with maximum accuracy, and Nanopore offers a cost-efficient alternative for large-scale runs where rapid turnaround and lower cost per construct are prioritized.

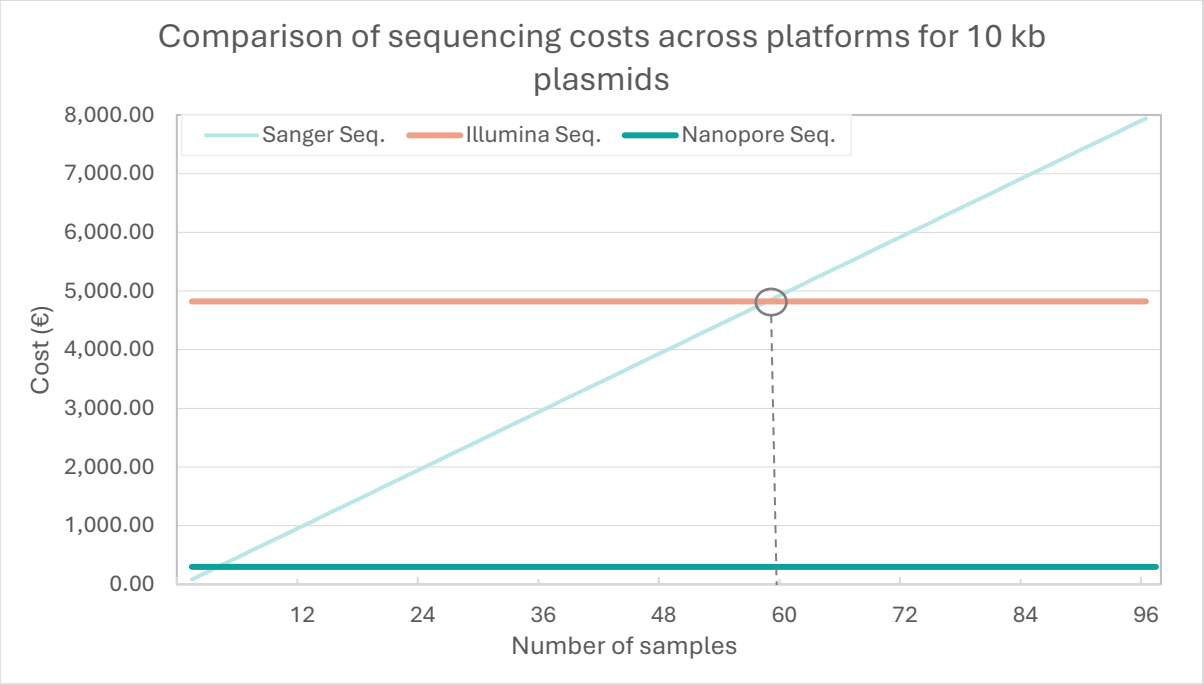

Figure 3.17: Cost comparison of Sanger, Illumina, and Nanopore sequencing across sample throughput for 10 kb plasmids. *The total sequencing cost was calculated for plasmids of 10 kb, representative of rAAV constructs, as a function of the number of samples per run. Sanger sequencing scales linearly with sample number, since each plasmid must be sequenced individually. Illumina (iSeq100) and Nanopore sequencing instead follow a constant cost profile, as the per-run expense is fixed regardless of the number of pooled samples. For Illumina, the crossover point with Sanger occurs at 60 constructs: below this threshold, Sanger is more economical, while above it Illumina becomes more cost-effective. Nanopore shows a lower overall run cost than Illumina, making it the cheapest option when full runs (96 samples) of 10 kb plasmids are performed.*

Operational Cost and Personnel Requirements

In addition to reagent and instrument costs, Illumina sequencing carries specific operational and personnel considerations that influence its overall economic profile. Library preparation for the iSeq100 is more elaborate than Nanopore barcoding, involving multiple steps such as tagmentation, dual-indexed amplification, and size-selection cleanups, typically requiring several hours of hands-on work and careful execution by trained operators. Moreover, because we sequence plasmids and linear dsDNA rather than standard genomic libraries, the built-in Illumina analysis tools cannot be used in our Biofoundry context. Therefore, the obtained FASTQ data must be processed manually, requiring mapping to construct-specific references and visual inspection of coverage and consensus accuracy, an analytical workload comparable to Nanopore sequencing and which requires several hours of expert review. As with ONT, this analytical bottleneck is expected to improve with the development of an automated in-house pipeline tailored to plasmid and synthetic-DNA QC, which will streamline data processing and reduce manual inspection time.

3.3.5 Practical considerations for pipeline integration

While the cost analysis indicates when Illumina becomes economically advantageous, its practical role in a Biofoundry also depends on operational factors. The iSeq100 requires about 18 hours for a run, substantially longer than Nanopore sequencing (1 – 4 hours), although it can reliably achieve a high per-base accuracy (≥Q30). Because of this trade-off, Illumina serves best as a confirmatory platform rather than for primary screening: although slower, it provides the accuracy and confidence needed for final sequence verification. On the other hand, Nanopore sequencing is better suited to rapid, high-throughput screening, where speed and scalability matter more than maximum base accuracy.

Both platforms can process up to 96 samples per run, supporting high-throughput analysis of plasmids and IVT templates. However, Illumina's reliance on enzymatic tagmentation and PCR during library preparation introduced uneven coverage in GC-rich regions in our benchmarking experiments. Although consensus accuracy remained high,

these biases emphasize the value of complementary sequencing approaches and point to the need for workflow optimization in extreme sequence contexts.

Another important consideration is the availability of Illumina Connected Analytics (ICA), a cloud platform that could enable automated analysis once properly programmed and integrated with our bioinformatic workflow. ICA has the potential to streamline Illumina-based processes by reducing operator time, enhancing reproducibility, and ensuring standardization and traceability. If successfully implemented, it could play an important role in incorporating Illumina into regulated Biofoundry pipelines. Combined with the automation of library preparation through and automated liquid handling system, these developments may position Illumina sequencing as a robust QC platform, complementing Nanopore's flexibility with definitive, base-accurate verification of critical constructs.

3.3.6 Accuracy and coverage in complex sequence contexts and fragment extremities

Illumina sequencing demonstrated consistent performance across a range of challenging sequence contexts tested in the stress experiment. Constructs containing multiple ITR units were fully resolved at base level, including a plasmid with six tandem ITRs, where Sanger sequencing systematically failed. Similarly, homopolymeric stretches, particularly polyA tails up to 120 nucleotides, were sequenced without loss of accuracy, in contrast to the systematic errors observed with Nanopore. In the case of linear dsDNA fragments, Illumina sequencing also provided complete and uniform coverage across the full length, correctly resolving both the 5' and 3' extremities. This directly overcomes the recurrent limitation observed with Nanopore, where the first and last 12–14 bp of fragments remained underrepresented, and is particularly important for verifying cloning overhangs and polyA boundaries in IVT templates.

In GC-rich regions (>60%), the samples sequenced with Illumina showed localized coverage drops, which may reflect biases introduced during enzymatic tagmentation. Nevertheless, despite these depth variations, the final consensus sequences was consistently achieved with high accuracy (≥Q30), showing that the platform can still produce accurate assemblies even under challenging sequence contexts.

When benchmarked against other technologies, Illumina offered the highest base-level accuracy: it consistently outperformed Nanopore in homopolymers and fragment termini, and Sanger in structurally complex regions, while still achieving coverage uniformity sufficient for consensus calling. Nanopore, by contrast, provided greater depth in some GC-rich windows but at the cost of lower per-base fidelity (Q10), whereas Sanger retained its role as a reference method for short, standard constructs but proved inadequate for repetitive or secondary-structure-rich regions.

Overall, these results position Illumina as a reliable platform for confirmatory analysis of complex constructs, thereby complementing the broader coverage and structural accessibility provided by Nanopore. Even though Illumina sequencing may show localized coverage biases, its ability to deliver highly accurate consensus sequences across ITRs, polyA tracts, GC-rich regions, and fragment termini makes it particularly valuable for final verification, reinforcing its complementary role alongside Nanopore within an integrated quality control pipeline.

3.3.7 Implications for Biofoundry workflows and CGT applications

Collectively, the results highlight that Illumina sequencing, while less flexible in terms of turnaround time and run cost optimization, provides a level of accuracy that is critical for the definitive validation of complex DNA and RNA constructs. Its ability to resolve multiple ITR units, faithfully read long polyA tails, and correctly capture both the 5' and 3' extremities of linear fragments underscores its suitability as a confirmatory technology. This latter feature is particularly important for validating cloning overhangs and polyA boundaries in IVT templates, contexts in which Nanopore sequencing showed recurrent limitations. At the same time, the observed coverage variability in GC-rich regions illustrates the importance of integrating Illumina with Nanopore, whose long-read architecture offers more uniform coverage despite lower per-base accuracy.

This complementarity prompted us to establish a dual-platform strategy: Nanopore sequencing that to enable rapid screening and structural resolution and Illumina sequencing to reach the precision required for final sequence verification. In fact, Illumina accuracy is especially important for rAAV vectors, where ITR mutations can impair packaging efficiency and therapeutic performance, and for IVT mRNA therapies, where polyA tail integrity is essential for transcript stability and translation. By confirming

the correct synthesis of these critical sequences, including fragment termini, Illumina sequencing strengthens both the robustness and safety of next-generation cell and gene therapy products.

In this way, Illumina sequencing emerges not only as a technical complement to Nanopore but as a cornerstone for ensuring the fidelity of rAAV and IVT mRNA products, ultimately safeguarding the efficacy and safety of cell and gene therapies. Building on these complementary strengths, the next step in our work was to evaluate how the two platforms can be jointly integrated into a unified quality control pipeline, including their extension to the direct analysis of in vitro transcribed mRNA.

3.4 IVT mRNA Sequencing with Oxford Nanopore Technologies

3.4.1 Rationale and Experimental Design

As our Biofoundry expanded its activities from recombinant AAV vectors to IVT mRNAs, the need for a sequencing-based quality control strategy became increasingly evident. While methods such as UV spectrophotometry and electrophoretic profiling (e.g., TapeStation) are routinely used to assess RNA concentration, purity, and integrity, they cannot provide nucleotide-level information on the synthesized transcript. This represents a critical limitation in the context of therapeutic mRNA, where sequence accuracy directly affects protein expression, stability, and safety (Pardi et al., 2018; Sahin et al., 2014).

Although the linear DNA template used for IVT can be independently verified by sequencing prior to transcription, this may not guarantee that the resulting RNA product is entirely free from errors or alterations. Incomplete transcription events or sequence misincorporations during the IVT reaction can result in transcripts that diverge from the intended design. For IVT mRNA applications, it is therefore important to verify the sequence at the level of the final mRNA molecule, ensuring that its integrity is preserved throughout the entire synthesis process and that in vivo studies are conducted with constructs that faithfully represent the intended therapeutic design (Daniel et al., 2022; Karikó et al., 2005). Verifying the final RNA is particularly important in preclinical settings,

since sequence deviations could lead to misleading biological readouts, potentially underestimating or overestimating the efficacy and safety of candidate mRNA therapies.

Oxford Nanopore Technologies (ONT) direct RNA sequencing offers a unique opportunity to fill this gap, as it allows the analysis of native RNA molecules without the need for reverse transcription or amplification. By sequencing RNA directly, this approach preserves the physicochemical properties of the transcript and captures characteristic electrical current deviations associated with modified nucleotides. Recent updates in ONT's analysis software (MinKNOW/Dorado) include experimental models for the detection of N^6-methyladenosine (m^6A), suggesting that this modification can be inferred from raw nanopore signal. In contrast, reliable basecalling of pseudouridine is not yet achievable, despite the fact that pseudouridine produces detectable shifts in the raw signal. The ongoing progress in modification-calling algorithms, suggests that future developments may extend accurate detection to a wide range of RNA modifications, including Ψ.

In addition, since direct RNA sequencing enables the acquisition of full-length reads, we expect that this approach might provide information also on the polyA tail length, a feature that plays a pivotal role in transcript stability and translational efficiency (A. Sachs, 1990; Passmore & Coller, 2022). For these reasons, ONT direct RNA sequencing was selected as a candidate technology to extend our sequencing-based QC pipeline from DNA constructs to IVT mRNA products. In practice, the experimental design adopted in this work consisted of three sequential steps: (i) a positive control using a well-characterized yeast transcript, (ii) direct sequencing of unmodified IVT mRNAs, and (iii) evaluation of pseudouridine-modified IVT mRNAs, in order to systematically assess the strengths and limitations of the Nanopore approach.

3.4.2 Positive Control Experiment

To assess the feasibility of Oxford Nanopore direct RNA sequencing in our laboratory, we began with a control experiment using a well-characterized transcript: YHR174W, encoding Enolase II from Saccharomyces cerevisiae. Sequencing was carried out using the Direct RNA Sequencing Kit (SQK-RNA004) on an RNA-dedicated flow cell (FLO-MIN004RA) with a MinION Mk1B device. The run, which lasted approximately 23 minutes, was processed with the Fast basecalling model v3.0.1.

The sequencing generated a total of 29,030 reads, of which 27,956 passed the quality filter, with an estimated N50 read length of 1,335 bases and a mean read quality score of approximately Q12.6. This performance is consistent with the expected range for Oxford Nanopore's direct RNA sequencing and confirms the reliability of the workflow under standard laboratory conditions.

The resulting dataset produced a clean alignment to the reference sequence, confirming the correct execution of the protocol and demonstrating that the MinION device can generate high-quality data for unmodified RNA molecules. As shown in *Figure 3.18*, the full transcript was uniformly covered with an average coverage of approximately 6,000X, and the consensus sequence 100% matched the reference. While individual reads displayed a variability at base-level, typical of ONT sequencing, these errors were resolved in the consensus, which showed no systematic deviations. Importantly, this type of direct RNA analysis is not achievable with Sanger sequencing, which requires reverse transcription into cDNA and is inherently limited in read length. Thus, the experiment underscores ONT's unique capability to directly sequence full-length RNA molecules and verify their integrity at single-nucleotide resolution (Parker et al., 2020).

A noteworthy observation emerged at the 3' end of the alignment: although the digital reference sequence of YHR174W does not contain a polyA tail, the ONT direct RNA sequencing protocol requires one for adapter ligation. Accordingly, a short polyA stretch was consistently detected in the sequenced reads, preceded by a 27 bp sequence. This additional sequence most likely derives from the tailing reaction used to enzymatically attach the polyA tail to the transcript, thereby enabling compatibility with the ONT workflow. This observation highlights the dual role of the polyA tail in ONT sequencing: a biologically relevant feature for transcript stability and, at the same time, a technical prerequisite for successful library preparation.

Overall, the control experiment confirmed that the ONT direct RNA sequencing worklofw a robust protocol that is capable to generate reliable, full-length coverage of canonical RNA molecules. It also established a valuable technical baseline for subsequent experiments on in-house IVT mRNAs, including both unmodified and chemically modified transcripts.

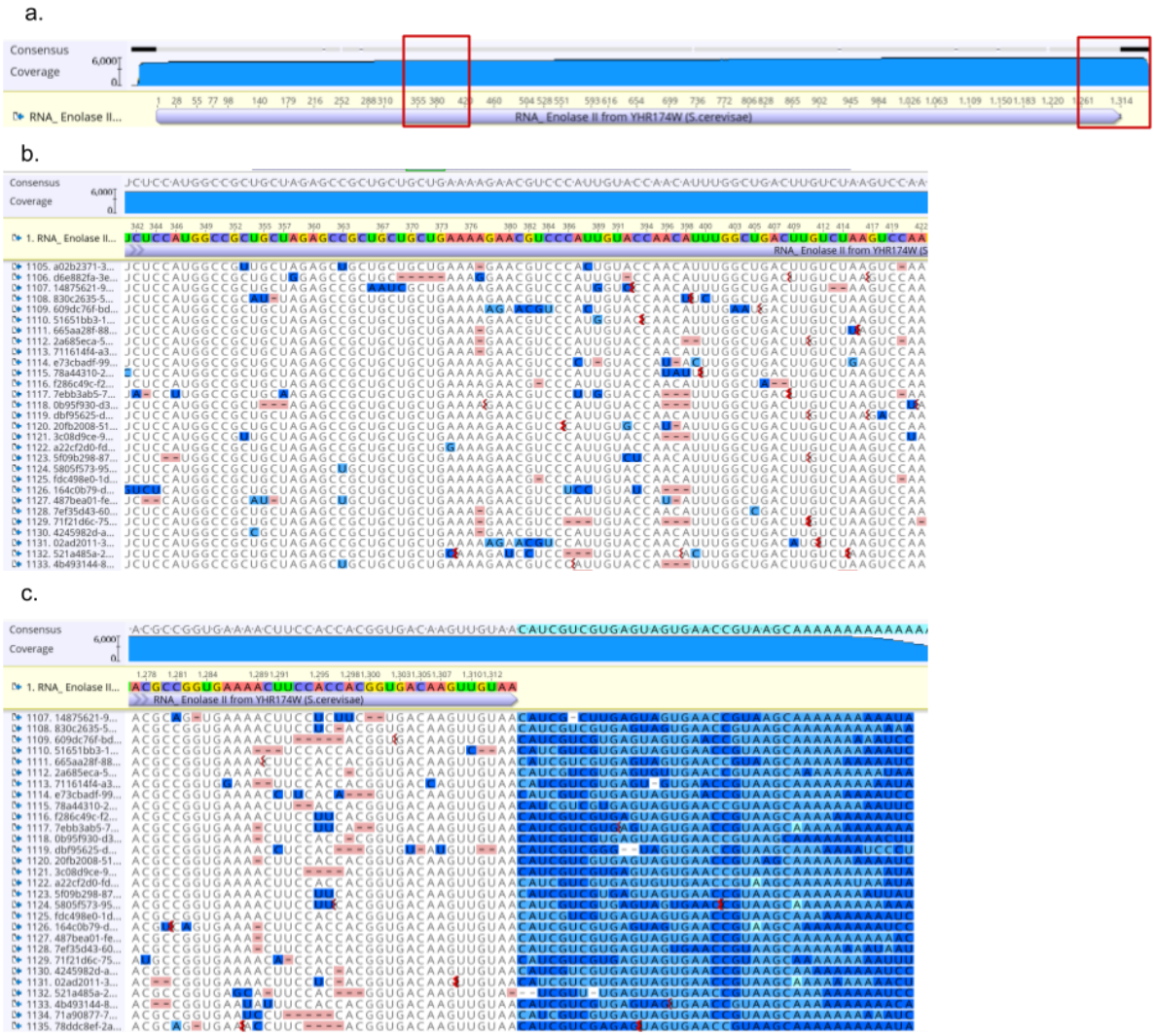

Figure 3.18: Alignment of the YHR174W transcript (Enolase II, S. cerevisiae) obtained with Oxford Nanopore direct RNA sequencing. The upper panel a) shows the full-length alignment displaying uniform coverage and a high-quality consensus sequence (average depth 6,000X; 100% identity to the reference). In the lower panels, b) provides a zoom-in of an internal region, illustrating minor base-level variability across individual reads yet a consistent consensus sequence, while c) highlights the 3' end of the transcript, where a short poly(A) tract is detected, preceded by a 27 bp sequence likely introduced during the tailing reaction used to add the poly(A) extension before library preparation.

3.4.3 IVT mRNA without modifications

As a first real-case application of the protocol, we sequenced a synthetic IVT mRNA produced for a customer. The transcript, approximately 900 nucleotides in length and lacking ribonucleotidic modifications, represents a typical class of unmodified mRNAs commonly used in preclinical studies. Sequencing was performed on a MinION Mk1B device using the Direct RNA Sequencing Kit (SQK-RNA004) and the Fast basecalling model v3.0.1. The run lasted approximately 1 hour and 20 minutes and generated a total of 12,156 reads, of which 12,083 passed the quality filter. The estimated N50 read length

was 1,065 bases, with a mean read quality score of around Q12.7, in line with expected values for Nanopore direct RNA sequencing.

The ONT direct RNA sequencing produced uniform alignment across the full transcript (*Figure 3.19*, panel a) with an average coverage of 11,800X, demonstrating that even relatively short IVT mRNA molecules can be fully captured and reconstructed. The consensus sequence was nearly identical to the reference, with only two local differences detected: a putative uridine deletion at position 268 and an apparent guanine/ambiguous base (G/N) insertion at position 294 (*Figure 3.19*, panel b). It's important to point out that both occurred within short homopolymeric tracts, a context where Nanopore sequencing is prone to indel artefacts. In fact, given their stochastic distribution across reads and the known ONT error profile, these discrepancies are most likely sequencing artefacts rather than genuine mutations. Despite the presence of these regions with artefactual calls, the overall consensus sequence faithfully aligned with the intended transcript reference, demonstrating that ONT direct RNA sequencing can provide rapid and reliable verification of unmodified IVT mRNAs. This represents a clear advantage over conventional quality control methods, which cannot access sequence-level information. At the 3' end, the polyA tail could not be fully reconstructed (*Figure 3.19*, panel c). This reflects a recognized limitation of ONT technology, which frequently underestimates the length and composition of long homopolymeric tracts such as polyA. While this prevents definitive characterization of the tail, computational approaches such as Nanopolish polyA or EPI2ME workflows can be applied to raw signal data to provide approximate estimates. Although these methods are not yet fully standardized, they may complement sequence verification by offering insights into transcript stability and translational potential. In summary, this first application of the pipeline to a customer-derived IVT mRNA demonstrated the feasibility of integrating ONT direct RNA sequencing into manufacturing workflows. Compared to the positive control, the unmodified synthetic transcript achieved an equally clean alignment and a consensus in near-complete concordance with the reference sequence, with only minor deviations in homopolymeric regions that are best interpreted as sequencing artefacts. These results confirm the robustness of the method as a quality control strategy for IVT mRNAs intended for in vivo testing in Cell and Gene Therapy (CGT).

Figure 3.19: Allignment of a direct RNA sequencing run on an unmodified IVT mRNA (around 900 nt). a) The uniform coverage along the full transcript (11,852X) was obtained by the alignment of the reads to the entire reference sequence. b) A closer look revealed two local differences: a putative uridine deletion at position 268 and an apparent guanine/ambiguous base (G/N) insertion at position 294. Both occurred within short homopolymeric tracts, a sequence context where Nanopore frequently introduces indel artefacts. c) A zoom-in of the 3' region showed that the polyA tail could not be completely resolved, consistent with ONT's known difficulty in accurately reconstructing long homopolymeric stretches.

3.4.4 IVT mRNA with Pseudouridine (Ψ)

Pseudouridine (Ψ) is the most abundant naturally occurring RNA modification, predominantly found in non-coding RNAs such as tRNAs, rRNAs, and snRNAs. Although endogenous mRNAs rarely carry Ψ under normal physiological conditions,

pseudouridylation can occur at specific sites in response to stress (Carlile et al., 2014). In the context of synthetic biology and therapeutic development, however, Ψ is deliberately introduced into IVT mRNAs by substituting uridine triphosphate with pseudouridine triphosphate (rΨTP) during transcription. This modification has become a cornerstone of therapeutic mRNA design because it improves transcript stability and translational efficiency while simultaneously reducing activation of innate immune pathways (Karikó et al., 2005). For this reason, our customers systematically request rΨTP to be incorporated during IVT reactions in our Biofoundry, generating capped and polyadenylated mRNA molecules intended for preclinical applications.

With Oxford Nanopore Technologies (ONT) Direct RNA sequencing, RNA molecules can be analyzed in their native form, eliminating the need for reverse transcription or amplification. This makes it an attractive option for quality control of IVT mRNAs, including those containing chemical modifications such as Ψ. We therefore sought to test how the platform performs when sequencing pseudouridine-modified transcripts.

The experiment was performed using the Direct RNA Sequencing Kit (SQK-RNA004) on a MinION Mk1B device with the Fast basecalling model v3.0.1. The run lasted approximately 1 hour and 24 minutes and generated a total of 6,004 reads, of which 5,931 passed the quality filter. The estimated N50 read length was 1,082 bases, with a mean read quality score of approximately Q12.7, consistent with the expected performance of Nanopore direct RNA sequencing.

The sequencing produced full-length reads, however, when aligned to the reference sequence, the resulting alignment was noticeably “dirty”, with clusters of mismatches and undefined bases. As shown in *Figure 3.20*, regions without pseudouridine were correctly basecalled, while stretches containing Ψ produced locally noisy alignments that obscured the correct sequence.

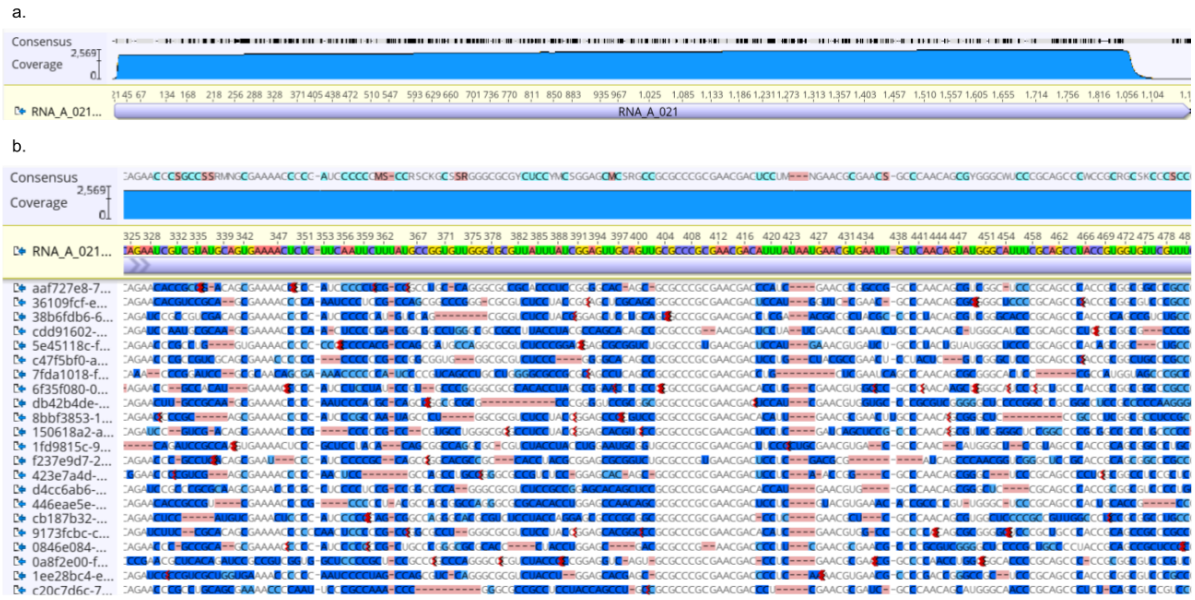

Figure 3.20: Alignment of direct RNA sequencing reads from an IVT transcript containing pseudouridine (Ψ). Panel (a) shows the full-length alignment, confirming that ONT Direct RNA sequencing can generate complete transcript coverage. Panel (b) provides a zoomed-in detail of a Ψ-containing region, where locally noisy and error-prone alignments emerge. While regions without Ψ were basecalled accurately, stretches containing Ψ produced clusters of mismatches and undefined positions. This reflects the effect of Ψ-induced signal perturbations across overlapping k-mers, as described by Wang et al. (2024) and Makhamreh et al. (2024)

To understand this outcome, it is important to consider how ONT basecalling operates. Each measurement of ionic current corresponds not to a single nucleotide but to overlapping k-mers, typically spanning five nucleotides. When a pseudouridine is present within a k-mer, it perturbs the signal of the entire unit, causing errors not only at the modified site but also in the neighboring nucleotides within the same window. Current basecalling models, which are not trained to interpret these altered signals, therefore introduce clusters of miscalls (Z. Wang et al., 2024). This principle is illustrated in *Figure 3.21*: in canonical transcripts, k-mers are consistently recognized and correctly assigned by the basecaller (green ticks), whereas in Ψ-modified transcripts, the altered current leads to systematic errors across the affected k-mer (red crosses).

Figure 3.21: Schematic representation of the effect of pseudouridine on ONT basecalling. a) in canonical transcripts (rUTP), each overlapping 5-mer is consistently recognized (green ticks), yielding accurate basecalls. b) when Ψ is incorporated (rΨTP), the altered ionic current perturbs the entire 5-mer (red crosses), leading to miscalls not only at the modified position but also in adjacent nucleotides. This mechanistic explanation, consistent with Wang et al. (2024), accounts for the "dirty" alignments observed in Ψ-rich regions (see *Figure 13.8*).

Our results are in line with previous studies suggesting that pseudouridine can generate signal deviations in nanopore sequencing, which may extend beyond the modified base and affect the surrounding sequence context. This makes accurate quantification particularly difficult without dedicated controls (Makhamreh et al., 2024). As a result, Ψ-containing regions appear as clusters of erroneous bases rather than being directly identified (Begik et al., 2021).

However, it is important to underline that recent advances in RNA modification analysis with nanopore sequencing have mainly focused on N^6-methyladenosine (m^6A), which is currently the best-characterized modification for ONT data. Several computational tools, such as xPore, m6Anet and MINES, can infer m^6A from direct RNA signals by leveraging large training datasets and established sequence motifs. In contrast, equivalent methods for pseudouridine remain limited, despite the fact that Ψ clearly induces detectable current shifts. This imbalance reflects a broader gap, not in the sequencing chemistry, but in the computational technology: while ONT chemistry is sensitive to multiple modifications, robust basecalling or prediction models currently exist only for a subset of them. Therefore, there is substantial room for improvement in developing Ψ-focused algorithms, which would greatly enhance the interpretability of direct RNA sequencing for therapeutically relevant transcripts.

In conclusion, ONT Direct RNA sequencing appears applicable to IVT mRNAs containing pseudouridine, providing full-length reads and supporting the accuracy of unmodified regions of the transcript. However, the reliable detection of Ψ, as well as neighboring nucleotides within the same k-mer window, still represents a challenge for current basecalling models. The observed “dirty” alignments are not due to experimental limitations but reflect a technological constraint: while the nanopore detects the altered current, the computational layer is unable to interpret it correctly. With the development of modification-aware basecallers, the direct detection of Ψ, along with improved accuracy for surrounding bases, may become achievable, potentially increasing the value of nanopore sequencing for therapeutic mRNA quality control.

3.4.5 Discussion of Strengths and Limitations

The evaluation of ONT Direct RNA sequencing across three experimental conditions, a positive control transcript, Biofoundry-generated IVT mRNAs without modifications, and IVT mRNAs containing pseudouridine, highlighted both the robustness of the technology and its current constraints.

The positive control experiment confirmed that the protocol was correctly executed, providing a clean alignment and a full-length coverage. This indicated that ONT can generate accurate consensus sequences from the sequencing of unmodified RNA molecules, and it provided an essential baseline for the subsequent analyses.

Sequencing of IVT mRNAs with unmodified ribonucleotides further indicated that the method can be applied to the verification of synthetic transcripts, as full-length reads aligned correctly to the reference. At the same time, the experiments confirmed ONT’s known weakness in homopolymeric regions, particularly within the polyA tail, which could not be reliably resolved. Since polyA length is biologically relevant for mRNA stability and translation, this remains an important limitation in the context of IVT mRNA quality control.

The sequencing analysis obtained from IVT mRNA with pseudouridine highlighted the most pronounced shortcomings: while the sequencing produced full-length reads, the regions containing pseudouridine displayed noisy alignments with clusters of

mismatches and undefined bases. This outcome reflects the principle of k-mer–based basecalling: each ionic current signal corresponds to overlapping nucleotide contexts, so the presence of pseudouridine perturbs not only the modified site but also the neighbouring nucleotides. Because current basecalling models are not trained to interpret these altered signals, these perturbations manifest in systematic local miscalls and reduced sequence reliability in modification-rich regions. Consistent with this, Gunter et al., (2023) reported a distinctive pattern of Ψ→C substitution artefacts in nanopore datasets from pseudouridine-containing mRNA constructs, indicating that these mismatches arise from incomplete modelling of modified nucleotide signals rather than true sequence changes. In line with their observations, the data presented in this thesis show that ONT performs robustly on unmodified IVT mRNA, whereas pseudouridine-modified transcripts require cautious interpretation due to persistent miscalling biases in current basecalling algorithms.

Another important limitation for our Biofoundry is that ONT Direct RNA sequencing is not a high-throughput method. Unlike DNA sequencing, where up to 96 barcoded plasmid or linear DNA samples can be processed in a single run, ONT has not released a barcoding kit for RNA. This means that only one RNA sample can be sequenced per run, which significantly restricts throughput. Moreover, this constraint increases the cost of RNA sequencing: a single run costs approximately € 270 per sample, making the approach less scalable for routine quality control when multiple constructs must be analyzed in parallel.

Taken together, these results demonstrate that ONT Direct RNA sequencing is effective for rapid verification of unmodified IVT transcripts, where it provides full-length coverage and reliable alignment. At the same time, its limitations are clear: inaccurate resolution of polyA tails, the inability to correctly identify pseudouridine and adjacent nucleotides within affected k-mers, and the absence of multiplexing capacity, which limits throughput and increases per-sample cost. These shortcomings stem from the state of the technology rather than from experimental execution, and they currently restrict the extent to which ONT alone can provide comprehensive quality control of therapeutic mRNAs.

3.4.6 Implications and Future Perspectives

The results obtained in this work highlight both the promise and the limitations of ONT Direct RNA sequencing as a tool for IVT mRNA quality control. The most immediate

implication is that the technology can be applied with confidence for rapid verification of unmodified transcripts, providing full-length coverage without the need for reverse transcription or amplification. This represents a clear advantage in Biofoundry workflows, where turnaround time is critical and multiple constructs often need to be screened in parallel.

At the same time, several limitations reduce the broader applicability of ONT in therapeutic mRNA pipelines. A particularly important one for industrial use is the lack of RNA sample multiplexing: while DNA constructs can be barcoded and pooled (up to 96 per flow cell), RNA sequencing can only be performed with one sample per flow cell. This restriction lowers the throughput and increases the costs, with each run amounting to about € 270 per sample. For our Biofoundry that needs to process multiple constructs, this cost barrier currently limits the scalability of ONT RNA sequencing compared to DNA sequencing.

Moreover, biological features, essential for therapeutic efficacy, remain unresolved by ONT: polyA tails, which influence mRNA stability and translational efficiency, could not be reliably quantified; the regions containing pseudouridine, despite being readily detected as a perturbation of the ionic current, could not be correctly basecalled, leading to noisy alignments both at the modified sites and at adjacent nucleotides. These findings underline that ONT, in its present state, provides only partial information: it can flag problematic regions but cannot yet deliver definitive confirmation of their sequence or modification status.

Looking forward, these limitations point to two complementary paths. On the one hand, Illumina sequencing, although indirect and reliant on reverse transcription, offers the high per-base accuracy and homopolymer resolution required to confirm coding sequence and polyA tails. Its integration into Biofoundry workflows could therefore provide the definitive verification that ONT currently lacks, albeit at the cost of discarding information on RNA modifications. On the other hand, advances in ONT basecalling are expected to improve direct RNA sequencing itself. Recent studies have demonstrated that machine learning approaches and the use of synthetic training datasets can enhance modification-aware basecalling, opening the possibility of directly identifying

pseudouridine and other chemical modifications in native transcripts (Makhamreh et al., 2024; Z. Wang et al., 2024).

In conclusion, ONT Direct RNA sequencing presently serves as a rapid but partial QC tool for IVT mRNAs: highly valuable for initial screening of unmodified constructs, but insufficient for comprehensive analysis of modified transcripts and polyA features. Future perspectives involve both the technological evolution of ONT and the strategic integration of complementary sequencing platforms such as Illumina, ultimately aiming to provide Biofoundries with workflows that are not only fast and scalable but also accurate and modification-aware, essential qualities for supporting the development of mRNA-based therapeutics in Cell and Gene Therapy.

4 Conclusions

Traditionally, the identification and verification of fully correct rAAV plasmids or DNA templates for IVT mRNA can require several months to over a year, as repeated cloning and Sanger sequencing rounds are often necessary to resolve ITRs, poly(A) tails, and GC-rich regions. By implementing an integrated Oxford Nanopore–Illumina sequencing pipeline, our Biofoundry can complete this step in approximately one month, thus drastically reducing both time and cost before in vivo testing and clinical translation. This acceleration directly addresses one of the main bottlenecks in Cell and Gene Therapy (CGT) development, where delays in obtaining sequence-perfect DNA and RNA molecules translate into significant setbacks in therapeutic discovery pipelines.

The first part of this thesis demonstrated that Oxford Nanopore Technologies (ONT) sequencing can overcome the intrinsic limitations of Sanger sequencing by spanning entire plasmids, including helper and packaging constructs up to 30 kb with inverted terminal repeats (ITRs), repetitive regions, and high GC content. By adopting barcoding strategies, ONT sequencing enabled the simultaneous screening of up to 96 colonies in a single run, reducing both costs and turnaround time. However, a recurrent limitation emerged: the per-base accuracy was low, with a median Q-score of Q12, with systematic errors particularly in long homopolymeric regions. For this reason, we decided to evaluate Illumina sequencing, a technology known for its high accuracy, to determine whether it could complement ONT and strengthen our QC pipeline.

In the second part of this work, Illumina sequencing was evaluated not as a replacement for ONT, but as a complementary technology able to resolve sequence regions where ONT encountered some limitations. The results confirmed that, when applied to DNA molecules, Illumina delivers precise base-level resolution in critical motifs, including ITRs, GC-rich regions, and long homopolymeric tracts such as the 120A poly(A) tail encoded in our proprietary IVT mRNA backbone. However, Illumina also comes with some drawbacks: longer turnaround times (18 hours per run compared with only a few hours for ONT), higher costs unless runs are fully multiplexed, and reduced coverage in regions of very high GC content. These factors limit its utility for rapid screening but reinforce its value as a confirmatory platform for definitive sequence verification.

The third part extended sequencing-based QC to RNA molecules. ONT Direct RNA sequencing successfully captured full-length unmodified IVT mRNAs, providing rapid transcript-level verification beyond conventional spectrophotometric and electrophoretic assays. However, when applied to pseudouridine-modified transcripts, ONT showed significant shortcomings. The altered ionic current generated by the pseudouridine interfered with k-mer based basecalling, which caused miscalls not only at the modified sites but also at the neighbouring bases. As a result, the alignments appeared noisy and the full sequence was harder to resolve. The combination of ONT's lack of RNA sample multiplexing and its continued difficulty in accurately reconstructing poly(A) tails, currently prevent its application to modified IVT mRNAs, although the method remains a reliable QC tool for unmodified transcripts.

Taken together, the results of this thesis show that no single sequencing technology alone can provide comprehensive QC for rAAV plasmids and IVT mRNAs. Instead, their complementarity supports a robust integrated strategy: ONT for rapid, high-throughput screening and direct RNA sequencing, Illumina for definitive base-level verification of complex DNA motifs, and Sanger sequencing reserved for targeted confirmatory checks of simpler constructs. After the validation on thousands of DNA samples and multiple IVT mRNAs, this combined pipeline offers a scalable and cost-effective sequencing solution tailored to the needs of Cell and Gene Therapy companies.

Beyond the technical outcomes, this thesis underscores the industrial and translational impact of sequencing-based QC. By compressing a process that typically takes several months into approximately one month, our Biofoundry directly accelerates preclinical discovery, reduces the risk of costly failures, and enables faster access to high-quality molecules. For CGT companies, this translates into shorter development cycles, improved reliability, and a more seamless transition from design to in vivo testing and eventually to clinical evaluation.

4.1 Future Perspectives

The next phase of this work will focus on further automation and scalability of the QC pipeline. With the recent acquisition of a TECAN Fluent liquid handler, the library preparation for both ONT and Illumina sequencing can be fully automated. This will ensure reproducibility, minimize operator variability, and will significantly increase the throughput. In parallel, automated bioinformatic pipelines should be developed to streamline data processing, reducing manual intervention and enabling standardized reporting suitable for industrial adoption.

Another important avenue will be the development of modification-aware basecalling algorithms, capable of correctly interpreting altered ionic current signals caused by pseudouridine and other ribonucleoside modifications. Achieving reliable basecalling of chemically modified IVT mRNAs would extend direct RNA sequencing to clinically relevant transcripts and overcome one of the main technological barriers currently limiting QC for therapeutic mRNAs.

Finally, the full integration of Illumina sequencing into our Biofoundry QC workflow will consolidate its role as a confirmatory platform, enabling definitive sequence verification of constructs containing ITRs, GC-rich regions, and long poly(A) tracts. Along with recent developments, this will establish an industrially robust, high-throughput sequencing pipeline that is fast, accurate, and cost effective. Such a system will not only form a cornerstone of our Biofoundry 2.0 but will also position it at the cutting edge of Cell and Gene Therapy innovation, providing an essential platform for the next generation of therapeutic products.

5 Acknowledgments

I would like to sincerely thank my supervisor, Prof. Flavio Rizzolio, for his guidance and encouragement throughout the course the past three years. I owe my deepest gratitude to my co-supervisor, Dr. Davide De Lucrezia, without whose support this PhD journey would not have taken place. Thank you, Davide, for believing in me from the beginning and for continually supporting my personal and professional growth. For three years, you've been the captain, patiently and wisely leading the Officinae Bio crew, and you remain our guiding light across every sea. I am extremely grateful also to my thesis reviewers, Dr. Matthew Perkett and Prof. Marco Scocchi: thank you for accepting to dedicate some of your precious time to review my thesis. Your corrections and suggestions will be invaluable not only for the improvement of this work, but also for my ongoing personal and academic growth.

A heartfelt thank you goes to my colleague Sota, whose invaluable guidance has profoundly influenced my approach to science. Thank you, Sota, for always sharing your knowledge with me unconditionally and for always encouraging me to give my best in every little thing I do. I am also sincerely grateful to the whole Officinae Bio crew for being the funniest weird nerds I could ever wished to work with, and for continuously fostering my growth and confidence. Thank you because, with your craziness, cheerfulness and also infinite intelligence, you always manage to create a joyful and highly valuable working environment. A special thanks also goes to Vicky for her profound human, professional, and artistic sensitivity. Thank you for sharing with me your boundless passion for The Lord of the Rings, gardening, and everything food related.

Special thanks to my friends, Arianna, Valentina, Clarissa and Roberta, who, despite the distances that separate us since we finished university, have been a constant source of encouragement and joy. Thank you for always being by my side and sharing the most special moments together. The memories we have created together, and those yet to come, are priceless to me.

I am forever grateful to my parents for their endless love and encouragement. Thank you mamma e papà for all the sacrifices you made to guarantee me to follow always my

aspirations. Your unconditional support in everything I do is the source of my strength. A warm thank you goes to Angelo and Serena, my younger brothers who have always been an important source of strength and joy in my life. Thank you for being my first supporters since you were kids. I am extremely grateful to life, and to mom and dad, for having you in my life.

Finally, but above all, the most heartfelt gratitude and love are reserved for my husband, Daniele. Thank you for always being my guiding light throughout these years, for always believing in me, and for supporting me through the most difficult times with your infinite patience and joy. I could never find enough space to write down everything I want to thank you for, after 15 years together.

6 Bibliography

A. Sachs. (1990). The role of poly(A) in the translation and stability of mRNA. *Current Opinion in Cell Biology*, *2*, 1092–1098. https://doi.org/10.1016/0955-0674(90)90161-7

Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., & Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, *12*(2). https://doi.org/10.1186/gb-2011-12-2-r18

Baek, R., Coughlan, K., Jiang, L., Liang, M., Ci, L., Singh, H., Zhang, H., Kaushal, N., Rajlic, I. L., Van, L., Dimen, R., Cavedon, A., Yin, L., Rice, L., Frassetto, A., Guey, L., Finn, P., & Martini, P. G. V. (2024). Characterizing the mechanism of action for mRNA therapeutics for the treatment of propionic acidemia, methylmalonic acidemia, and phenylketonuria. *Nature Communications*, *15*(1). https://doi.org/10.1038/s41467-024-47460-9

Barton E. Slatko, Andrew F. Gardner, & Frederick M. Ausubel. (2018). Overview of Next Generation Sequencing Technologies. *Physiology & Behavior*, *176*(12), 139–148. https://doi.org/10.1002/cpmb.59

Begik, O., Lucas, M. C., Pryszcz, L. P., Ramirez, J. M., Medina, R., Milenkovic, I., Cruciani, S., Liu, H., Vieira, H. G. S., Sas-Chen, A., Mattick, J. S., Schwartz, S., & Novoa, E. M. (2021). Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nature Biotechnology*, *39*(10), 1278–1291. https://doi.org/10.1038/s41587-021-00915-6

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, *40*(10). https://doi.org/10.1093/nar/gks001

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517

Bowater, R. P., Bohálová, N., & Brázda, V. (2022). Interaction of Proteins with Inverted Repeats and Cruciform Structures in Nucleic Acids. In *International Journal of Molecular Sciences* (Vol. 23, Issue 11). MDPI. https://doi.org/10.3390/ijms23116171

Byrne, B. J., Flanigan, K. M., Matesanz, S. E., Finkel, R. S., Waldrop, M. A., D'Ambrosio, E. S., Johnson, N. E., Smith, B. K., Bönnemann, C., Carrig, S., Rossano, J. W., Greenberg, B., Lalaguna, L., Lara-Pezzi, E., Subramony, S., Corti, M., Mercado-Rodriguez, C., Leon-Astudillo, C., Ahrens-Nicklas, R., … George, L. A. (2025). Current clinical applications of

AAV-mediated gene therapy. In *Molecular Therapy* (Vol. 33, Issue 6, pp. 2479–2516). Cell Press. https://doi.org/10.1016/j.ymthe.2025.04.045

Carlile, T. M., Rojas-Duran, M. F., Zinshteyn, B., Shin, H., Bartoli, K. M., & Gilbert, W. V. (2014). Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature*, *515*(7525), 143–146. https://doi.org/10.1038/nature13802

Chen, Y., Hu, S., Lee, W., Walsh, N., Iozza, K., Huang, N., Preston, G., Drouin, L. M., Jia, N., Deng, J., Hebben, M., & Liao, J. (2024). A Comprehensive Study of the Effects by Sequence Truncation within Inverted Terminal Repeats (ITRs) on the Productivity, Genome Packaging, and Potency of AAV Vectors. *Microorganisms*, *12*(2). https://doi.org/10.3390/microorganisms12020310

Cheng, C., & Xiao, P. (2022). Evaluation of the correctable decoding sequencing as a new powerful strategy for DNA sequencing. *Life Science Alliance*, *5*(8). https://doi.org/10.26508/lsa.202101294

Daniel, S., Kis, Z., Kontoravdi, C., & Shah, N. (2022). Quality by Design for enabling RNA platform production processes. In *Trends in Biotechnology* (Vol. 40, Issue 10, pp. 1213–1228). Elsevier Ltd. https://doi.org/10.1016/j.tibtech.2022.03.012

Delahaye, C., & Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. *PLoS ONE*, *16*(10 October). https://doi.org/10.1371/journal.pone.0257521

Ersing, I., Rego, M., Wang, C., Zhang, Y., Harten DeMaio, K., Petrozzi, M., Fava, A., Clouse, G., Patrick, M., Guerin, K., & Fan, M. (2023). Quality Control for Adeno-Associated Viral Vector Production. In *Neuromethods* (Vol. 195, pp. 77–101). Humana Press Inc. https://doi.org/10.1007/978-1-0716-2918-5_5

F. Sanger, A.R. Coulson, G.F. Hong, D.F. Hill, & G.B. Petersen. (1982). Nucleotide Sequence of Bacteriophage λ DNA. In *J. Mol. Biol* (Vol. 162). https://doi.org/10.1016/0022-2836(82)90546-0

Fang, E., Liu, X., Li, M., Zhang, Z., Song, L., Zhu, B., Wu, X., Liu, J., Zhao, D., & Li, Y. (2022). Advances in COVID-19 mRNA vaccine development. In *Signal Transduction and Targeted Therapy* (Vol. 7, Issue 1). Springer Nature. https://doi.org/10.1038/s41392-022-00950-y

Granados-Riveron, J. T., & Aquino-Jarquin, G. (2021). Engineering of the current nucleoside-modified mRNA-LNP vaccines against SARS-CoV-2. In *Biomedicine and Pharmacotherapy* (Vol. 142). Elsevier Masson s.r.l. https://doi.org/10.1016/j.biopha.2021.111953

Grieger, J. C., & Samulski, R. J. (2005). Packaging Capacity of Adeno-Associated Virus Serotypes: Impact of Larger Genomes on Infectivity and Postentry Steps. *Journal of Virology*, *79*(15), 9933–9944. https://doi.org/10.1128/jvi.79.15.9933-9944.2005

Gunter, H. M., Idrisoglu, S., Singh, S., Han, D. J., Ariens, E., Peters, J. R., Wong, T., Cheetham, S. W., Xu, J., Rai, S. K., Feldman, R., Herbert, A., Marcellin, E., Tropee, R., Munro, T., & Mercer, T. R. (2023). mRNA vaccine quality analysis using RNA sequencing. *Nature Communications*, *14*(1). https://doi.org/10.1038/s41467-023-41354-y

Heba H. Mostafa. (2024). An evolution of Nanopore next-generation sequencing technology: implications for medical microbiology and public health. *Journal of Clinical Microbiology*, *62*(5). https://doi.org/10.1128/jcm.01576-23

Hillson, N., Caddick, M., Cai, Y., Carrasco, J. A., Chang, M. W., Curach, N. C., Bell, D. J., Le Feuvre, R., Friedman, D. C., Fu, X., Gold, N. D., Herrgård, M. J., Holowko, M. B., Johnson, J. R., Johnson, R. A., Keasling, J. D., Kitney, R. I., Kondo, A., Liu, C., Freemont, P. S. (2019). Building a global alliance of biofoundries. In *Nature Communications* (Vol. 10, Issue 1). Nature Publishing Group. https://doi.org/10.1038/s41467-019-10079-2

Hou, X., Zaks, T., Langer, R., & Dong, Y. (2021). Lipid nanoparticles for mRNA delivery. In *Nature Reviews Materials* (Vol. 6, Issue 12, pp. 1078–1094). Nature Research. https://doi.org/10.1038/s41578-021-00358-0

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 1–11. https://doi.org/10.1186/s13059-016-1103-0

Jan Kieleczawa. (2006). Fundamentals of Sequencing of Difficult Templates -- An Overview. *Journal of Biomolecular Techniques*, *17*, 207–217.

Karikó, K., Buckstein, M., Ni, H., & Weissman, D. (2005). Suppression of RNA recognition by Toll-like receptors: The impact of nucleoside modification and the evolutionary origin of RNA. *Immunity*, *23*(2), 165–175. https://doi.org/10.1016/j.immuni.2005.06.008

Khalil, A. S., & Collins, J. J. (2010). Synthetic biology: Applications come of age. *Nature Reviews Genetics*, *11*(5), 367–379. https://doi.org/10.1038/nrg2775

Kontogiannis, T., Braybrook, J., McElroy, C., Foy, C., Whale, A. S., Quaglia, M., & Smales, C. M. (2024). Characterization of AAV vectors: A review of analytical techniques and critical quality attributes. In *Molecular Therapy Methods and Clinical Development* (Vol. 32, Issue 3). Cell Press. https://doi.org/10.1016/j.omtm.2024.101309

Kovaka, S., Hook, P. W., Jenike, K. M., Shivakumar, V., Morina, L. B., Razaghi, R., Timp, W., & Schatz, M. C. (2025). Uncalled4 improves nanopore DNA and RNA modification detection via fast and accurate signal alignment. *Nature Methods*, *22*(4), 681–691. https://doi.org/10.1038/s41592-025-02631-4

Lee, V. V., Judd, L. M., Jex, A. R., Holt, K. E., Tonkin, C. J., & Ralph, S. A. (2021). *Direct Nanopore Sequencing of mRNA Reveals Landscape of Transcript Isoforms in Apicomplexan Parasites*. https://doi.org/10

Makhamreh, A., Tavakoli, S., Fallahi, A., Kang, X., Gamper, H., Nabizadehmashhadtoroghi, M., Jain, M., Hou, Y. M., Rouhanifard, S. H., & Wanunu, M. (2024). Nanopore signal deviations from pseudouridine modifications in RNA are sequence-specific: quantification requires dedicated synthetic controls. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-72994-9

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA (DNA chenistry/dimethyl sulfate cleavage/hydrazine/piperidine). In *Biochemistry* (Vol. 74, Issue 2). https://www.pnas.org

Mihovilovic, M., Hagerty, N., & Stein, D. (2012). *The Statistics of DNA Capture by a Solid-State Nanopore*. https://doi.org/10.1103/PhysRevLett.110.028102

Mingozzi, F., & High, K. A. (2011). Therapeutic in vivo gene transfer for genetic disease using AAV: progress and challenges. In *Nature Reviews Genetics* (Vol. 12). http://www.nature.com/nrg/journal/

Mohammadi, M. M., & Bavi, O. (2022). DNA sequencing: an overview of solid-state and biological nanopore-based methods. *Biophysical Reviews*, *14*(1), 99–110. https://doi.org/10.1007/s12551-021-00857-y

Morais, P., Adachi, H., & Yu, Y. T. (2021). The Critical Contribution of Pseudouridine to mRNA COVID-19 Vaccines. In *Frontiers in Cell and Developmental Biology* (Vol. 9). Frontiers Media S.A. https://doi.org/10.3389/fcell.2021.789427

Mroske, C., Rivera, H., Ul-Hasan, T., Chatterjee, S., & Wong, K. K. (2012). A capillary electrophoresis sequencing method for the identification of mutations in the inverted terminal repeats of adeno-associated virus. *Human Gene Therapy Methods*, *23*(2), 128–136. https://doi.org/10.1089/hgtb.2011.231

Murlidharan, G., Samulski, R. J., & Asokan, A. (2014). Biology of adeno-associated viral vectors in the central nervous system. In *Frontiers in Molecular Neuroscience* (Vol. 7). Frontiers Research Foundation. https://doi.org/10.3389/fnmol.2014.00076

Naso, M. F., Tomkowicz, B., Perry, W. L., & Strohl, W. R. (2017). Adeno-Associated Virus (AAV) as a Vector for Gene Therapy. *BioDrugs*, *31*(4), 317–334. https://doi.org/10.1007/s40259-017-0234-5

Pardi, N., Hogan, M. J., Porter, F. W., & Weissman, D. (2018). mRNA vaccines-a new era in vaccinology. In *Nature Reviews Drug Discovery* (Vol. 17, Issue 4, pp. 261–279). Nature Publishing Group. https://doi.org/10.1038/nrd.2017.243

Parker, M. T., Knop, K., Sherwood, A. V., Schurch, N. J., Mackinnon, K., Gould, P. D., Hall, A. J. W., Barton, G. J., & Simpson, G. G. (2020). Nanopore direct RNA sequencing maps the complexity of arabidopsis mRNA processing and m6A modification. *ELife*, *9*. https://doi.org/10.7554/eLife.49658

Passmore, L. A., & Coller, J. (2022). Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression. In *Nature Reviews Molecular Cell Biology* (Vol. 23, Issue 2, pp. 93–106). Nature Research. https://doi.org/10.1038/s41580-021-00417-y

Pereira, F., Carneiro, J., & van Asch, B. (2010). A Guide for Mitochondrial DNA Analysis in Non-Human Forensic Investigations. *The Open Forensic Science Journal*, *3*(2), 33–44. https://doi.org/10.2174/1874402801003020033

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Pérez Marc, G., Moreira, E. D., Zerbini, C., Bailey, R., Swanson, K. A., Roychoudhury, S., Koury, K., Li, P., Kalina, W. V., Cooper, D., Frenck, R. W., Hammitt, L. L., … Gruber, W. C. (2020). Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *New England Journal of Medicine*, *383*(27), 2603–2615. https://doi.org/10.1056/nejmoa2034577

Rang, F. J., Kloosterman, W. P., & de Ridder, J. (2018). From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. In *Genome Biology* (Vol. 19, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-018-1462-9

Ribarska, T., Bjørnstad, P. M., Sundaram, A. Y. M., & Gilfillan, G. D. (2022). Optimization of enzymatic fragmentation is crucial to maximize genome coverage: a comparison of library preparation methods for Illumina sequencing. *BMC Genomics*, *23*(1). https://doi.org/10.1186/s12864-022-08316-y

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., & Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, *14*(5). https://doi.org/10.1186/gb-2013-14-5-r51

Sahin, U., Karikó, K., & Türeci, Ö. (2014). mRNA-based therapeutics-developing a new class of drugs. In *Nature Reviews Drug Discovery* (Vol. 13, Issue 10, pp. 759–780). Nature Publishing Group. https://doi.org/10.1038/nrd4278

Samulski, R. J., & Muzyczka, N. (2014). AAV-mediated gene therapy for research and therapeutic purposes. *Annual Review of Virology*, *1*(1), 427–451. https://doi.org/10.1146/annurev-virology-031413-085355

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). *DNA sequencing with chain-terminating inhibitors (DNA polymerase/nucleotide sequences/bacteriophage 4X174)* (Vol. 74, Issue 12). https://www.pnas.org

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B. H., & Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, *321*, 674–679. https://doi.org/https://doi.org/10.1038/321674a0

Srivastava, A. (2016). In vivo tissue-tropism of adeno-associated viral vectors. In *Current Opinion in Virology* (Vol. 21, pp. 75–80). Elsevier B.V. https://doi.org/10.1016/j.coviro.2016.08.003

Wang, J. H., Gessler, D. J., Zhan, W., Gallagher, T. L., & Gao, G. (2024). Adeno-associated virus as a delivery vector for gene therapy of human diseases. In *Signal Transduction and Targeted Therapy* (Vol. 9, Issue 1). Springer Nature. https://doi.org/10.1038/s41392-024-01780-w

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). *Nanopore sequencing technology, bioinformatics and applications* (Vol. 39, Issue 11). https://doi.org/10.1038/s41587-021-01108-x.Nanopore

Wang, Z., Fang, Y., Liu, Z., Hao, N., Zhang, H. H., Sun, X., Que, J., & Ding, H. (2024). Adapting nanopore sequencing basecalling models for modification detection via incremental learning and anomaly detection. *Nature Communications*, *15*(1). https://doi.org/10.1038/s41467-024-51639-5

Weber, J. S., Adnan Khattak, M., Carlino, M. S., Meniawy, T., Taylor, M. H., Ansstas, G., Kim, K. B., McKean, M., Sullivan, R. J., Faries, M. B., Tran, T., Lance Cowey, C., Medina, T., Margaret Segar, J., Atkinson, V., Thomas Gibney, G., Luke, J. J., Iannotti Buchbinder, E., Long, G. V, & Meehan, R. S. (2025). *Individualized neoantigen therapy mRNA-4157 (V940) plus pembrolizumab in resected melanoma: 3-year update from the mRNA-4157-P201 (KEYNOTE-942) trial*.

Zhang, J., Yu, X., Chrzanowski, M., Tian, J., Pouchnik, D., Guo, P., Herzog, R. W., & Xiao, W. (2024). Thorough molecular configuration analysis of noncanonical AAV genomes in AAV vector preparations. *Molecular Therapy Methods and Clinical Development*, *32*(1). https://doi.org/10.1016/j.omtm.2024.101215

Zhang, T., Li, H., Jiang, M., Hou, H., Gao, Y., Li, Y., Wang, F., Wang, J., Peng, K., & Liu, Y. X. (2024). Nanopore sequencing: flourishing in its teenage years. In *Journal of Genetics and Genomics* (Vol. 51, Issue 12, pp. 1361–1374). Institute of Genetics and Developmental Biology. https://doi.org/10.1016/j.jgg.2024.09.007