UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Evaluation and Application of Tools for mRNA Isoform Detection Using Nanopore Sequencing Data

by

Keenan Ikking

(2174355)

Dissertation submitted in fulfilment of the requirements for the degree

Master of Science

in Molecular and Cell Biology

in the Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa

Supervisor: Dr Nikki Gentle

June 2025

Declaration

I, Keenan Ikking (2174355), am a student registered for the degree of Master of Science in the academic year 2025.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their
- permission and/or without acknowledging the original source) is wrong.

- I confirm that the work submitted for assessment for the above degree is my own unaided work except where explicitly indicated otherwise and acknowledged.

- I have not submitted this work before for any other degree or examination at this
- or any other University.

- The information used in the Dissertation HAS NOT been obtained by me while employed by, or working under the aegis of, any person or organisation other than the University.

- I have followed the required conventions in referencing the thoughts and ideas of others.

- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Name: Keenan Ikking Date: 06 June 2025

Abstract

Alternative splicing contributes to RNA isoform diversity, enhancing functional complexity in biological systems. Recent advancements in long-read sequencing technologies, notably those by Oxford Nanopore Technologies (ONT), facilitate comprehensive analysis of RNA isoforms through the generation of longer reads capable of capturing full-length transcripts. However, such advancements introduce challenges associated with complex long-read data, particularly in terms of accurate basecalling and robust isoform identification.

In this study, we benchmarked two basecalling algorithms, Guppy and Dorado, utilizing real ONT sequencing data, finding that Dorado greatly enhanced read length and quality, thus improving downstream isoform analysis accuracy. Subsequently, we evaluated five isoform identification and quantification tools (Bambu, IsoQuant, FLAIR, StringTie2, and lr-kallisto) using spike-in RNA datasets as a ground truth. To further dissect how sequencing parameters (read length, read accuracy, sequencing depth, and transcript length) impact tool performance, we generated 96 synthetic ONT RNA datasets. We identified distinct strengths and weaknesses for each tool in isoform discovery depending on varying sequencing qualities and provided guidance for selecting the most suitable tool based on specific research questions. Among the evaluated tools, Bambu exhibited superior performance under the specific conditions of our datasets. Applying the optimized workflow, we compared transcriptional profiles of lung and bladder epithelial cells following exposure to IFN-β, an antiviral cytokine. Notably, we discovered that a truncated form of RIG-I, a key receptor detecting dsRNA viruses such as influenza and coronaviruses, is specifically expressed in lung epithelial cells but not in bladder epithelial cells in response to IFN-β.

These findings underscore the critical importance of selecting appropriate basecalling algorithms and isoform identification tools tailored to the characteristics of the sequencing data, ultimately enabling the identification of cell-type-specific isoform variations with potential therapeutic implications for viral infections.

Acknowledgements

Completing this dissertation has been an incredible journey, one that I could not have navigated without the unwavering support of many wonderful people.

To my wife, Julia, for falling in love with my research and sharing enthusiasm for pretty plots. We took this journey together and experienced every twist and turn. Your unwavering support, help with planning, willingness to listen to my ramblings, and shared appreciation made this journey not only manageable but inspired me to come up with the next best thing.

To my supervisor, Nikki—your passion for RIGI has infected me, and I'm grateful for every symptom. Thank you for the long hours, numerous presentations, and meticulous feedback on my writing. Your approach taught me how to craft a story out of technical writing.

To my mom and dad, for laying the groundwork that allowed me to embark on such a fulfilling and extended journey in pursuit of my passions. Your shared eye for aesthetically pleasing craftsmanship, whether it be a weld or a font, inspired me to pay close attention to the details.

To Paul and Mary Rose, for the daily walks spent deliberating the story of my MSc, and for your unconditional support and guidance in shaping this research.

To Cathy, Dale, Emi, and Mike—for your prayers, but most importantly for enduring countless defeats in Bananagrams. Your sacrifices provided essential fuel for my writing.

To Rama, Monde, Dylan, and Aidan, for sharing each step of this journey, sitting patiently through my practice presentations, and making the process lighter and more enjoyable.

Research Outputs

Presentations

- Ikking, K. and Gentle, N.L. RNA Isoform Profiling in Lung Cells: Implications for COVID-19. University of the Witwatersrand Cross-Faculty Symposium. 2024, Johannesburg. (2nd Place)
- Ikking, K. and Gentle, N.L. Performance assessment of Nanopore sequencing tools for transcriptome analysis. SASG/SASBi Conference. 2024, Pretoria.

Posters

- Ikking K. and Gentle N.L. Exploring the Functional Implications of Alternative Splicing in A549 (lower respiratory tract epithelial cells) Inflammation. Molecular Biology Research Trust Symposium. 2024, Johannesburg.
- Ikking K. and Gentle N.L. Evaluating Tools For Transcript Identification And Quantification From Nanopore Sequencing Data. SASHG. 2024, Sun City. (3rd Place)

Table of Contents

Declaration i
Abstract ii
Acknowledgements iii
Research Outputs iv
Table of Contents v
List of Abbreviations vii
List of Tables ix
List of Figures x
1. INTRODUCTION 1
1.1 The Splicing Machinery 1
1.2 Alternative Splicing 3
1.3 Alternative Splicing in Immunity and Disease 5
1.4 The Retinoic Acid-like Receptors (RLRs) 6
1.5 Next Generation Sequencing Approaches for Studying Alternative Splicing 11
1.6 The History, Fundamentals and Capabilities of ONT Sequencing 13
1.7 Library Preparation Strategies Supported by ONT 14
1.8 Signal Interpretation and Basecalling 14
1.9 Recent Advances in ONT Sequencing 16
1.10 Algorithms for Isoform Discovery Using Long-Read Sequencing 17
1.11 Study Rationale 18
2. METHODS 20
2.1 Data Acquisition 20
2.2 Basecaller and Library Type Comparison 22
2.3 RNA Data Simulations 22
2.3.1 Isoform Length 24
2.3.2 Read Count 24
2.3.3 Read Length 24
2.3.4 Read Accuracy 25

2.4 Genome Alignment Using MiniMap2 25
2.4.1 Minimizers for Efficient Mapping..... 26
2.4.2 Handling Gaps and Spliced Alignment 27
2.5 Isoform Discovery Tools 28
2.5.1 StringTie2..... 31
2.5.2 Bambu..... 32
2.5.3 IsoQuant..... 34
2.5.4 FLAIR 36
2.5.5 lr-kallisto 38
2.6 Data Availability 39
3. RESULTS..... 40
3.1 The Effect of Basecaller Selection on Sequence Data Quality 40
3.2 The Relationship Between Library Type and Sequencing Data Quality 41
3.3 Evaluation of Tools for Transcript Discovery Across Different Library Types 43
3.4 Evaluation of Tools for Transcript Discovery Using Simulated Data..... 44
3.4.1 Sequence Quality..... 45
3.4.2 Transcript and Sequencing Read Length 47
3.5 Evaluation of Tools for Transcript Quantification Using Simulated Data 48
3.6 RLR Transcript Discovery and Quantification 49
4. DISCUSSION 51
REFERENCES 56
Supplementary Figures 67

List of Abbreviations

ASG	Alternative Splicing Graph
AWS	Amazon Web Services
BAM	Binary Alignment/Map Format
BBP	Branch Point Binding Protein
BED	Browser Extensible Data Format
CARD	Caspase Activation and Recruitment Domain
cDNA	Complementary DNA
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CRF	Conditional Random Field
CTC	Connectionist Temporal Classification
CTD	C-terminal Domain
dsRNA	Double-stranded RNA
EG	Exon Graph
GPU	Graphics Processing Unit
GTF	Gene Transfer Format
IFN	Interferon
IFN-β	Interferon Beta
IRF	Interferon Regulatory Factor
ISG	Interferon-Stimulated Gene
LGP2	Laboratory of Genetics and Physiology 2
MAVS	Mitochondrial Antiviral-Signaling Protein
MDA5	Melanoma Differentiation-Associated Gene 5
NMD	Nonsense-Mediated Decay
ONT	Oxford Nanopore Technologies

PCR	Polymerase Chain Reaction
RIG-I	Retinoic Acid-inducible Gene I
RLR	RIG-I–like Receptor
RNN	Recurrent Neural Network
RNA	Ribonucleic Acid
SG-NEx	Singapore Nanopore Expression
SIRV	Spike-in RNA Variant
SJG	Splice Junction Graph
SMRT	Single Molecule Real-Time
snRNA	Small Nuclear RNA
snRNP	Small Nuclear Ribonucleoprotein
SR	Serine/Arginine-rich
SRA	Sequence Read Archive
TCC	Transcript Compatibility Class
TES	Transcription End Site
TPM	Transcripts Per Million
TSS	Transcription Start Site
ASG	Alternative Splicing Graph
AWS	Amazon Web Services
BAM	Binary Alignment/Map Format

List of Tables

Table 2.1: Summary of ONT data used in this study. 11
Table 2.2: MiniMap2 splice-aware parameters used in this study, compared to the aligner's default settings. 13
Table 2.3: Summary of isoform discovery tool methods compared in this study. 15

List of Figures

Figure 1.1: Assembly of the spliceosome enables the removal of intronic regions from pre-mRNA. 2

Figure 1.2: The major mechanisms of alternative splicing. 4

Figure 1.3: Domain structures of the three RLRs and MAVS. 6

Figure 1.4: Activation and downstream signalling cascade of RIG-I and MDA5. 8

Figure 1.5: Alternative splicing of RIG-I regulates activation of MAVS. 10

Figure 1.6: The ONT sequencing platform and its output. 13

Figure 1.7: The basic neural network architecture of basecalling. 15

Figure 2.1: Overview of long-read RNA data simulation pipeline. 23

Figure 2.2: Minimizer selection from k-mers. 27

Figure 2.3: Overview of alternative splicing graph assembly and the application of the maximum flow algorithm by StringTie2. 29

Figure 2.4: Overview of the read class creation and assignment process used by Bambu. 33

Figure 2.5: Overview of the splice junction graph (SJG) and exon graph (EG) approaches for assigning reads to known transcripts. 35

Figure 2.6: Overview of the steps involved in the FLAIR-correct and FLAIR-collapse modules. 37

Figure 3.1: Read length and read accuracy distribution comparison between Guppy and Dorado. 41

Figure 3.2: Read length, read accuracy and read count distributions of three different ONT RNA library preparation techniques. 42

Figure 3.3: The precision and sensitivity scores of each isoform detection tool against different ONT library types. 43

Figure 3.4: Mean F1 scores of different isoform detection tools used against different ONT library types. 44

Figure 3.5: The performance of isoform detection tools varies across different read lengths and read accuracy distributions. 46

Figure 3.6: Heatmap of Precision and Sensitivity Scores for Isoform-Detection Tools Across Simulated Transcript and Read Lengths. 48

Figure 3.7: Proportion of Transcripts Quantified Within 10 %, 20 %, and 30 % of Ground-Truth Abundance. 49

Figure 3.8: The expression profiles of RIG-I and MDA5 isoforms in A549 and HT1376 cells before and after treatment with IFN-B. 50

Figure S1: Characteristics of sequencing data from simulated values. 67

1. INTRODUCTION

1.1 The Splicing Machinery

Splicing refers to the precise removal of introns from precursor messenger RNA (pre-mRNA; Black, 2003). The boundaries between introns and exons are demarcated by conserved sequences. The 5' splice site of introns is typically marked by a conserved "GU" dinucleotide, while the 3' splice site ends with an "AG" dinucleotide. Additionally, the branch point sequence, approximately 20–40 nucleotides upstream of the 3' splice site, includes a conserved adenine nucleotide crucial for catalytic activity. Splicing takes place within the spliceosome, a highly dynamic ribonucleoprotein complex composed of five small nuclear ribonucleoproteins (snRNPs) (U1, U2, U4, U5, and U6) and numerous associated proteins. Each snRNP contains a small nuclear RNA (snRNA) ranging from 100–300 nucleotides in length, which folds into secondary structures that facilitate specific interactions with splice junctions and other components of the spliceosome (Black, 2003).

Spliceosome assembly occurs in a stepwise manner (Figure 1.1). It begins with U1 snRNP recognizing the 5' splice site, a process stabilized by serine/arginine-rich (SR) proteins (Will and Lührmann, 2011). Simultaneously, the branch point adenine is bound by the branch point binding protein (BBP), and U2 auxiliary factor (U2AF) associates with the 3' splice site (E Complex, Figure 1.1). U2 snRNP then replaces BBP, securing the branch point and positioning the spliceosome for catalysis (A Complex, Figure 1.1). Next, U4, U5, and U6 snRNPs associate with the pre-mRNA, and the release of U1 and U4 triggers a structural rearrangement, transitioning the complex into its catalytically active conformation (B Complex, Figure 1.1). The U6 snRNP facilitates two transesterification reactions: first, the hydroxyl group of the branch point adenine attacks the phosphate backbone at the 5' splice site, cleaving it and forming a lariat structure, and U5 snRNP stabilizes the spliced exon during this step. In the second catalytic step, the hydroxyl group at the 3' end of the upstream exon attacks the phosphate group at the 3' splice site, ligating the two exons into a continuous coding sequence and releasing the intron lariat. The lariat structure is

subsequently debranched and degraded, while the mature mRNA is exported to the cytoplasm for translation (Will and Lührmann, 2011).

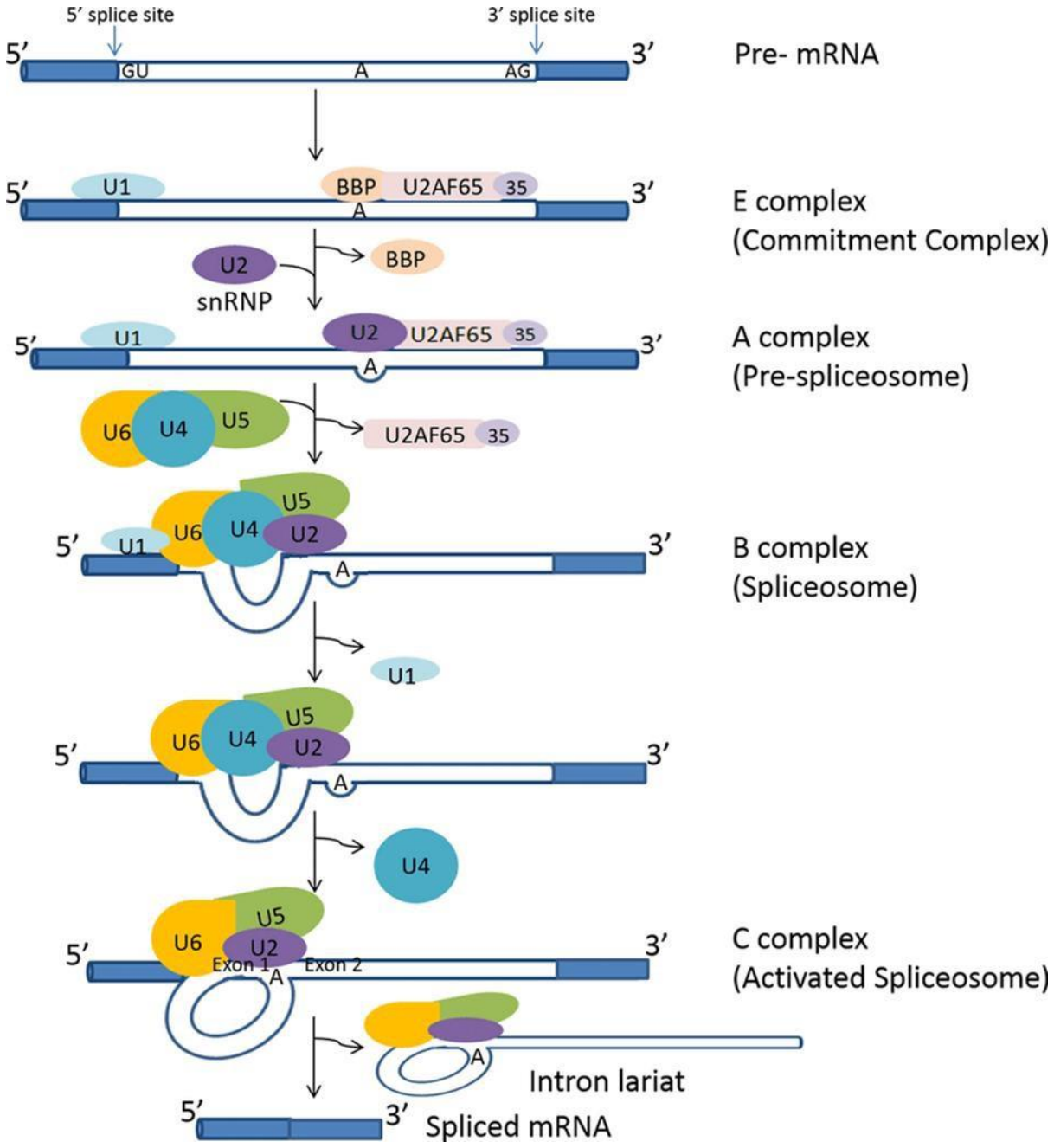


Figure 1.1: Assembly of the spliceosome enables the removal of intronic regions from pre-mRNA. Pre-mRNA (blue) is transcribed from DNA in the nucleus. Dark blue bars represent exons and light blue bars represent introns. This illustration depicts the sequential assembly and catalytic steps of the spliceosome during pre-mRNA splicing. The process begins with U1 snRNP binding to the 5′ splice site, while U2AF65/35 and the branchpoint binding protein (BBP) recognize the branch point. U2 snRNP then replaces BBP, creating a

stable interaction at the branch point. Next, the U4, U5, and U6 snRNPs are recruited, leading to rearrangements that bring the splice sites into proximity. Catalysis proceeds through cleavage of the 5′ splice site to form an intron lariat, followed by cleavage at the 3′ splice site and exon ligation. The final product is a mature mRNA with the intron removed as a lariat structure. Image adapted from Lai *et al.*, (2021).

1.2 Alternative Splicing

Splicing is a highly coordinated process that ensures the precise removal of introns and the correct assembly of exons, which is essential for generating functional mRNAs. However, the dynamic nature of the spliceosome allows for subtle shifts in splicing site usage, often resulting in alternative splicing events that expand proteomic diversity and enable context-specific gene regulation (Manuel *et al.*, 2023). Alternative splicing is a major driver of RNA diversity in eukaryotes, enabling a single gene to produce multiple transcript variants (Liu *et al.*, 2017). These RNA variants, referred to as isoforms, can have distinct functionalities. In humans, more than 94% of protein-coding genes undergoing alternative splicing, makingthis regulatory mechanism a key driver of proteome complexity (Castle *et al.*, 2008). By enabling a single gene to generate multiple protein products, alternative splicing contributes to approximately four times more protein diversity than the number of protein-coding genes alone. The differential usage of isoforms serves as a key mechanism for cell-specific responses (Trapnell *et al.*, 2010), influencing signaling pathways (Kim *et al.*, 2018) and cellular function (Liao and Garcia-Blanco, 2021).

Alternative splicing regulates gene expression by altering exon combinations through one of six major mechanisms (Figure 1.2). As a result, exons can be either constitutive (Figure 1.2A), where they are present in all isoforms of the gene, or non-constitutive (Figure 1.2B), i.e., either spliced out or retained in the different isoforms (Nilsen and Graveley, 2010). The use of alternative 5′ and 3′ splice sites within exons (Figure 1.2C and 1.2D) can alter coding regions, resulting in isoforms with similar protein structures, but with differing allosteric effects (Nilsen and Graveley, 2010). Additionally, gene expression can be regulated through the retention of intron regions (Figure 1.2E), as this often results in nonsense mediated decay of the affected transcript. Lastly, transcript isoforms can form combinations distinct from each other, where the presence of exons in the transcript are mutually exclusive from each other

(Figure 1.2F). Thus, depending on regulation at the spliceosome, isoforms can exhibit functionally significant differences due to their exon combinations (Nilsen and Graveley, 2010).

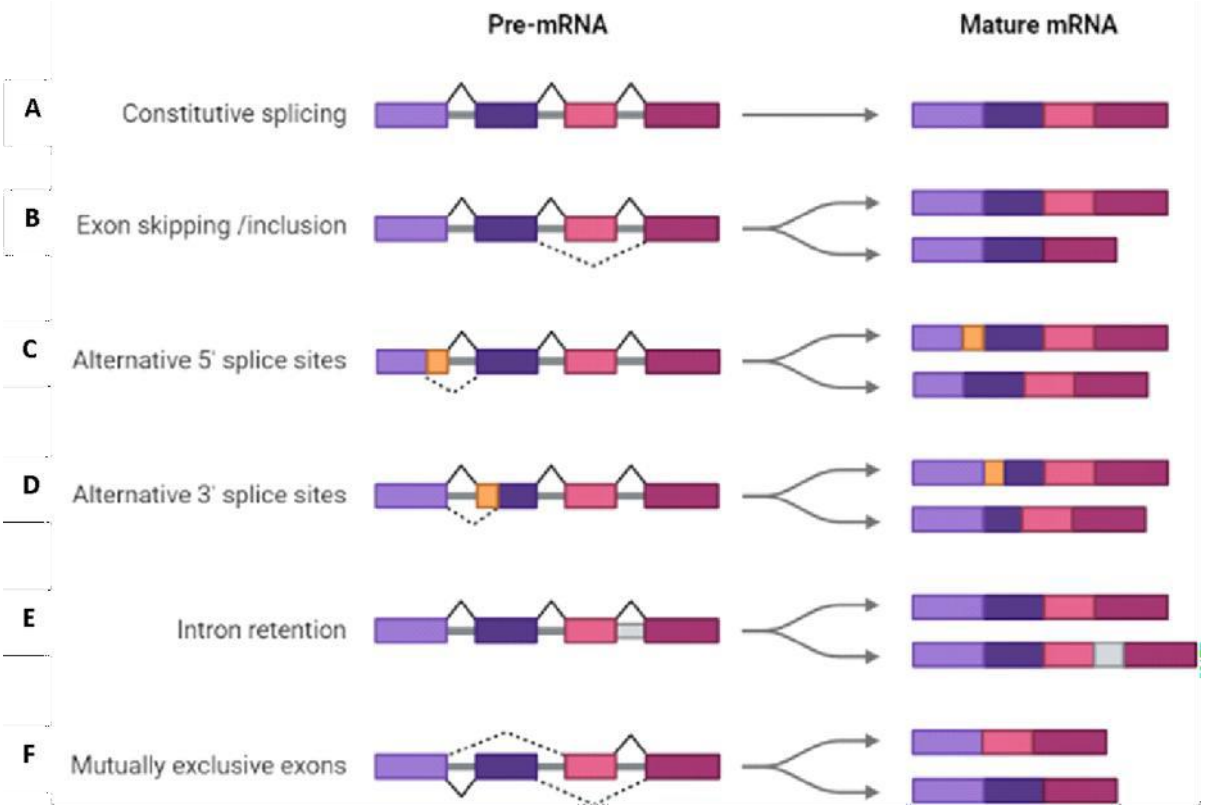


Figure 1.2: The major mechanisms of alternative splicing. A) Constitutive splicing results in all exons always appearing in the mature mRNA. B) Exon skipping or inclusion results in non-constitutive exons either being retained or spliced out. C) The use of alternative 5′ and D) 3′ splice-sites results in either the inclusion or exclusion of a portion of the affected exon. E) Intron retention results in isoforms containing intron regions. F) The inclusion/exclusion of mutually exclusive exons results in isoforms with consistently distinct exonic combinations. Image created with BioRender.com. Adapted from Black, 2003.

Where there is greater organismal complexity, reflected in the number of unique cell types, a higher proportion of alternative splicing events can be observed. In humans, for example, a substantially larger fraction of transcripts undergo alternative splicing compared to less complex organisms, underscoring the critical role of this mechanism in generating both cell-specific and condition-specific responses (Manuel *et al.*, 2023). These tightly regulated splicing patterns emerge from interactions between

specialized regulatory elements and RNA-binding proteins, which can selectively enhance or suppress the splicing of individual transcripts, thus enabling precise control of gene expression in different cell types and under varying physiological conditions (Arias *et al.*, 2010). While alternative splicing has been extensively investigated in neural development (Su *et al.*, 2018), it is also increasingly recognized as a key determinant in adaptive immunity (Zikherman and Weiss, 2008). Understanding how these splicing mechanisms operate in immune cells sets the stage for exploring their broader significance in health and disease.

1.3 Alternative Splicing in Immunity and Disease

Immunity is fundamental for host defense but must be tightly regulated to balance a strong enough response to prevent infection, but also a modulated response to prevent collateral tissue damage (Mihaylova *et al.*, 2018). This regulation often takes place at the tissue level, where distinct cellular compositions and local microenvironments shape immune responses. For example, lung cells adopt a predominantly cytoprotective response compared to nasal epithelial cells, thereby preserving the delicate pulmonary tissue (Mihaylova *et al.*, 2018). This nuanced immune modulation is frequently orchestrated by alternative splicing, which generates diverse protein isoforms that fulfill specialised roles across various tissues and contexts. For example, macrophages can produce specific isoforms that dampen cytokine production and mitigate the risk of an overactive inflammatory reaction (De Arras and Alper, 2013). At the epithelial cell level, alternative splicing involves switching important pathogen sensing membrane-bound receptors, to their soluble and less active forms, a process that helps restore homeostasis by curbing pro-inflammatory signaling once an acute response is no longer necessary (Schaub and Glasmacher, 2017). Given that alternative splicing also underlies key aspects of cellular differentiation, understanding its role in regulating these pathogen sensing receptors in cell-specific contexts can offer valuable insights into the mechanisms of immune regulation in disease.

1.4 The Retinoic Acid-like Receptors (RLRs)

One of the major families of pathogen recognition receptors at the forefront of innate immunity are retinoic acid-like receptors (RLRs). RLRs are cytosolic receptors that recognise pathogen-associated molecular patterns (PAMPs) and trigger an inflammatory cascade in response to non-self-nucleic acids (Kato and Fujita, 2015), including a variety of viruses (Cárdenas *et al.*, 2006; Linehan *et al.*, 2018; Zhang *et al.*, 2016). The RLR family of receptors includes retinoic-inducible gene-I (RIG-I), melanoma differentiation-associated protein 5 (MDA5), and Laboratory of Genetics and Physiology 2 (LGP2; Takeuchi and Akira, 2010). RIG-I and MDA5 possess two caspase activation and recruitment domains (CARDs; Figure 1.3), which mediate signal transduction (Rehwinkel and Gack, 2020). LGP2, lacking CARD domains, modulates the activities of RIG-I and MDA5 by either enhancing or suppressing their signaling (Rehwinkel and Gack, 2020).

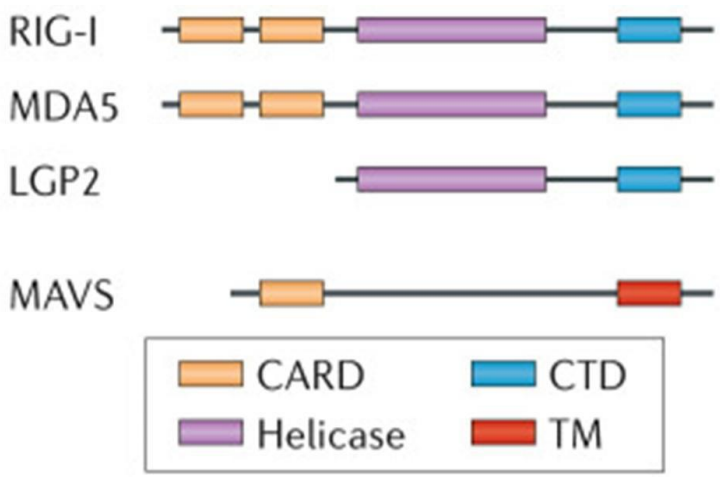


Figure 1.3: Domain structures of the three RLRs and MAVS. Shown are the key domains of the three RIG-I–like receptors (RIG-I, MDA5, and LGP2) and mitochondrial antiviral-signaling protein (MAVS), with their color-coded regions: the N-terminal CARD domains in orange, helicase domains in purple, and the CTD in blue. Adapted from Rehwinkel and Gack, 2020.

Both MDA5 and RIG-I detect viral RNA in the cytoplasm and trigger antiviral signaling, yet they differ somewhat in their RNA-binding preferences. RIG-I generally recognizes shorter 5′ triphosphate–bearing double-stranded RNA (dsRNA) or partially double-stranded structures that arise often from single-stranded RNA viruses that form hairpins, whereas MDA5 preferentially binds longer dsRNA molecules (Matsumiya and Stafforini, 2010; Rehwinkel and Gack, 2020; Ren *et al.*, 2019). Each receptor contains a helicase domain with ATP-hydrolyzing activity and a C-terminal domain (CTD; Figure 1.3) that directly engages viral RNA (Rehwinkel and Gack, 2020). In RIG-I, the CTD has high affinity for the 5′ triphosphate and adjacent unmethylated nucleotides, which is critical for distinguishing viral from host RNA (Ren *et al.*, 2019). When viral RNA binds to the CTD, the helicase domain can undergo conformational changes, driven by ATP hydrolysis, that wrap the receptor around the RNA (Kell and Gale, 2015). In their resting (closed) conformation, RIG-I and MDA5 both keep their CARDs masked by the helicase region (Loo and Gale, 2011). RNA binding induces a shift to an open conformation that exposes these CARDs (Loo and Gale, 2011), which then require ubiquitination to interact with the CARD domain of the mitochondrial antiviral-signaling protein (MAVS; Figure 1.3). This interaction initiates downstream antiviral pathways resulting in the production of type I and type III interferons (Liu *et al.*, 2017).

RLRs function as pattern recognition receptors (PRRs) that detect viral RNA and initiate immune signaling (Figure 1.4) by triggering activation of the MAVS. Downstream of MAVS signaling, the transcription factors, IRF3 and IRF7, along with NF-κB, become activated, leading to an innate immune response (Figure 1.4; Kell and Gale, 2015; Rehwinkel and Gack, 2020). IRF3 and IRF7 are typically dormant until they are phosphorylated, at which point they translocate to the nucleus and promote the expression of type I and type III interferon (IFN) genes (Figure 1.4) (Yanai *et al.*, 2018). These IFNs, once secreted, bind to interferon receptors (IFNARs for type I IFNs and IFNLRs for type III IFNs) and initiate signaling via tyrosine and janus kinases (TYK2 and JAK1), which in turn induces numerous interferon-stimulated genes (ISGs) with antiviral and inflammatory functions (McNab *et al.*, 2015). A positive-feedback loop is then established, because IFNs act on the same cells that produce them and on neighboring cells, thereby amplifying the antiviral state of surrounding tissue (Stanifer *et al.*, 2020).

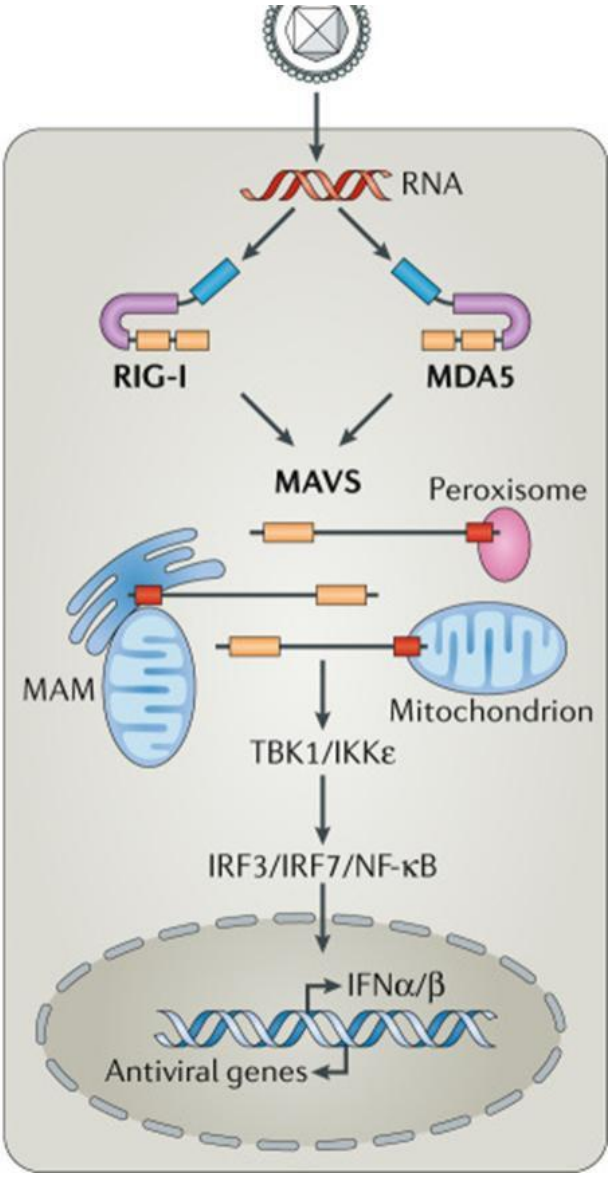


Figure 1.4: Activation and downstream signalling cascade of RIG-I and MDA5. RIG-I and MDA5 detect viral RNA and converge on MAVS to initiate antiviral responses. Upon viral infection, double-stranded RNA is recognized by RIG-I or MDA5, which undergo conformational changes and recruit MAVS. MAVS, localized to both the mitochondrial membrane and peroxisomes, then activates the kinase's TBK1 and IKKε, leading to the phosphorylation and activation of the transcription factors, IRF3, IRF7, and NF-κB. These factors translocate to the nucleus to induce the expression of type I (IFNα/β) or type III (IFN-λ) interferons and other antiviral genes, thereby establishing an antiviral state within the cell. Adapted from Rehwinkel and Gack, 2020.

RIG-I–like receptor (RLR) signaling requires careful regulation to prevent excessive immune activation and inflammation (Rehwinkel and Gack, 2020). In particular, RIG-I can trigger a large array of immunogenic genes through a positive-feedback loop largely driven by interferons, making precise control of its activity essential for immune homeostasis. Multiple layers of regulation have been characterized (Gack *et al.*, 2008, 2007), including post-translational modifications (e.g., ubiquitination and phosphorylation), post-transcriptional mechanisms such as alternative splicing, competitive inhibition of ligand binding, and the controlled downregulation of *DDX58* (the gene encoding RIG-I). Under basal conditions, RIG-I is expressed at relatively low levels (Matsumiya and Stafforini, 2010), but IFN stimulation, especially via early IFN-β production by other pattern-recognition receptors, greatly increases RIG-I abundance and readies cells for a robust antiviral response (Kell and Gale, 2015; Matsumiya and Stafforini, 2010). Once activated, RIG-I amplifies a positive-feedback loop, activating additional interferons and IRFs that further boost RIG-I expression (Matsumiya and Stafforini, 2010).

A crucial target of RIG-I regulation lies in its CARD domains, whose interaction with MAVS is required for antiviral signaling (Kell and Gale, 2015). Overexpression of these CARDs alone drives constitutive downstream activation (Matsumiya and Stafforini, 2010), while non-synonymous mutations or domain disruptions can entirely abrogate signaling (Matsumiya and Stafforini, 2010). Beyond protein-level control, alternative splicing of *DDX58* adds another dimension of regulation. Among at least ten known splice variants, only one yields the full-length, fully functional receptor (Liao and Garcia-Blanco, 2021). One notable variant lacks a CARD segment required for TRIM25-mediated ubiquitination, precluding effective MAVS engagement; however, it retains the RNA-binding C-terminal domain, thereby sequestering viral RNA and competitively inhibiting the full-length receptor (Figure 1.5; Gack *et al.*, 2008). Intriguingly, this particular isoform appears exclusively in response to IFN-β, and its overexpression leads to reduced NF-κB activation (Figure 1.5; Gack *et al.*, 2008). Given its evident role in modulating RIG-I activity, it is now important to quantify the expression of this isoform relative to the dominant form to determine its impact on the overall antiviral response.

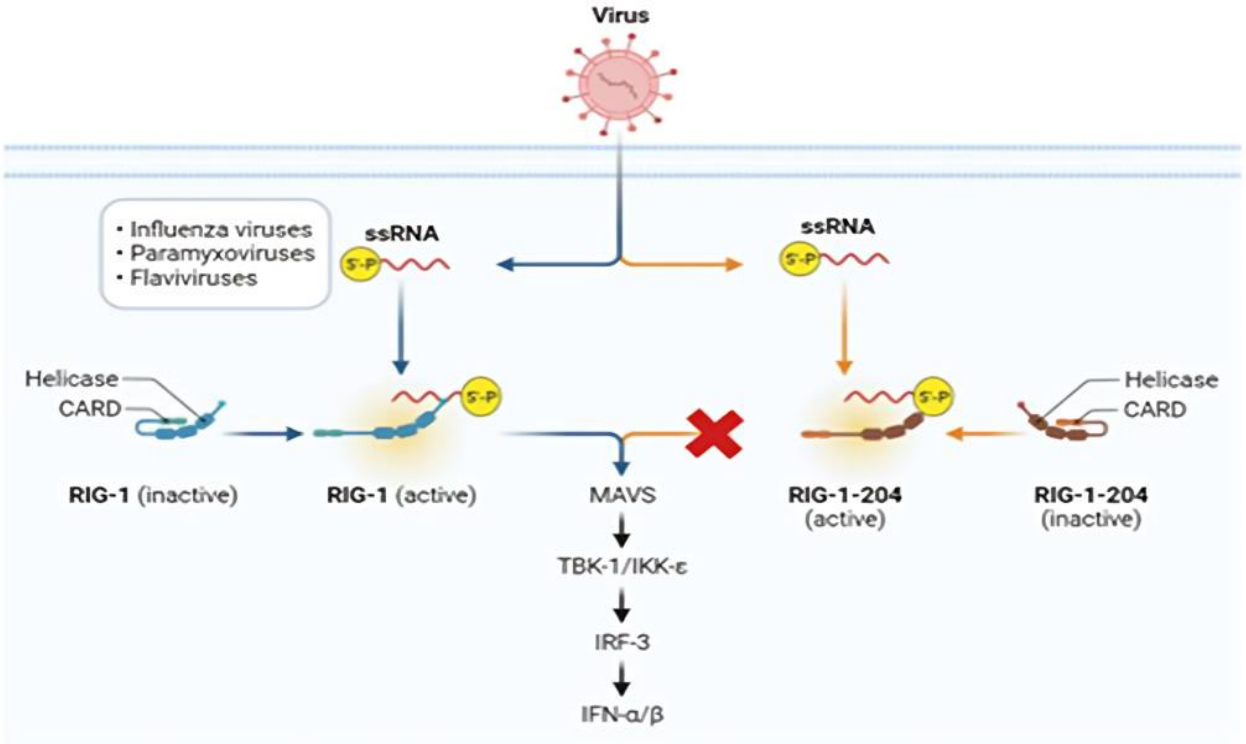


Figure 1.5:Alternative splicing of RIG-I regulates activation of MAVS. Upon binding viral ssRNA, RIG-I isoforms containing a functional CARD domain interact with MAVS, activating downstream immune signalling. Isoforms lacking this domain are unable to interact with MAVS, thereby inhibiting immune signalling. RIG-1 isoforms containing and lacking the relevant CARD domain are shown in blue and red, respectively. Adapted from Gack *et al.*, 2008.

1.5 Next Generation Sequencing Approaches for Studying Alternative Splicing

Over the past two decades, short-read sequencing technologies (e.g., Illumina) have been pivotal in transcriptome profiling (Goodwin *et al.*, 2016). Their high base-calling accuracy (often exceeding 99.5%) and the availability of mature bioinformatics analysis pipelines have made them a reliable standard (Trapnell *et al.*, 2010). However, short fragments (75–150 bp) are inadequate for fully capturing the complexity of alternative splicing, particularly in transcripts with extensive exon–intron structures or closely related isoforms (Savarese *et al.*, 2018). The core limitation lies in reconstructing entire isoforms, which is crucial for distinguishing minor sequence variations and assigning reads unambiguously across repetitive regions (Chen *et al.*, 2021a). Ultimately, short reads lack the resolving power required for a comprehensive assessment of isoform diversity.

Long-read sequencing technologies, including PacBio's Single Molecule Real-Time (SMRT) sequencing and Oxford Nanopore Technology (ONT) sequencing, have emerged to address the length restrictions of short reads (Jain *et al.*, 2018). Unlike short-read approaches, which must reconstruct transcripts from overlapping fragments, the PacBio and ONT technologies can generate reads spanning tens of kilobases, allowing near-complete coverage of full-length isoforms in a single pass (Amarasinghe *et al.*, 2020). PacBio SMRT sequencing accomplishes this via zero-mode waveguides and real-time detection of nucleotide incorporations, whereas ONT measures fluctuations in ionic current as nucleotides traverse a biological nanopore (Goodwin *et al.*, 2016). Crucially, these extended read lengths enable more accurate delineation of exon–intron boundaries and splicing events, reducing reliance on computational assembly steps. This increased resolution has facilitated research into alternative splicing in complex eukaryotic systems, aiding in the discovery of crucial isoforms (Volden *et al.*, 2022).

Although long-read platforms promise a deeper look into the transcriptome, they remain susceptible to lower base-calling accuracies, typically 80–95% (Amarasinghe *et al.*, 2020). Insertions and deletions (indels) in particular can corrupt the interpretation of splice junctions, potentially creating false positives or negatives in isoform detection (Slaff *et al.*, 2021). Additionally, library preparation protocols

themselves can introduce artifacts, such as partial cDNA synthesis or incomplete strand termination, which complicate downstream analyses (Tang *et al.*, 2020). Error-correction algorithms, including Racon (Vaser *et al.*, 2017), Medaka (https://github.com/nanoporetech/medaka), or LoRDEC (Salmela and Rivals, 2014), can refine read quality, but these steps add computational overhead and can inadvertently remove rare yet biologically relevant transcripts if applied too aggressively (Lee *et al.*, 2021). Thus, while long-read methods provide a transformative improvement over short reads, researchers must remain vigilant about error sources and biases and potentially use short read tools to validate their discoveries.

PacBio SMRT sequencing has long reported raw read accuracies of ~99 %, and its newer HiFi circular-consensus reads push per-base accuracy even higher, making the platform well-suited to projects that demand exceptionally high-fidelity data (Udaondo et al., 2021). In contrast, ONT has grappled with higher error rates and a steeper learning curve for bioinformatics support compared to the more established PacBio environment (Udaondo *et al.*, 2021). ONT's first commercial product, the MinION, provided unprecedented portability, but initially faced throughput and stability challenges (Kasianowicz *et al.*, 1996). Over time, improved flow cell designs, with thousands of nanopores, and optimized chemistries yielded more robust data, culminating in high-volume platforms like the GridION and PromethION (Espinosa *et al.*, 2024).

In parallel, significant progress in basecalling algorithms has addressed the high signal-to-noise ratio, historically a major bottleneck that hindered ONT's ability to translate raw signals into accurate nucleotide sequences (Wick *et al.*, 2019). Over the past five years, significant improvements in basecalling algorithms have further enhanced ONT's signal decoding capabilities (Pagès-Gallego and de Ridder, 2023). As a result, ONT excels at producing ultra-long reads (exceeding 100 kb in some cases), which are particularly useful for spanning repetitive regions and capturing full-length isoforms. Moreover, ONT's direct RNA sequencing capabilities allow investigation of epitranscriptome features and poly(A) tail lengths without amplification biases (Garalde *et al.*, 2018). Therefore, in this study, the choice was made to focus on data obtained using ONT sequencing, due to its rapidly evolving throughput, relatively lower costs, and full-length transcript capturing approaches that hold promise for fully characterising the alternative splicing landscape.

1.6 The History, Fundamentals and Capabilities of ONT Sequencing

ONT was established in 2005, building on research that pioneered the use of nanopores for single-molecule detection (Deamer *et al.*, 2016). Early prototypes demonstrated the feasibility of threading DNA or RNA strands through protein pores under an applied voltage, using subtle ionic current changes to infer base composition (Kasianowicz *et al.*, 1996). ONT sequences nucleic acids by measuring current changes across a ~180 mV gradient (Figure 1.6A), as negatively charged DNA or RNA passes through a 1.2 nm pore embedded in a synthetic membrane (Deamer *et al.*, 2016). An enzyme, typically phi29 DNA polymerase, slows translocation from microseconds per base to hundreds of microseconds (Figure 1.6B), improving the signal-to-noise ratio (Deamer *et al.*, 2016). Each nucleotide (A, C, G or T) presents a slightly different impedance signature, with multiple adjacent bases often simultaneously contributing to the measured current (Deamer *et al.*, 2016). Unlike Illumina or PacBio, ONT does not require optical detection, relying instead on purely electrical measurements.

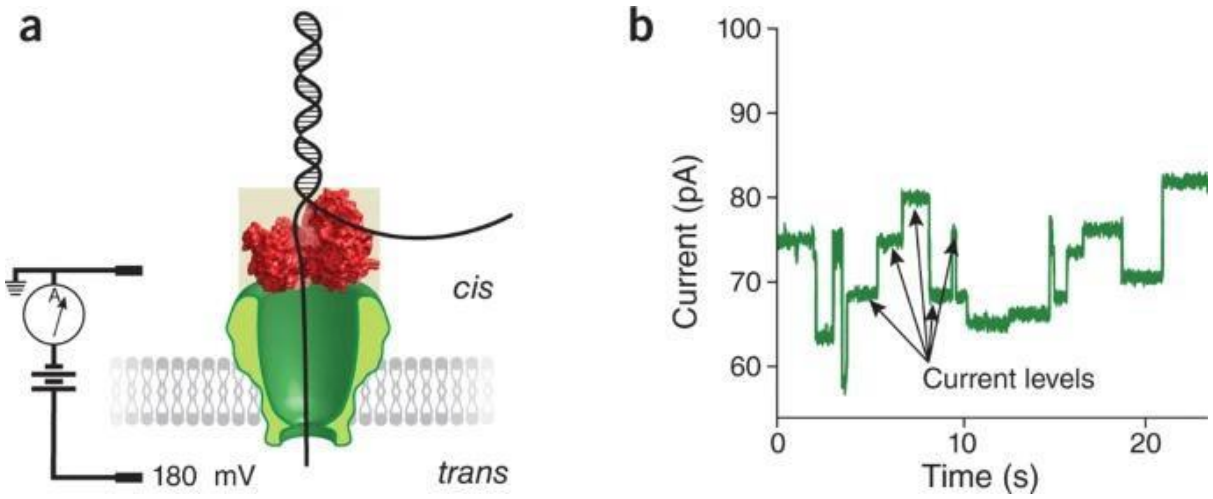


Figure 1.6: The ONT sequencing platform and its output. A) ssDNA is motored through the nanopore (green) at a constant speed by an enzyme (red). B) A signal graph is generated by measuring the impedance of nucleotides passing through the pore. Figure adapted from Deamer, Akeson and Branton, 2016.

1.6.1 Library Preparation Strategies Supported by ONT

ONT provides two principal routes for RNA sequencing library preparation: direct RNA sequencing, which preserves RNA modifications (e.g., m6A), in the absence of reverse transcription (Workman *et al.*, 2019), and cDNA-based approaches, which convert RNA into complementary DNA for more robust yields and the removal of base modifications that could potentially confound signal interpretation (Huang *et al.*, 2022). cDNA libraries can either be sequenced directly, or PCR amplified to aid in the detection of low-abundance isoforms, although each amplification step introduces the possibility of bias or chimera formation (Oikonomopoulos *et al.*, 2016).

Because direct RNA and cDNA libraries differ fundamentally, each has its pitfalls. Direct RNA libraries can suffer from reduced throughput and require specialized basecalling models that handle modified bases (Wang *et al.*, 2024). Conversely, cDNA approaches rely on reverse transcription, potentially leading to incomplete coverage of very long transcripts or biases against certain secondary structures (Chen *et al.*, 2021b). Furthermore, PCR amplified libraries risk exponential amplification of certain transcripts at the expense of others (Aird *et al.*, 2011). Consequently, the choice of library approach must weigh the trade-offs between depth of coverage, sample complexity, and the desire for information about native RNA modifications.

1.6.2 Signal Interpretation and Basecalling

Basecalling is the computational process that transforms raw nanopore signals (i.e. fluctuations in ionic current) into digital nucleotide sequences. Each nucleotide, or small group of nucleotides, temporarily obstructs the pore, altering the current in a signature, characteristic manner (Albrecht *et al.*, 2017). Although these signals are highly informative, overlapping bases and inherent electrical noise complicate decoding, causing higher reliance on machine learning algorithms rather than static algorithms to differentiate subtle current differences across multiple contiguous nucleotides (Albrecht *et al.*, 2017).

Many modern basecallers employ deep neural networks, such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), or hybrids thereof, to interpret the dynamic, time-series data generated by nanopore sequencers (Pagès-Gallego and de Ridder, 2023). In these models, the signal is segmented according to size and position to form input nodes (Figure 1.7). Weighted connections between the input nodes and a hidden layer then generate the most probable base as the output. During training, known outputs are used to iteratively adjust the weights through backpropagation, thereby improving the model's overall accuracy (Wen *et al.*, 2021).

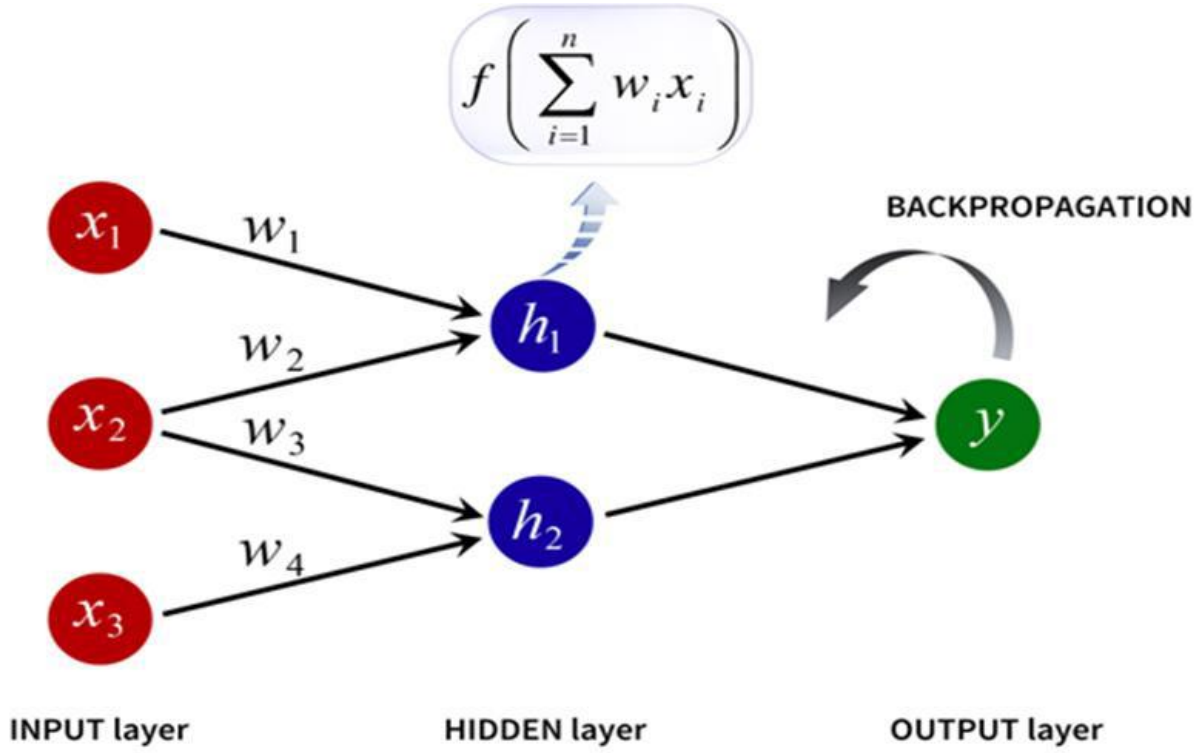


Figure 1.7: The basic neural network architecture of basecalling. Red nodes represent input vectors (size, position, slope etc.). Blue nodes represent the hidden layer that processes that vector according to the weighting of the branch (w). Green represents the output node that presents the most likely base call associated with the input information. Backpropagation is used during model training to alter the weightings (w) for more accurate input value interpretation. Figure adapted from Wen, Dematties and Zhang, 2021.

The general architecture of modern basecalling pipelines typically involves three main components. First, an encoder, often implemented as a CNN, transforms the incoming current signals into a compressed or “binary” representation (Rang *et al.*, 2018). CNNs, widely used in image recognition, are adept at detecting local features by applying convolutional filters across the signal, thereby helping to distinguish meaningful signal patterns from random noise (Rang *et al.*, 2018) Second, a predictive model, commonly a RNN, processes this encoded sequence over time and leverages hidden states to incorporate information from previous time steps, enabling the model to account for context-dependent variations in the nanopore signal (Xu *et al.*, 2021). Finally, a decoder, frequently a Connectionist Temporal Classification (CTC) or Conditional Random Field (CRF), converts the time-series output of the predictive model into a final base-by-base call. Decoders such as CTC and CRF handle variable-length outputs and fine partial or noisy predictions from the RNN into coherent nucleotide sequences (Pagès-Gallego and de Ridder, 2023). Together, these three stepwise components (encoder, predictive model, and decoder) form a robust pipeline that systematically reduces noise and ultimately translates raw nanopore data into high-confidence genomic reads.

1.6.3 Recent Advances in ONT Sequencing

Historically, the computational demands of decoding raw nanopore signals, an inherently neural network–driven process, have constrained ONT basecalling pipelines (Wick *et al.*, 2019). Central processing units (CPUs), with relatively few cores and limited parallel processing capabilities, often struggled under this heavy load, leading to prolonged run times and high computational overhead. By contrast, graphics processing units (GPUs) feature thousands of specialized cores optimized for the calculations important for neural network computations, thereby enabling faster throughput and more sophisticated decoding algorithms (Buber and Diri, 2018). Among these advancements are CRF-based decoders, which can incorporate information from previous basecalls to improve accuracy (Goyal *et al.*, 2018; Pagès-Gallego and de Ridder, 2023). Building on these GPU-focused innovations, ONT released Dorado (https://github.com/nanoporetech/dorado), an open-source basecaller designed to leverage the parallel architecture of GPUs . While its primary benefit lies in increasing basecalling speed, how Dorado affects overall sequence quality remains largely unexplored.

1.7 Algorithms for Isoform Discovery Using Long-Read Sequencing

The development of isoform discovery tools tailored for long-read sequencing has focused on three key objectives: error correction, transcript clustering, and structural reconstruction (Chen *et al.*, 2023; Kovaka *et al.*, 2019; Prjibelski *et al.*, 2023; Tang *et al.*, 2020). Long-read errors, particularly at splice junctions, must first be corrected to ensure accurate isoform detection. Once error correction is performed, reads are grouped based on shared splice-junction patterns, enabling the identification of distinct transcript variants. The final step involves the reconstruction of isoform structures, often using alternative splice graphs (ASGs) or other computational models that differentiate between closely related isoforms (Chen *et al.*, 2023; Kovaka *et al.*, 2019; Prjibelski *et al.*, 2023; Tang *et al.*, 2020).

Several isoform discovery tools have been developed, each employing distinct strategies for transcript reconstruction. StringTie2 and IsoQuant utilize ASGs to cluster and distinguish isoforms based on shared exon–intron boundaries (Kovaka *et al.*, 2019; Prjibelski *et al.*, 2023). FLAIR employs a reference-guided approach in which long-read sequences are first used to construct a high-confidence transcriptome assembly, followed by realignment of reads to this reference to refine isoform predictions (Tang *et al.*, 2020). Bambu incorporates machine-learning techniques to dynamically adjust correction windows and filtering thresholds based on sequence quality, thereby improving sensitivity to isoform detection (Chen *et al.*, 2023). Other tools, such as lr-kallisto, adapt short-read quantification methodologies to long-read data, achieving high efficiency in alignment and quantification, but sacrificing sensitivity in the detection of novel isoforms (Loving *et al.*, 2025). Despite these advancements, distinguishing true isoforms from sequencing artifacts remains a significant challenge, particularly in complex transcriptomes with extensive alternative splicing

1.8 Study Rationale

Despite the rapid evolution of ONT, including GPU-optimized basecalling algorithms and refined sequencing protocols, no recent, comprehensive benchmarking effort has evaluated the implications of these improvements for isoform detection (Chen *et al.*, 2021; Dong *et al.*, 2023; Su *et al.*, 2024). This gap leaves researchers uncertain about whether re-basecalling older ONT datasets would yield meaningful improvements, and how best to accommodate differences in library preparation methods when analyzing isoforms. By addressing these uncertainties, the present study aims to clarify whether updated pipelines can unlock new insights into alternative splicing and to provide practical guidelines for selecting an optimal isoform detection strategy across diverse ONT sequencing conditions.

Furthermore, although several studies have compared the strengths and weaknesses of isoform discovery tools, none have emerged as a definitive gold standard (Chen *et al.*, 2021; Dong *et al.*, 2023; Prjibelski *et al.*, 2023; Pardo-Palacios *et al.*, 2024; Su *et al.*, 2024; Loving *et al.*, 2025). Research using synthetic spike-in RNA variants (SIRVs) has highlighted varying degrees of precision, recall, and computational efficiency among different methods (Chen *et al.*, 2021; Dong *et al.*, 2023; Su *et al.*, 2024). For example, StringTie2 is known for high precision, whereas Bambu excels in capturing a broader range of isoforms (Dong *et al.*, 2023). More recently, IsoQuant outperformed both StringTie2 and Bambu on ONT Pore version 10.4 data but showed a sharper decline in accuracy under low read abundance (Prjibelski *et al.*, 2023). Despite these comparative assessments, a systematic investigation that rigorously accounts for key sequencing parameters, such as read accuracy, read length, and total read count, is still lacking, making it difficult to identify the most reliable tool for challenging sequencing environments.

In parallel, the growing feasibility of long-read sequencing has spurred interest in alternative splicing, particularly for genes involved in immune regulation. Previous methods were constrained by limited throughput, hindering simultaneous analysis of multiple transcripts. This shortcoming has left baseline expression levels of functionally important isoforms, such as those modulating RLRs, largely undefined, complicating efforts to detect aberrations in disease contexts. To bridge this gap, we systematically evaluated five isoform discovery and quantification tools (Bambu,

FLAIR, IsoQuant, lr-kallisto, and StringTie2) using both real and simulated ONT datasets. By focusing on RLR isoforms (RIG-I and MDA5) in human epithelial cells (A549 and HT1376) treated with or without IFN-β, our goal was to determine the most suitable approaches for identifying and quantifying these critical immune-related transcripts under varying experimental conditions. The above aims were achieved through the following objectives:

1. To assess the impact of basecalling models on ONT sequencing data quality
2. To assess the impact of choice of library type on ONT sequencing data quality
3. To assess the impact of choice of library type on tool performance with respect to isoform discovery
4. To assess the performance of tool performance with respect to isoform discovery using simulated ONT sequencing data
5. To assess the performance of tool performance with respect to isoform quantification using simulated ONT sequencing data
1. To identify the most suitable tool to identify and quantify RLR isoforms in A549 and HT1376 cells treated with and without IFN-β

2. METHODS

2.1 Data Acquisition

To evaluate the impact of basecaller selection on sequencing data and benchmark isoform discovery tools, datasets from the Singapore Nanopore Expression (SG-NEx) project, a large-scale collaboration to establish standardized benchmarking datasets for long-read sequencing, were accessed via Amazon Web Services (AWS) S3 bucket *arn:aws:s3:::sg-nex-data* (Chen *et al.*, 2021). Three direct cDNA samples, two PCR-amplified cDNA samples, and one direct RNA replicate from the HCT116 cell line, all treated with spike-in RNA and provided in FAST5 format, were downloaded to assess the impact of library preparation techniques on data quality (Table 2.1). The inclusion of spike-in RNA controls is pivotal for this analysis, as these exogenous transcripts with known sequences and concentrations act as ground-truth values for real data and enable the ability to determine whether isoform discovery tools can accurately recover and quantify expected transcript variants. By monitoring the recovery of these spike-in isoforms, we can effectively benchmark the sensitivity and reliability of different downstream isoform detection pipelines using real data. The three direct cDNA samples were used to compare basecaller model performance. Additionally, FASTQ files of A549 and HT1376 cells treated with IFN-β and mock-treated controls were obtained from the NCBI Sequence Read Archive (SRA) under accession PRJNA743928 (Table 2.1). These data were used to identify RLR isoforms specifically expressed following IFN-β treatment in lung epithelial cells compared to bladder epithelial cells (Banday *et al.*, 2022).

Table 2.1: Summary of ONT data used in this study.

Sample	Library	Flow cell	Sequencing Kit	Sequencing Platform	No. of Replicates	Reference
Hct116	Direct cDNA	FLO-MIN106D	SQK-DCS109	GridION	3	(Chen *et al.*, 2021)
Hct116	PCR cDNA	FO-MIN106D	SQK-PCS109	GridION	2	(Chen *et al.*, 2021)
Hct116	Direct RNA	FLO-MIN106	SQK-RNA002	GridION	1	(Chen *et al.*, 2021)
IFN-β Treated HT1376	Direct cDNA	FLO-MIN106D	SQK-DCS109	GridION	1	(Banday *et al.*, 2022)
Mock Treated HT1376	Direct cDNA	FLO-MIN106D	SQK-DCS109	GridION	1	(Banday *et al.*, 2022)
IFN-β Treated A549	Direct cDNA	FLO-MIN106D	SQK-DCS109	GridION	1	(Banday *et al.*, 2022)
Mock Treated A549	Direct cDNA	FLO-MIN106D	SQK-DCS109	GridION	1	(Banday *et al.*, 2022)

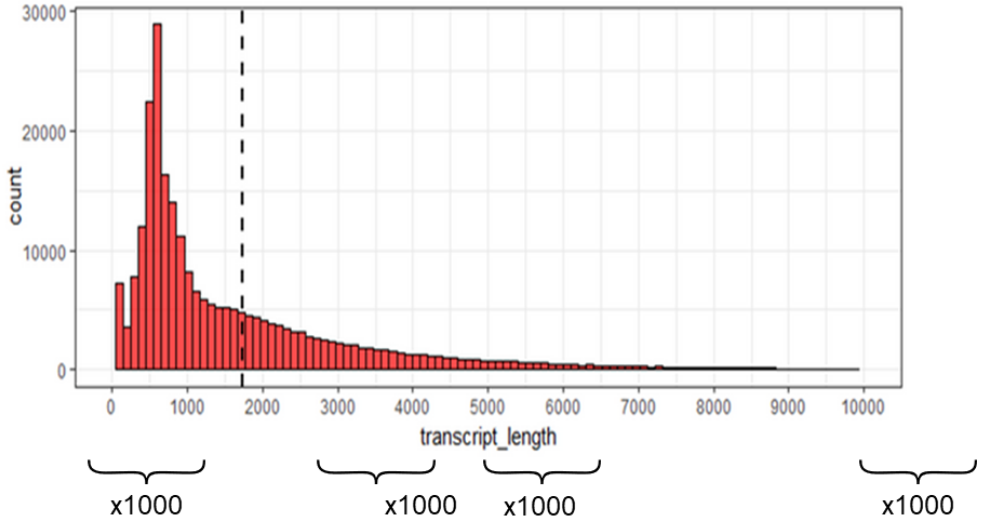
2.2 Basecaller and Library Type Comparison

FAST5 files were converted to the POD5 format using the Fast5toPod5 tool developed by ONT (https://pod5.nanoporetech.com/). This conversion accelerated basecalling speeds and reduced read/write bottlenecks during processing. Additionally, individual FAST5 files were consolidated into multiple smaller files (100 sequences per file) to optimise single-read processing performance by reducing memory overhead. To evaluate the effect of newer base-calling algorithms on data quality, the same datasets were processed with Guppy v6.5.7 (https://nanoporetech.com/document/Guppy-protocol) and Dorado v0.4.1 (https://github.com/nanoporetech/dorado). Guppy was run with the parameters specified for each sample by using the *--flowcell* and *-- kit* options. For Dorado, basecalling was performed using the High Accuracy R9.4.1 pore models for DNA (DNA_r9.4.1_hac@v3.3) and for RNA (rna004_130bps_hac@v3.0.1) in GPU mode. Differences in read-length and read-accuracy distributions between basecallers and library types were compared using direct-cDNA data from HCT116 cells (*n = 3*) generated by the SGNEX project (Table 2.1).

2.3 RNA Data Simulations

To evaluate the performance of isoform discovery tools under controlled conditions, RNA sequencing data was simulated using PBSIM3 v3.0.4 (Ono *et al.*, 2022). This approach was chosen to account for the significant variability in ONT sequencing data, which can arise due to factors such as library preparation methods, pore and flow-cell versions, and basecalling software (e.g., Guppy version). The use of simulated data allowed for two key advantages: The systematic assessment of how different sequencing features (read length, transcript length, read accuracy, and read count) impacted tool performance, and the provision of ground truth values, enabling direct comparison between reconstructed isoforms and known transcript structures. A total of 96 RNA data samples were generated (Figure S1), with specific parameters controlled to reflect typical sequencing characteristics, including read length, read count, read accuracy, and isoform length (Figure 2.1).

A



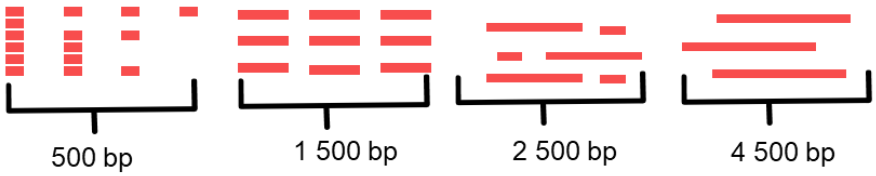
Count Value Assignment

B

ENST00001234	ACTGAAAAAAAATTTTTTTTCTGC	5
ENST00001234	ACTGAAAAAAAATTTTTTTTCTGC	5
...	....	
ENST00001234	ACTGAAAAAAAATTTTTTTTCTGC	5

....

C



D

Figure 2.1: Overview of long-read RNA data attributes generated by the simulation pipeline. (A) Transcript length distribution in the Gencode v45 annotation. The histogram depicts the four length categories sampled for the simulation: < 2000 bp, 2000–4000 bp, 4000–6000 bp, and > 10,000 bp. Brackets below the graph highlight the sampled categories. (B) Input file format. Each transcript ID and FASTA sequence was paired with a randomly assigned abundance value (1–10) drawn from a normal distribution. (C) Fragmentation process. Full transcript sequences (red bar) were fragmented into simulated reads of varying lengths, based on distributions defined for each category. Fragmentation involved random start positions and generated overlapping reads. (D) Read accuracy categories. Simulated reads were assigned to one of three accuracy categories: 75%, 85%, or 95%. Yellow blocks indicate errors, including substitutions, insertions, and deletions, introduced by the FIC_HMM error model to mimic ONT-specific sequence error patterns.

2.3.1 Isoform Length

Transcript lengths range from short non-coding RNAs to large multi-exon transcripts, reflecting the diversity of the transcriptome (Figure 2.1A). To evaluate how this variability affects isoform discovery, transcripts from the Gencode v45 annotation were sorted into four length categories: <2000 bp, 2000–4000 bp, 4000–6000 bp, and >10,000 bp. From each category, 1000 transcript IDs and their corresponding FASTA sequences were randomly subsampled, ensuring balanced representation of transcript length diversity.

2.3.2 Read Count

To mimic the differences in gene expression and sequencing depth observed in real sequencing data, count values were initially assigned to each transcript using a normal distribution ranging from 1 to 10. This approach captured the natural skew of gene expression, with a lower proportion of highly and lowly expressed transcripts. To then further simulate variations in sequencing depth, these initial count values were multiplied either 10 or 100-fold, creating two additional count categories within each transcript length group (Figure 2.1B).

2.3.3 Read Length

For each group of transcripts of given length and expression/depth, read lengths were modelled with mean read lengths set to 500 bp, 1500 bp, 2500 bp, or 4500 bp, with a

range of ±50 bp around each mean to create a uniform distribution. PBSIM3 sampled read lengths from this distribution, randomly selecting a start position on each transcript and extracting a fragment of the assigned length. This sampling process continued iteratively for each transcript until the total number of bases generated met or exceeded the transcript's nucleotide count multiplied by its assigned count value. For instance, a 4000 bp transcript with a count value of 5 under a 2500 bp read length distribution would generate fragments totalling approximately 5 x 4000 = 20 000 base pairs (Figure 2.1C).

2.3.4 Read Accuracy

Simulated reads were assigned accuracy levels of 75%, 85%, or 95%, with errors introduced. PBSIM3 supports two models for error introduction: nucleotide point errors and structural errors (insertions, deletions, and substitutions). Due to the common occurrence of structural errors in ONT data, a FIC_HMM error model, supplied by PBSIM3, was employed to introduce these errors. The FIC_HMM model was trained on two ONT datasets and introduced substitutions, insertions, and deletions at defined rates, ensuring a context-aware error distribution. For example, homopolymer stretches were more susceptible to deletions, while substitutions occurred more frequently near specific sequence motifs (Ono *et al.*, 2022).

2.4 Genome Alignment Using MiniMap2

In order to perform isoform discovery and quantification, sequencing reads were aligned to the human reference genome (GRCh38), modified to include spiked-in RNA variants, to enable their alignment and quantification. Alignment to the genome, rather than to the transcriptome, was chosen to enable the detection of alternative splicing events by preserving intronic regions and splice junctions in the analysis. Minimap2 v2.28 (Li, 2018) was selected for its efficient alignment capabilities and its support for spliced alignments, a critical feature for long-read RNA sequencing. Splice-aware alignment was enabled using the `-x splice` preset. Table 2.2 summarises the main parameter adjustments made by this preset, relative to the default settings.

Table 2.2: MiniMap2 splice-aware parameters used in this study, compared to the aligner's default settings.

Parameter	Default	Splice Mode
k-mer size (-k)	15	15
Window size (-w)	10	5
Match Score (-A)	2	1
Mismatch Score (-B)	4	2
Gap Open Score (-O)	4, 24	2, 32
Gap Extension (-E)	2, 1	1, 0

2.4.1 Minimizers for Efficient Mapping

Long-read alignment can be computationally intensive, given the substantial data that must be retained for each alignment. To mitigate these demands, Minimap2 employs minimizers to index both reference genomes and query sequences more efficiently, thereby reducing overall memory usage (Li, 2021). Minimizers are representative k-mers derived from a sliding window of sequence fragments. For instance, with a k-mer length of 4 and a window size of 6, all k-mers in the window are generated as depicted in Figure 2.2. The smallest k-mer is selected as the minimizer, and the window shifts one nucleotide to repeat the process. This approach ensures efficient representation of sequences while maintaining computational speed, enabling Minimap2 to identify potential alignment regions without excessive memory use. In this study, a window size of 5 and a k-mer length of 15 were employed. Although longer k-mers can enhance mapping accuracy, they may compromise sensitivity when analysing ONT reads, given the higher error rates associated with this sequencing platform (Li, 2021).

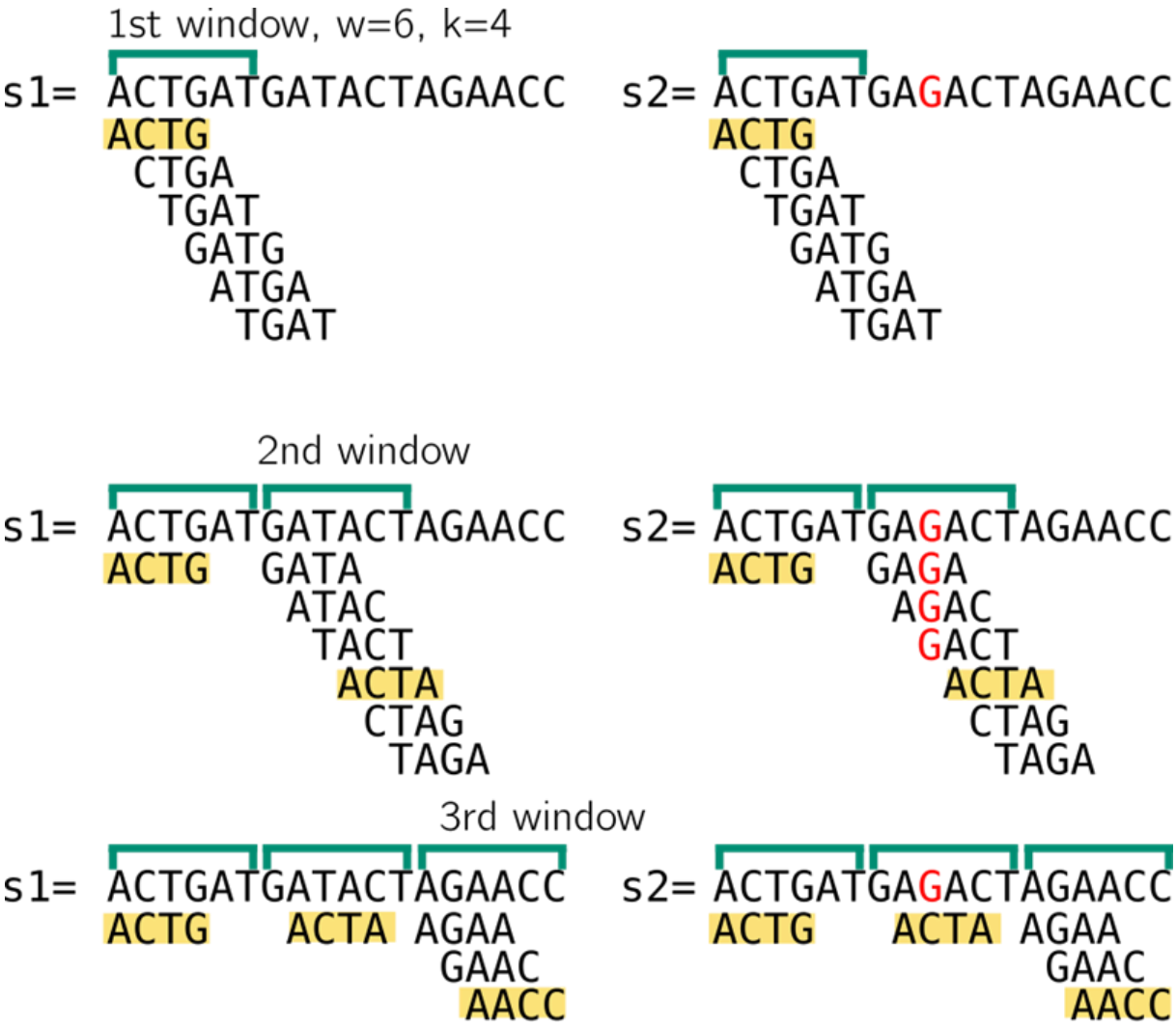


Figure 2.2: Minimizer selection from k-mers. Two sequences (S1 and S2) are broken into 4-mers. Green brackets represent sliding windows. The lexicographically smallest k-mer in each window is highlighted as the minimizer. A mismatch (G) between S1 and S2 can still result in the same minimizer.

2.4.2 Handling Gaps and Spliced Alignment

Minimap2 employs the seed-chain-align algorithm to narrow down alignment locations, enhancing accuracy and performance (Li, 2021). The process begins with identifying seeds, which are minimizers from the query sequence that match those from the reference. Chaining is then performed between these matching seeds, where each seed is evaluated in combination with all preceding seeds (Li, 2021). This step refines anchor alignment by linking each anchor to its predecessor while considering the gap size and number of matching bases. By accounting for indels and

misplaced anchors, this approach ensures robust alignment (Li, 2021). Once optimal chains are established, dynamic programming is used to align the sequences between the anchors. This extension step adjusts for gaps and mismatches, enabling precise calculation of alignment scores and identification of the most accurate alignment paths (Li, 2021).

In handling gaps, Minimap2 uses dynamic programming with a 2-piece affine gap cost to resolve gaps between anchors (Li, 2021). This method assigns higher penalties to shorter gaps than to longer ones, allowing Minimap2 to account for introns and indels commonly found in spliced RNA (Li, 2021). In splice-aware mode, penalties for short gaps and mismatches are reduced, the second gap penalty is omitted and match scores are slightly lowered compared to the default settings (Table 2.2). This adjustment enhances sensitivity in aligning exon structures, trading off some precision (Li, 2021).

To address ambiguous read assignments, Minimap2 classifies all high-scoring alignments as primary alignments and labels overlapping but lower-quality alignments as secondary alignments (Li, 2021). The mapping quality is then calculated by considering the number of anchors in the primary alignment and the quality of the secondary alignments. As a result, reads with fewer anchors or those with high-quality secondary alignments are assigned lower mapping quality scores, effectively filtering out unreliable alignments (Li, 2021).

2.5 Isoform Discovery Tools

Performance of the five isoform-detection tools (Table 2.3) was assessed using two complementary datasets: spike-in RNA libraries produced with diverse preparation protocols and in-silico simulated reads, allowing systematic variation of library type, data quality, and sequence features. For the real-data component, direct cDNA (n = 3), PCR-amplified cDNA (n = 2), and direct RNA (n = 1) samples from the HCT116 cell line were tested, each containing spiked-in RNA as ground truth. The human genome annotation file (GRCh38), the corresponding GTF annotation file (Gencode v45), and the transcriptome reference (Gencode v45) were all used where specified in the tool workflows These datasets provided a basis for comparing each tool's performance across distinct library preparation methods.

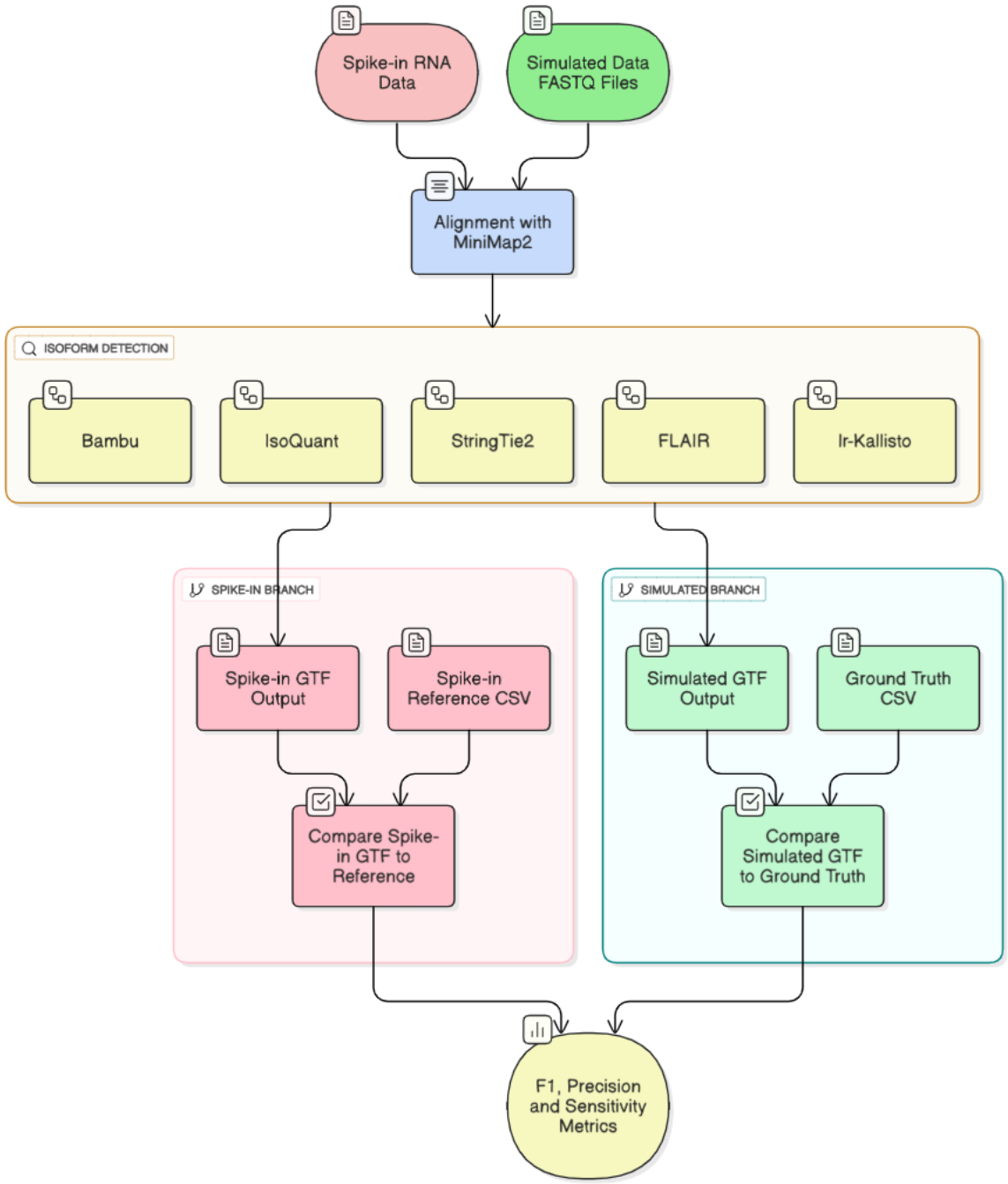


Figure 2.3: Benchmark workflow for evaluating long-read isoform-discovery tools. Spike-in RNA reads (red) and in-silico simulated FASTQ reads (green) are each aligned to the reference genome with Minimap2 (blue). The resulting BAM files are passed in parallel to five isoform-detection pipelines—Bambu, IsoQuant, StringTie2, FLAIR, and lr-kallisto (yellow). For the spike-in branch (left, red panel) each tool's GTF output is compared with a curated spike-in reference annotation, whereas for the simulated branch (right, green panel) the GTF is compared with the known ground-truth transcriptome. Performance is summarised across both branches with F1-score, precision and sensitivity metrics.

To further examine the influence of read length, read accuracy, read count, and transcript length, simulated data was used to assess tool performance by calculating sensitivity, precision, and F1 scores (Chen *et al.*, 2021; Dong *et al.*, 2023; Su *et al.*, 2024). Here, sensitivity (or the true positive rate) measured the proportion of true isoforms that each tool correctly identified. True positives (TP) were defined as isoforms with both observed and expected counts greater than zero, indicating successful detection from the ground-truth isoform list (either from the spike-in references in real data or the known transcript set in simulated data). Conversely, false negatives (FN) were isoforms present in the simulation or spike-in list but undetected by the tool (i.e., observed count of zero despite an expected count above zero). Precision evaluated the ratio of correct predictions among all predicted isoforms, highlighting each tool's ability to avoid false positives (FP). All metrics were calculated using the following formulas:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

To provide a balanced evaluation of performance, the F1 score, calculated as the harmonic mean of sensitivity and precision, was used. This metric is particularly advantageous over the arithmetic mean because it accounts for discrepancies between sensitivity and precision, such as high sensitivity paired with low precision, thereby offering a more comprehensive assessment of tool accuracy. The F1 score is particularly useful in situations where there are extreme differences in values, as it provides a single score that reflects both the tool's ability to correctly identify true positives (sensitivity) and its accuracy in identifying only true positives among all positive identifications (precision). The F1 score was computed using the formula:

$$F1\ Score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity}$$

Table 2.3: Summary of isoform discovery tool methods compared in this study.

Tool	SJ Alignment	Read-Binning	Multi-Mapped Reads	Quantification	Reference
Bambu	10 bp	Reads classified into distinct classes based on read-to-read similarity.	Keeps all multi-mapped reads to improve isoform discovery.	Sums read classes for gene- or transcript-level counts.	(Chen *et al.*, 2023)
IsoQuant	6 bp	Reads grouped to genomic loci on reference genome	Multi-mapped reads are considered for quantification but not for isoform discovery	Counts uniquely consistent reads per isoform and distributes ambiguous reads proportionally.	(Prjibelski *et al.*, 2023)
FLAIR	15 bp	Creates high-confidence isoform assembly using reference genome, then re-aligns reads.	Excludes all multi-mapped reads to ensure the highest accuracy in isoform construction	Tallies raw read counts by realigning all reads to a collapsed transcript reference.	(Tang *et al.*, 2020)
StringTie2	10 bp	Reads grouped to genomic loci on reference genome	If more than 95% of reads at a region are multi-mapped, that region is skipped.	Reports TPM values only (no raw counts), assuming transcripts are full length for length correction.	(Kovaka *et al.*, 2019)
lr-kallisto	None (pseudoalignment)	Groups reads by k-mer (TCC) groups	Multi-mapped reads are kept and included after the EM step, where they help improve quantification	Performs pseudoalignment with an EM algorithm and reports TPM.	(Bray *et al.*, 2016)

2.5.1 StringTie2

StringTie2 (Kovaka *et al.*, 2019) improves upon its predecessor, StringTie (Pertea *et al.*, 2015), which was originally designed for transcriptome assembly using short-read sequences generated by Illumina. StringTie assembled transcripts by combining short reads into longer, super-reads, enabling efficient reconstruction of expressed isoforms. StringTie2 advances this approach to accommodate long-read sequences, such as those generated by ONT and PacBio. The central method underlying both tools is the use of an alternative splicing graph (ASG) (Figure 2.3) and a maximum flow algorithm to iteratively find the most probable transcript paths within the graph. For long-read data, StringTie2 incorporates alignment corrections and modified parameters to handle the lower-confidence splice junction assignments often encountered in these datasets.

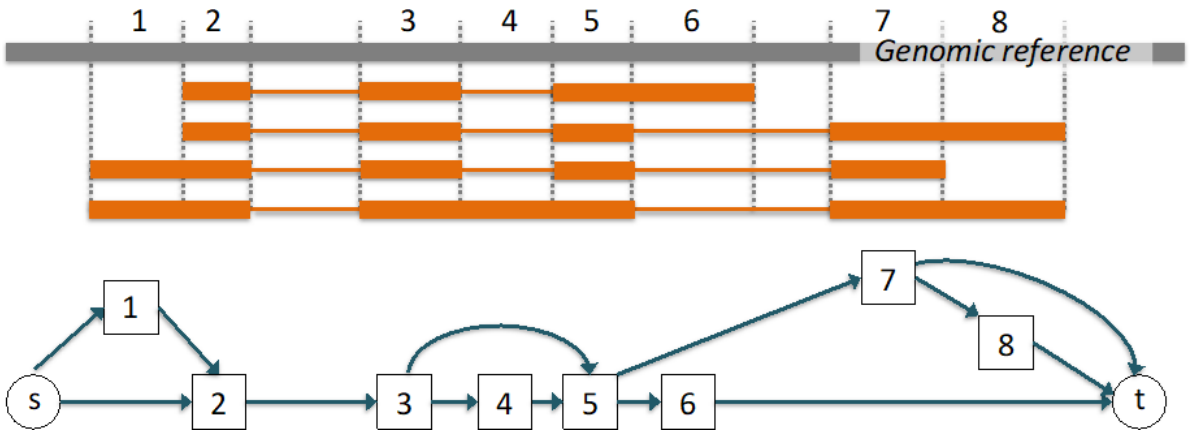


Figure 2.4: Overview of alternative splicing graph assembly and the application of the maximum flow algorithm by StringTie2. Aligned reads (orange) correspond to the numbered nodes (numbered blocks), with arrows indicating the possible paths between exons. Source (s) and Sink (t) represent the start and end of the path that the algorithm iterates through. Each line represents an alternative path that can be taken by the algorithm when resolving the combinations of nodes. Adapted from Kovaka *et al.*, 2019.

StringTie2 processes each gene locus individually by grouping reads aligned to a region for ASG creation (Figure 2.3). Splice junctions are classified as low-confidence if the supporting read has an alignment error rate exceeding 1%. Such splice junctions require 25% more supporting reads to be retained in the ASG without correction. If this

threshold is not met, low-confidence splice sites are corrected by aligning them to the nearest annotated splice junction within 10 base pairs. This alignment correction step improves the accuracy of splice site assignments and ensures more reliable transcript assembly. Following alignment correction, StringTie2 constructs an ASG for each region, using all reads aligned to the locus. Exonic regions are represented as nodes in the graph, while splice junctions are represented as edges. Coverage values are assigned to each node based on the number of reads supporting the corresponding exonic region (Kovaka *et al.*, 2019).

The maximum flow algorithm is applied to the ASG to identify the path with the highest read coverage, starting from the node with the highest count and extending through compatible paths to the source (s) and sink (t) (Figure 2.3). This path represents the primary transcript for the gene. StringTie2 then iteratively removes reads supporting this transcript, updates the node coverage values, and repeats the process to identify additional paths corresponding to alternative transcripts. By allowing separate reads from the same isoform to contribute to the same path, this method ensures accurate reconstruction of isoforms through the merging of reads into unified transcript paths (Kovaka *et al.*, 2019).

StringTie2 was run with the -L parameter to enable long-read mode using the GTF annotation. StringTie2 reports transcript abundance estimates as Transcripts per Million (TPM), under the assumption that assembled reads represent full-length transcripts and therefore do not require length normalisation. Raw count values are not reported, which precluded StringTie2 from inclusion in quantification benchmarking analyses.

2.5.2 Bambu

Bambu (Chen *et al.*, 2023) was run with default parameters, with isoform discovery and quantification enabled, using the annotation and reference genome. Bambu employs machine-learning algorithms trained on sequencing data to improve the sensitivity of splice-junction corrections, distinguishing it from most isoform tools that rely on strict boundary rules. Instead of directly correcting reads to an annotation, Bambu groups similar reads into "read classes" and assigns the read classes to known transcripts in the annotation (Figure 2.4). Alignment error correction in Bambu begins

by training a decision tree-based machine-learning model using all reads in the sample to correct misplaced splice junctions. This model evaluates the distances between read and annotated splice junctions, strand information, and read support for misplaced junctions to develop a sample-specific error profile. Unlike rigid boundaries that may miss novel splice sites, Bambu's approach enables adaptable corrections across sequence qualities. Splice junctions are categorised as either high-confidence or low-confidence based on model predictions. High-confidence junctions remain unchanged, while low-confidence ones are corrected to the nearest reference splice junction within a 10 bp threshold, unless no reference junction is nearby, in which case the alignment is left unaltered.

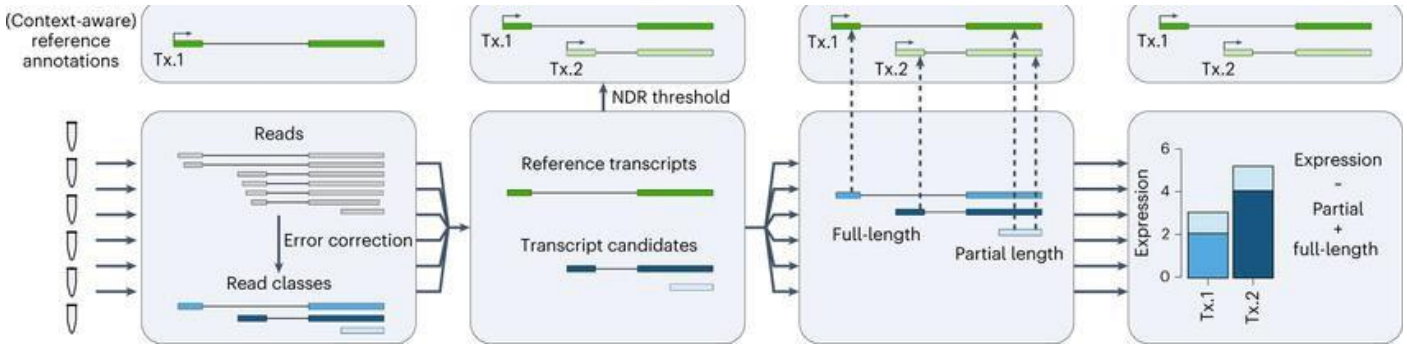


Figure 2.5: Overview of the read class creation and assignment process used by Bambu. Error correction of reads leverages reference annotations (green) to predict and correct splice junctions. Reads are grouped into read classes (blue) based on shared splice junction patterns, independent of the reference. These read classes are subsequently assigned to transcripts, which are then used collectively for quantification. Adapted from Chen *et al.,* 2023.

After performing alignment corrections, Bambu groups reads with identical splice junction patterns into "read classes" (Figure 2.4). These classes represent sets of reads that share consistent structural features. Each read class is then matched to transcripts that overlap by at least 35 base pairs (bp) for every exon, ensuring alignment accuracy prior to quantification. To evaluate whether a read class corresponds to a valid transcript, Bambu employs multiple predictive features. One key feature is the number of reads assigned to the transcript, with higher read counts

interpreted as stronger evidence of validity. To mitigate biases introduced by fragmented or degraded reads, Bambu calculates the gene proportion, defined as the ratio of reads assigned to the read class relative to the total number of reads for the gene. Further, Bambu analyzes the transcription start site (TSS) distances among reads within the same class to detect inconsistencies. It also examines strand proportions to address sequencing artifacts, such as reverse transcription biases. Lastly, Bambu evaluates the abundance of A/T-rich regions near the endpoints of reads, as these sequences are prone to causing transcription slippage and producing truncated reads (Chen *et al.*, 2023).

Bambu performs quantification at both the gene and transcript levels. For transcript-level quantification, read classes are assigned to a transcript if they share a continuous set of splice junctions within a 10 bp threshold. Read classes failing this criterion are assigned to the gene level. Gene-level quantification requires at least a 35 bp overlap between the gene and exon. All read classes are used in the final quantification.

2.5.3 IsoQuant

IsoQuant (Prjibelski *et al.*, 2023) employs two complementary approaches for transcript assembly, (1) assigning reads to known transcripts and (2) constructing intron graphs from unassigned reads to detect novel isoforms. Each gene is processed individually by extracting all reads aligning to its genomic region from the BAM file. This approach ensures that reads are processed in the context of specific gene boundaries, allowing for more precise isoform identification. IsoQuant uses two graph structures for assigning reads to isoforms: the splice junction graph (SJG) and the exon graph (EG) (Figure 2.5). Splice junctions and exons from each annotated isoform are assigned unique numbers based on their genomic position. For each read, a vector is created to capture its alignment to these features. A value of 1 indicates that a read's splice junction or exon matches the annotation, -1 signifies that the read skips a splice junction present in the annotation, and 0 indicates that the read does not reach a particular splice junction. Reads are marked as consistent with an isoform if the combined vector profile sums to zero, indicating alignment with all expected features. Ambiguous reads, which match multiple isoforms, are flagged accordingly.

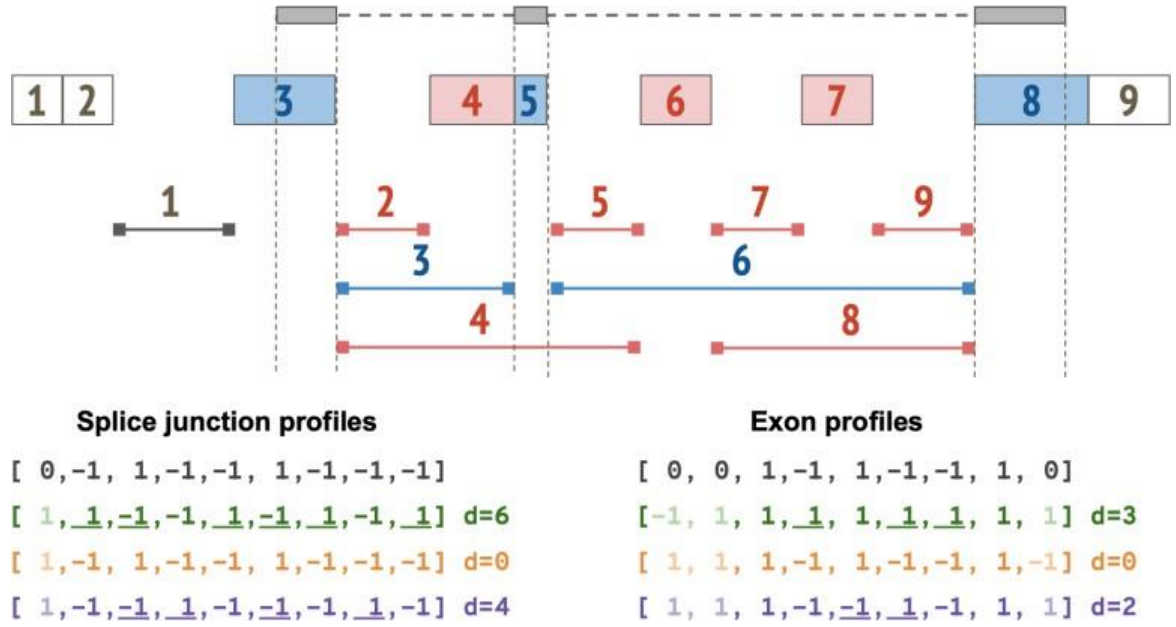


Figure 2.6: Overview of the splice junction graph (SJG) and exon graph (EG) approaches for assigning reads to known transcripts used by IsoQuant. The aligned read (grey) maps to the gene model represented by numbered exons (boxes) and splice junctions (lines connecting exons). Each splice junction is numbered and listed from A to C below the gene model. Splice junction profiles (SJG) are depicted as lines spanning the gaps between exons, while exon profiles (EG) capture exon alignments. The read profile (top row) is compared against annotated isoform profiles (subsequent rows) for both SJG and EG. The distance (d) between the read profile and each isoform profile is calculated and noted, with consistent reads having a distance of 0. This systematic comparison determines whether the read aligns uniquely, ambiguously, or inconsistently to known isoforms. Adapted from Prjibelski *et al.*, 2023.

Inconsistent reads that fail to align within six base pairs of annotated splice junctions, are used to construct transcript models (Prjibelski *et al.*, 2023). This is under the assumption that reads that are inconsistent with the annotation could be novel and thus require novel isoform discovery techniques. IsoQuant creates intron graphs for novel isoform detection using the same node-and-edge principles as StringTie2, where exonic regions are represented as nodes and splice junctions as edges. Unique and ambiguous consistent reads are excluded from transcript model construction but serve as quality controls for annotated splice junctions. To filter out low-confidence paths, IsoQuant requires a minimum of five full-length supporting transcripts for novel

isoforms to be reported. For quantification, only reads with unique and consistent mappings were used, with each read counted as a single transcript. Ambiguous reads were proportionally distributed across all compatible isoforms. Reads marked as inconsistent were excluded from the quantification process but flagged as potential contributors to novel isoform detection. This methodology balances accuracy in assigning reads to known isoforms with flexibility in detecting novel splicing patterns.

IsoQuant takes a sqlite3 database (db) file as an annotation file for “complex manipulation of hierarchical features” such as transcript and exon relationships (Dale, 2025). Although IsoQuant offers the ability to automatically convert GTF to db, duplicated exon names in the spiked-in dataset required custom db building (as described in https://github.com/Kikking/nano-pipe) using gffutils v0.13 (Dale, 2025). IsoQuant was run using both the reference genome and db annotation file and specifying the datatype as “nanopore”.

2.5.4 FLAIR

FLAIR (Tang *et al.*, 2020) consists of several modular steps designed for long-read transcriptome analysis, beginning with read alignment and extending to applications like differential splicing analysis. For the purposes of isoform discovery in this study, only two modules of FLAIR were utilized: Correct and Collapse (Figure 2.6). During the alignment correction step, BAM files generated with MiniMap2 were converted into BED12 format using BedTools v2.31.0 (Quinlan and Hall, 2010), as required by the FLAIR-correct module. In this module, read alignments were adjusted to the nearest splice junction in the reference genome within a 15 bp window (Figure 2.6, red dots). These corrected reads were then passed to the FLAIR-collapse module for further processing.

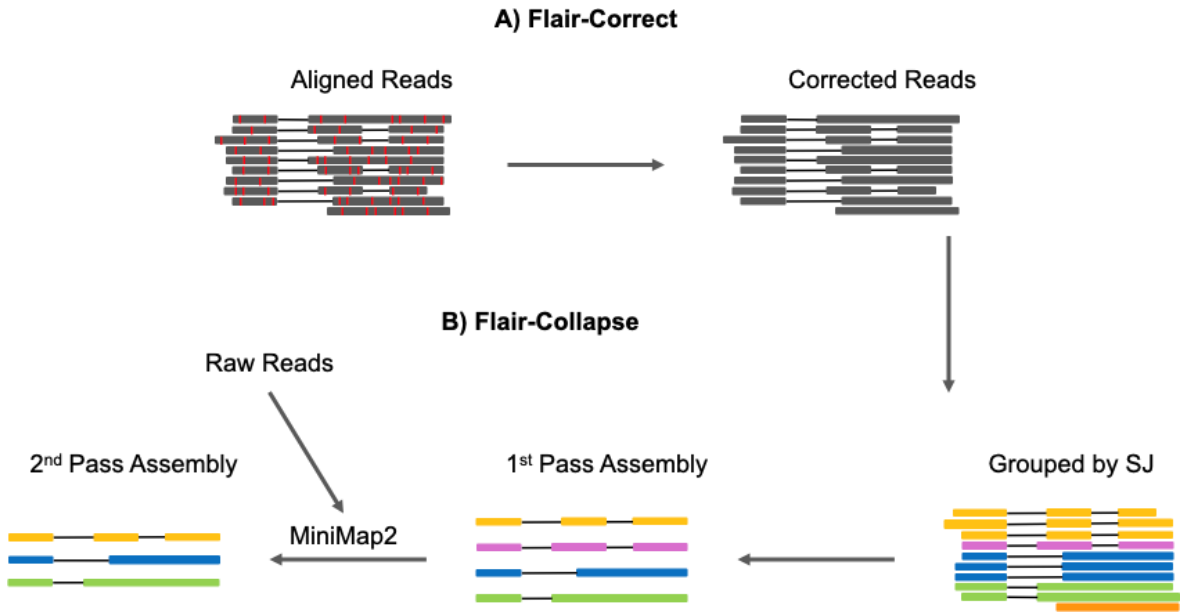


Figure 2.7: Overview of the steps involved in the FLAIR-correct and FLAIR-collapse modules. (A) Reads aligned with MiniMap2 are corrected within 15 bp of the nearest reference splice junction. (B) In the FLAIR-collapse module, corrected reads are grouped by shared splice junctions, and a first-pass assembly is generated using high-confidence transcripts. Subsequently, all raw reads are realigned to the new assembly using MiniMap2, producing a high-confidence second assembly. Adapted from Tang *et al.*, 2020.

A unique feature of FLAIR, implemented during the collapse step, is the creation of a high-confidence reference assembly using high-quality corrected reads, followed by the re-alignment of raw reads to this assembly (Tang *et al.*, 2020). Initially, corrected reads sharing splice junctions were grouped to identify candidate isoforms. To refine the isoform discovery process, transcription start sites (TSS) and transcription end sites (TES) were determined within 100 bp windows of read coverage, selecting the most frequent start and end positions within each window. This approach reduces the impact of truncated reads on the assignment of transcription boundaries. From these groups, a single isoform was selected to represent each cluster of reads, forming a first-pass transcriptome consisting of high- confidence isoforms (Figure 2.6). Raw, uncorrected reads were then realigned to this first-pass transcriptome using MiniMap2 (Figure 2.6). To maximise sensitivity and enable the detection of isoforms with low read counts in the simulated dataset, the minimum number of supporting reads was set to one. Additionally, the recommended `--stringent` flag was omitted, as it is optimised for full-length ONT reads and would have excluded the fragmented reads typical of the simulated dataset.

For quantification, FLAIR reports transcript-level counts by tallying all unique read assignments (i.e., MAPQ > 1) from the alignment to the reference assembly (the first pass assembly using sample data). The FLAIR-quantify module was not utilised in this analysis, as it normalises count values across multiple samples, which was unnecessary for this single-sample per condition dataset. Instead, raw count values were reported directly to facilitate consistent comparison across tools.

2.5.5 lr-kallisto

Kallisto's algorithm (Bray *et al.*, 2016), initially designed for short-read data, has been adapted for long-read sequencing, resulting in long-read Kallisto (lr-Kallisto) (Loving *et al.*, 2024). Unlike other long-read transcript discovery methods that rely on MiniMap2 and a reference genome, lr-Kallisto applies pseudoalignment using a reference transcriptome, similar to its short-read approach. Key enhancements for long-read sequencing include increasing the k-mer length from 31 bp to 63 bp, which greatly improves accuracy, especially for low-quality data with high error rates like those seen in ONT.

First, the reference transcriptome was indexed using lr-Kallisto's index command. Then lr-Kallisto was run in quant mode with *–long* and *-P ONT* to enable long-read isoform quantification. In lr-Kallisto's pseudoalignment process, the reference transcriptome is broken into overlapping 63 bp k-mers. These k-mers are then grouped and ordered by their shared transcript origins. Each k-mer is associated with a compatibility set—a list of transcripts it maps to. To align each read, lr-Kallisto breaks the read into overlapping 63 bp k-mers and intersects the compatibility sets from each k-mer in the read. This intersection forms the Transcript Compatibility Class (TCC), which represents the transcripts most compatible with the read sequence. If this intersection is empty, as might occur with low-quality sequences or novel isoforms, lr-Kallisto assigns the most common compatibility set as the read's origin (Loving *et al.*, 2024).

To adjust for lower coverage of longer transcripts, lr-Kallisto calculates effective transcript length as the sum of all read lengths aligned to the transcript, divided by the number of aligned reads, minus the k-mer length. For quantification, Kallisto employs an Expectation Maximization (EM) algorithm, with slight modifications: multi-mapping

reads are initially distributed uniformly across all possible transcripts. After running the EM iterations, uniquely mapped reads are added to the estimated abundance, reducing the influence of ambiguously assigned multi-mapped reads on the final abundance value.

2.6 Data Availability

All analysis scripts and corresponding file metadata are available on GitHub at https://github.com/Kikking/nano-pipe. The Hct116 data samples treated with spiked-in RNA can be accessed via the Amazon AWS S3 bucket (arn:aws:s3:::sg-nex-data), and the Mock and IFN- β-treated A549 and HT1376 samples are available from SRA under accession PRJNA743928. Reference annotation files can be found at the Gencode website https://www.gencodegenes.org/human/release_45.html.

3. RESULTS

3.1 The Effect of Basecaller Selection on Sequence Data Quality

To determine whether ONT signal files generated using older sequencing chemistries could benefit from updated basecalling algorithms, we compared the performance of Dorado with its predecessor, Guppy (Figure 3.1) using direct cDNA data derived from the Hct116 cell line (n = 3), sequenced as part of the SGNEX project. Dorado produced a broader distribution of read lengths than Guppy across all samples analyzed, suggesting an improved ability to resolve longer sequences. In addition, Dorado consistently yielded higher accuracy reads, demonstrating a clear improvement in overall quality of the sequencing data obtained. In light of these findings, Dorado v0.4.1 was used for basecalling the sequencing data used in all subsequent analyses.

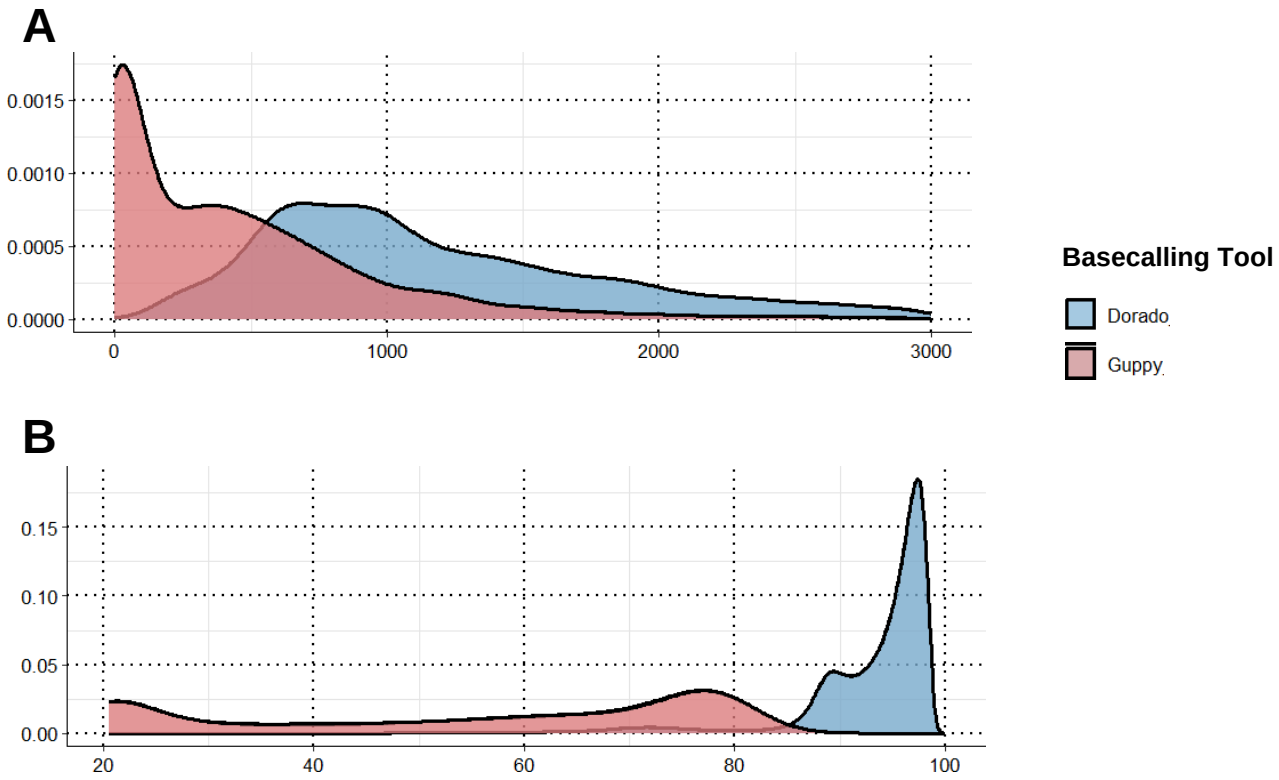


Figure 3.1: Read length and read accuracy distribution comparison between Guppy and Dorado. The density plots illustrate the distribution of A) read lengths (measured as the number of base pairs per read) and B) read accuracy (quantified on a Phred score scale) obtained when performing basecalling using either Guppy (Red) or Dorado (Blue).

3.2 The Relationship Between Library Type and Sequencing Data Quality

In order to evaluate the quality and quantity of the sequencing data obtained with the three ONT RNA library-preparation strategies, the read-length, read-accuracy, and read-count distributions were assessed. This analysis focused on direct cDNA (n = 3), PCR-amplified cDNA (n = 2), and direct RNA (n = 1) libraries from HCT116 cells sequenced in the SGNEX project (Figure 3.2).The PCR- amplified cDNA libraries were found to exhibit the shortest average read length and the lowest read quality across the three library types, though these samples had read counts similar to the direct cDNA library samples. In contrast, the single direct RNA

library yielded shorter reads than the direct cDNA libraries, but had comparable read accuracy. However, the direct RNA library sample contained fewer reads in total, compared to the other library type samples.

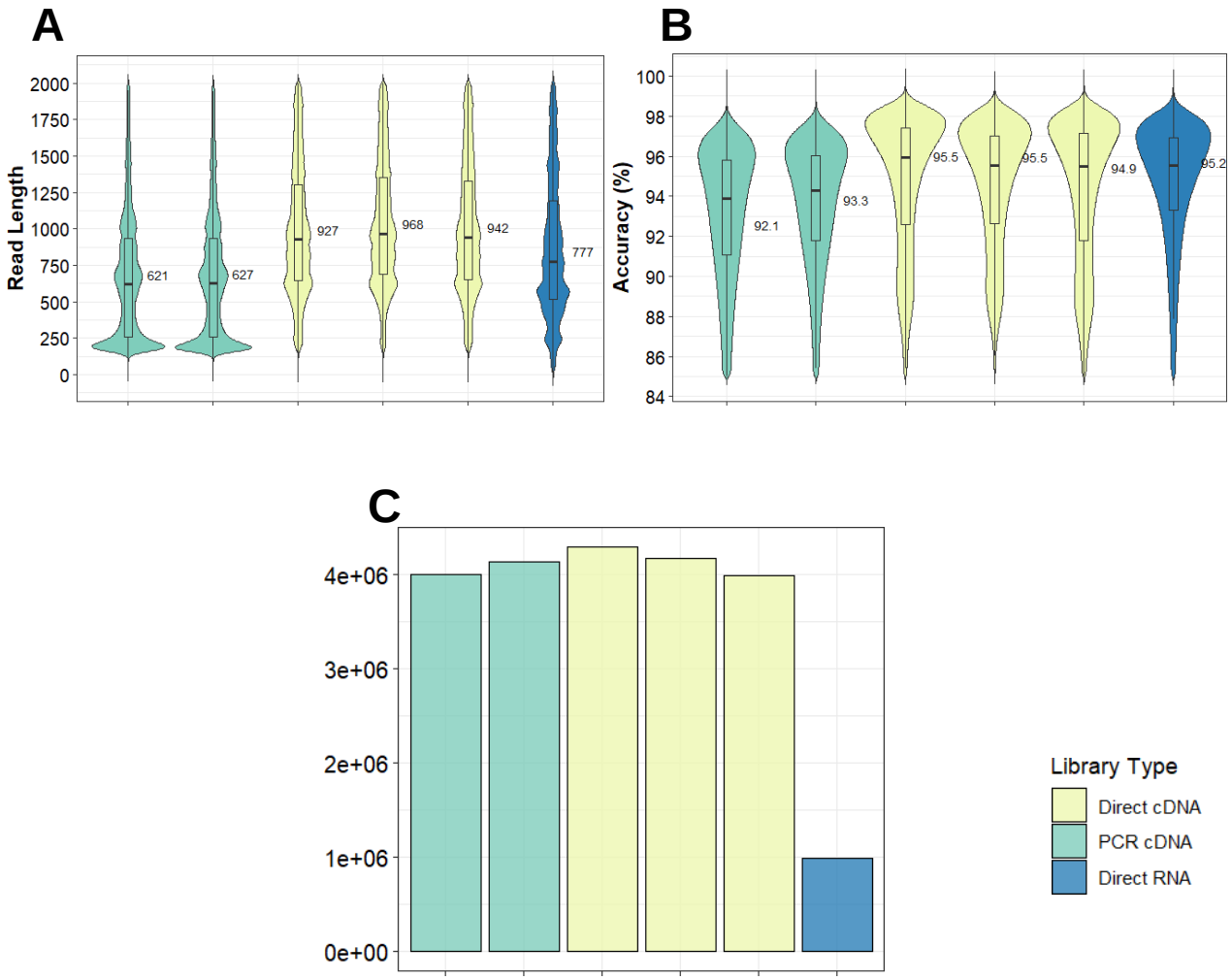


Figure 3.2: Read length, read accuracy and read count distributions of three different ONT RNA library preparation techniques. The (A) read length, (B) read accuracy, and (C) read counts distributions for data derived from direct cDNA (yellow; n = 3), PCR-amplified cDNA (green; n = 2), and direct RNA (blue; n = 1) libraries are shown.

3.3 Evaluation of Tools for Transcript Discovery Across Different Library Types

The ability of five isoform-discovery tools—Bambu, FLAIR, IsoQuant, lr-kallisto, and StringTie2—to recover known spiked-in transcripts was assessed. The evaluation used HCT116 datasets from the SGNEX project generated with three library preparations: direct cDNA (n = 3), PCR-amplified cDNA (n = 2), and direct RNA (n = 1). For each tool, precision, sensitivity, and the F1 score (the harmonic mean of precision and sensitivity) were calculated. When comparing the balance of precision and sensitivity across the tools, Bambu and IsoQuant emerged as the best-performing tools when using data derived from direct cDNA libraries (Figure 3.3). For PCR amplified cDNA libraries, the sensitivity and precision of Bambu and IsoQuant were lower compared to direct cDNA libraries, such that they showed comparable performance with lr-kallisto. All tools performed poorly when resolving spike-in transcripts from direct RNA libraries.

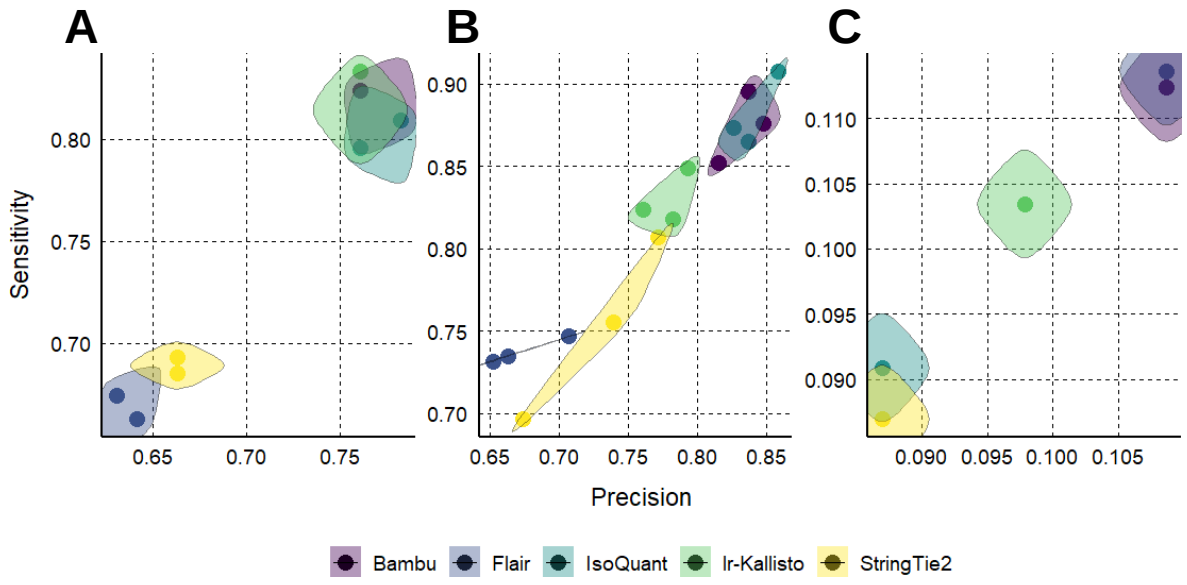


Figure 3.3: The precision and sensitivity scores of each isoform detection tool against different ONT library types. The performance of Bambu (purple), FLAIR (dark blue), IsoQuant (light blue), lr-kallisto (green) and StringTie2 (yellow) when using data derived from (A) PCR- amplified cDNA (n = 2), (B) Direct cDNA (n = 3), and (C) direct RNA (n = 1) libraries is shown.

The F1 scores for all tools were highest when using direct cDNA libraries, indicating superior recovery of spike-in transcripts from this library type (Figure 3.4). In contrast, the F1 scores for all tools was consistently below 0.2 when using data derived from the single direct RNA library. Notably, lr-kallisto was the most robust method with respect to cDNA libraries, displaying the smallest performance gap between PCR-amplified and direct cDNA libraries.

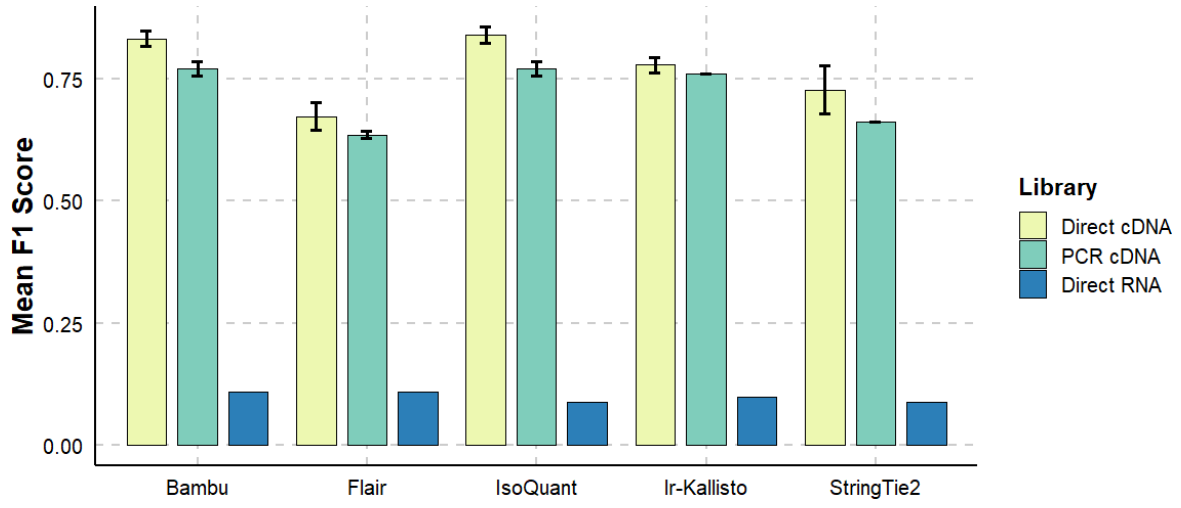


Figure 3.4: Mean F1 scores of different isoform detection tools used against different ONT library types. The mean F1 score for each tool when using data derived from direct cDNA (yellow; n = 3), PCR-amplified cDNA (green; n = 2), and direct RNA (blue; n = 1) libraries is shown.

3.4 Evaluation of Tools for Transcript Discovery Using Simulated Data

While useful in some respects, spiked-in RNA transcripts provide a limited perspective on the nuanced challenges of isoform discovery. First, spike-in RNA represents only a small fraction of the overall dataset, limiting the ability to capture the full diversity of isoforms present in real biological samples. Secondly, spike-in transcripts are synthetic, resulting in relatively low complexity (with respect to alternative splicing events and sequence diversity) compared to cellular RNA. Consequently, while spike-

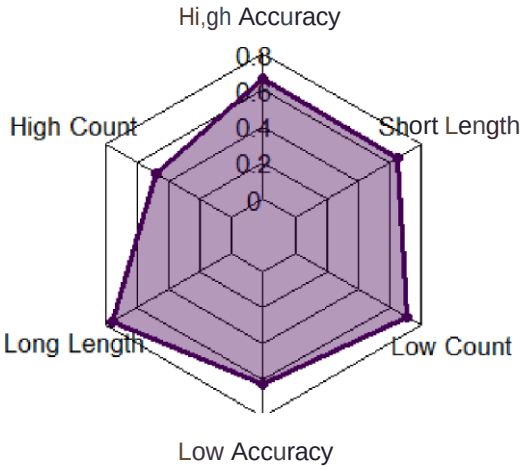
in RNA is useful for initial tool benchmarking, it may not fully reflect a tool's performance in complex biological contexts. Therefore, to comprehensively evaluate how sequence quality and isoform diversity influence the performance of five isoform-discovery tools, 96 RNA datasets were simulated with PBSIM 3.04. Key parameters—read length, transcript length, read accuracy and read count—were varied to create controlled datasets with known ground-truth isoforms. For each tool, precision, sensitivity, and the F1 score (the harmonic mean of precision and sensitivity) were calculated. By systematically altering each simulatedparameter, each tool's ability to accurately detect isoforms under diverse sequencing scenarios could be directly measured.

3.4.1 Sequence Quality

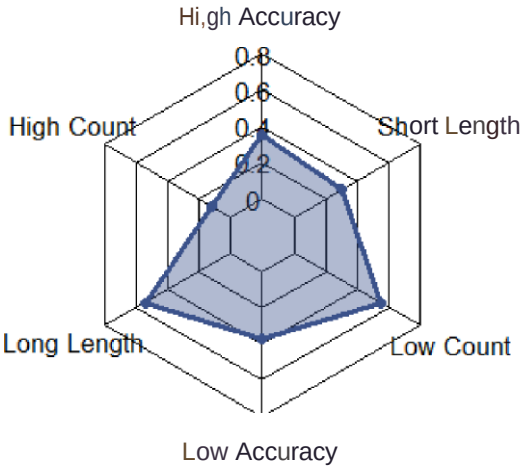
To compare isoform discovery tool performance under a broad range of sequencing conditions, we evaluated two "extreme" parameter sets (75% vs. 95% accuracy, 500 bp vs. 4,500 bp read length, and read count of 10 vs. 100) and classified these parameters into categories. By placing each tool's performance in these discrete bins, we could systematically chart how effectively each tool adapted to challenging (lower-quality) versus more ideal (higher-quality) scenarios (Figure 3.5).

IsoQuant was the most robust to different sequencing conditions, with only a minor drop in performance at higher read counts (Figure 3.5). Bambu showed a similar performance, but with lower overall scores compared to IsoQuant. FLAIR performed better with longer read lengths and lower read counts, but was unaffected by read accuracy. StringTie2 showed very high performance with longer, more accurate and more abundant reads compared to lower quality data, while lr-kallisto performed best with high accuracy reads, but performed very poorly with lower accuracy reads. Longer reads and lower count values were more suitable for lr-kallisto's performance.

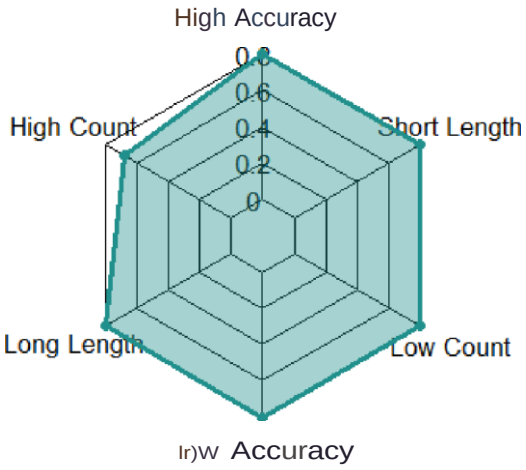
Bambu



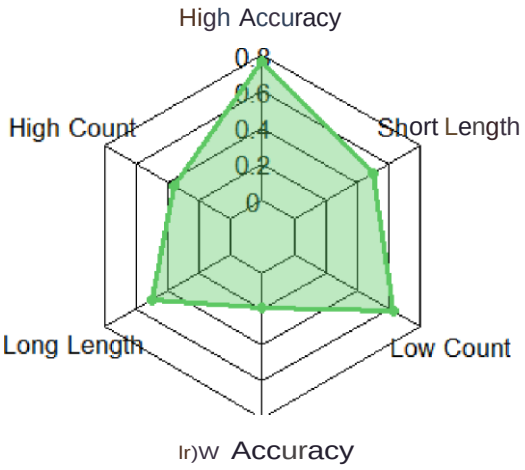
Flair



IsoQuant



Lr-Kallisto



StringTie2

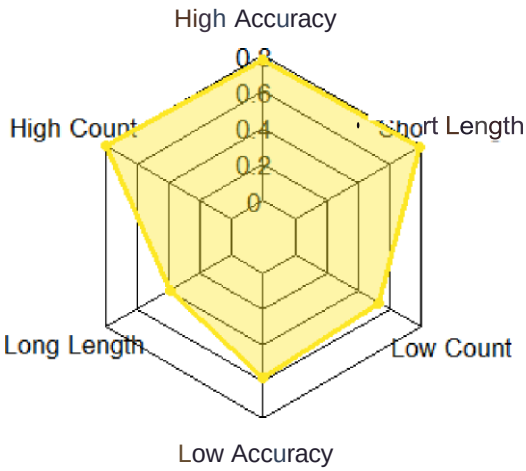


Figure 3.5: The performance of isoform detection tools across the key simulated parameters. Radar plot illustrating the performance of five isoform detection tools – Bambu (purple), FLAIR (dark blue), IsoQuant (light blue), lr- kallisto (green) and StringTie2 (yellow) – across six categories: short length, long length, low accuracy, high accuracy, low count, and high count. Performance was measured using F1 scores, with higher values indicating better performance. Categories are based on simulated variables: low accuracy for high sequencing errors, high accuracy for low errors, low count for lower transcript abundances, and high count for higher abundances.

3.4.2 Transcript and Sequencing Read Length

The transcriptome exhibits a broad diversity in isoform lengths, adding complexity to transcript reconstruction as both transcript lengths and the fragment sizes sequenced vary considerably. According to GENCODE v45, human mRNA transcript lengths span a wide range, from short (<2000 bp) to very long (>10 000 bp), with a median transcript length of 1800 bp and a substantial number (28 000) of transcripts under 2000 bp. Given this breadth, it was important to assess the both impact of transcript length on tool performance and the effect of read length on reconstructing these different sized isoforms.

To assess the impact of read length on isoform reconstruction, the precision and sensitivity of isoform discovery tools was measured across different transcript lengths. All five tools were seen to vary with respect to their effectiveness at reconstructing transcripts depending on both transcript and sequencing read length (Figure 3.6). StringTie2, lr-kallisto and FLAIR all showed poor resolution of transcripts longer than 10 000 bp and also had provided insufficient reconstruction of isoforms when read lengths were under 2000 bp, as seen by both poor sensitivity and precision in these ranges (Figure 3.6A and Figure 3.6B). In contrast, Bambu consistently demonstrated high sensitivity across all transcript lengths and fragment sizes, although it displayed reduced precision when reconstructing shorter transcripts (<2000 bp) and when faced with shorter read lengths (500 bp). IsoQuant demonstrated the best performance with respect to both precision and sensitivity in the short transcript range.

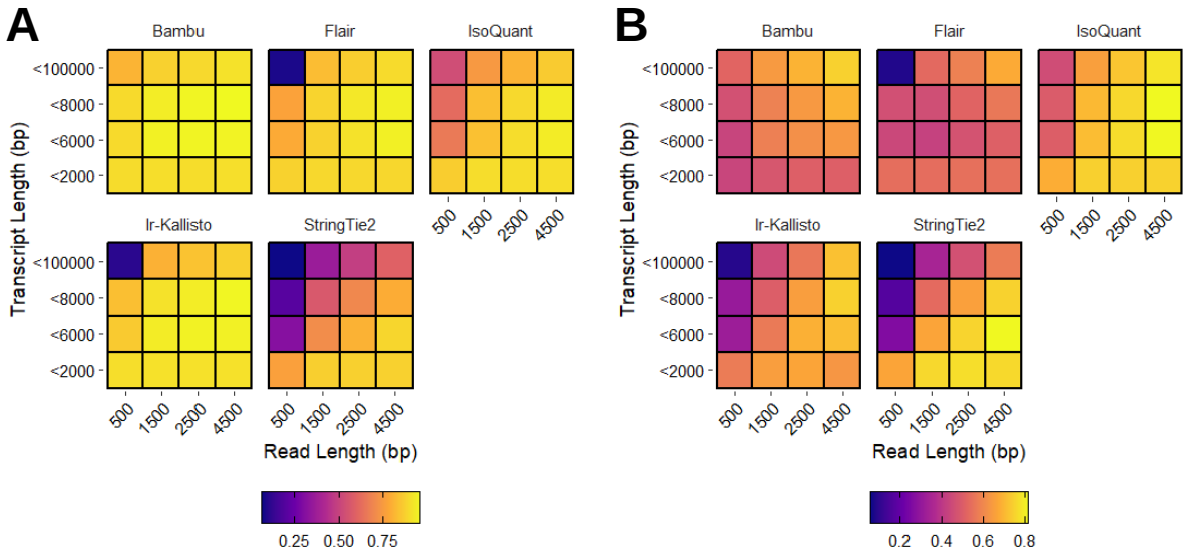


Figure 3.6: Heatmap of Precision and Sensitivity Scores for Isoform-Detection Tools Across Simulated Transcript and Read Lengths. The heatmaps show (A) the sensitivity scores and (B) the precision of each isoform discovery tool at varying transcript and sequencing read lengths. Sensitivity and precision are scored from 0 to 1, with blue indicating lower values and yellow indicating higher values. The y-axis denotes transcript length categories, while the x-axis represents simulated read lengths.

3.5 Evaluation of Tools for Transcript Quantification Using Simulated Data

In most experimental scenarios, the goal is not only to identify transcripts, but also to quantify changes in isoform expression across conditions, i.e., to identify instances of differential transcript usage. Therefore, the accuracy of quantification estimates for five tools (Bambu, FLAIR, IsoQuant, lr-kallisto and StringTie2) was also evaluated by assessing how effectively they estimated isoform counts within 10%, 20%, or 30% of the true simulated transcript counts. While increased counts did not dramatically affect quantification accuracy for any of the tools, read length was an important factor (Figure 3.7). Among the tools, Bambu consistently delivered the highest accuracy, estimating counts within 10% of the true value more frequently than other tools, which generally achieved this level of accuracy only within the 30% threshold. This underscored Bambu's robustness in producing accurate abundance estimates across varying read lengths.

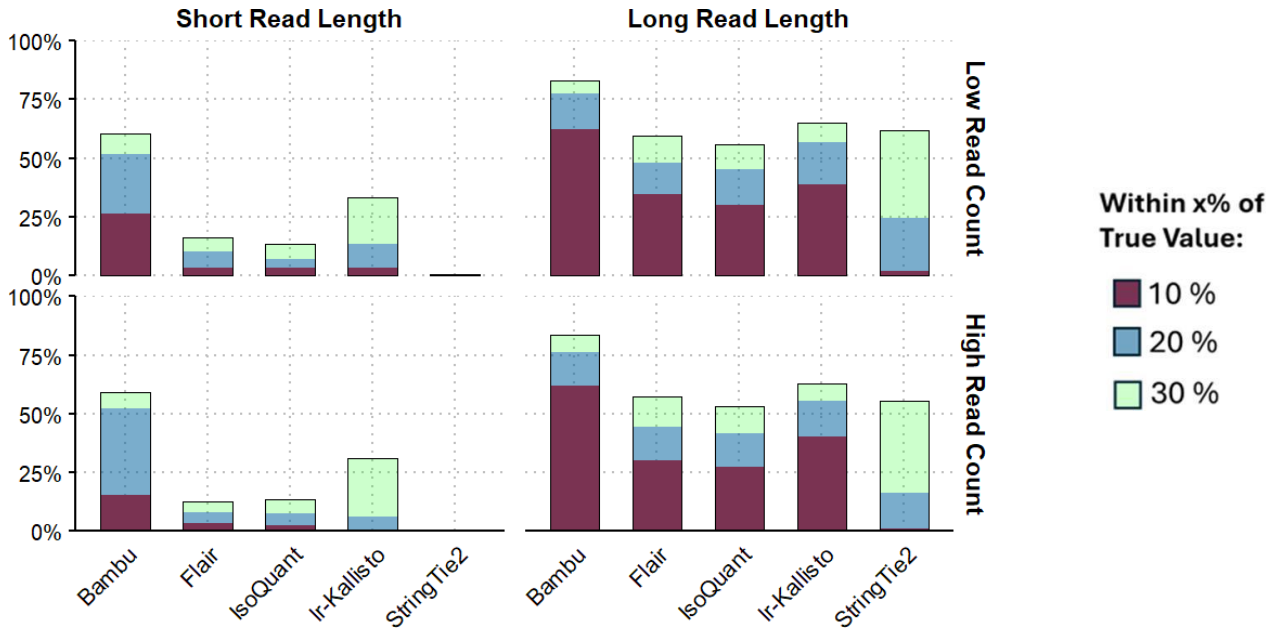


Figure 3.7: Proportion of Transcripts Quantified Within 10 %, 20 %, and 30 % of Ground-Truth Abundance. The figure shows accuracy of isoform detection tools across increasing read count (top to bottom) and read length (left to right). The y-axis shows the proportion of observed values that fall within 10%, 20%, and 30% of the true value.

3.6 RLR Transcript Discovery and Quantification

In light of the findings with respect to tool performance, Bambu was used to assess isoform expression patterns of the RLRs RIG-I (*RIGI*) and MDA5 (*IFIH1*) in A549 and HT1376 cells treated with and without IFN-β (Figure S2). Both genes were subject to alternative splicing and underwent changes in isoform expression patterns in response to IFN-β treatment (Figure 3.8). The primary RIG-I isoform (RIGI-202) showed minimal expression in mock-treated A549 and HT1376 cells but markedly increased once IFN-β was applied. By contrast, a splicing variant (RIGI-204) appeared exclusively in A549 cells after IFN-β treatment. For MDA5, both the dominant isoform (IFIH-206) and an alternative isoform (IFIH-208) were detected in both cell lines under IFN-β treatment.

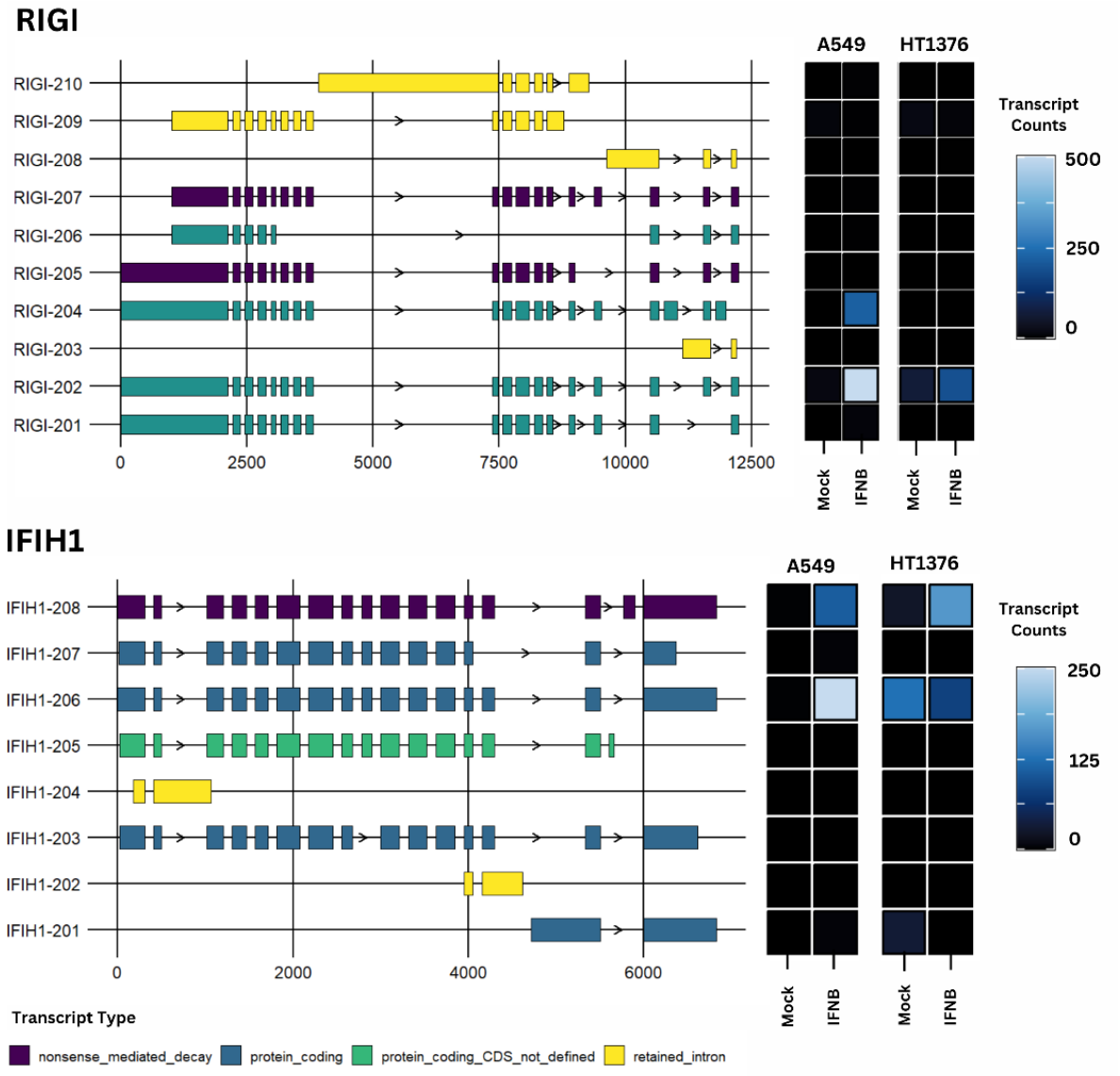


Figure 3.8: The expression profiles of RIG-I and MDA5 isoforms in A549 and HT1376 cells before and after treatment with IFN-β. The heatmap shows expression in IFN-β- and mock-treated A549 and HT1376 cells, with black representing no expression and blue indicating expression (increasing from dark to light). Transcript structures were obtained using Gencode v45.

4. DISCUSSION

This study evaluated long-read sequencing tools to develop an accurate isoform identification pipeline for human ONT data, specifically to quantify the alternative splicing landscape of RLRs following IFN-β treatment. We demonstrated that basecalling software greatly impacts read length and accuracy, influencing downstream isoform detection. Using simulated data, we highlighted the strengths and limitations of different isoform discovery tools, depending on sequencing quality. Based on this evaluation, Bambu was selected as the most suitable tool for this dataset, though other tools may be preferable in different contexts. Our analysis revealed alternative isoforms of both MDA5 and RIG-I, suggesting alternative splicing likely plays a role in regulating the function of these proteins.

In terms of basecalling, Dorado represents a major leap forward for ONT, delivering both a substantial performance boost and marked improvements in overall sequencing quality. In this study, Dorado consistently outperformed its predecessor, Guppy, in terms of sequence accuracy, a crucial factor for isoform detection tools that depend on precise splice-junction information (Prjibelski *et al.*, 2023). Moreover, Dorado produced longer average read lengths, likely owing to more effective signal decoding rather than prematurely terminating reads. A core innovation underlying Dorado's performance is its use of a CRF-based decoder, in contrast to the CTC model employed by Guppy. CRF-based approaches can outperform CTC methods by considering adjacent base probabilities, thereby mitigating systematic errors such as insertions and deletions (Pagès-Gallego and de Ridder, 2023). Although CRF decoding is computationally demanding, Dorado's GPU-optimized architecture allows it to handle this complexity while maintaining high throughput and accuracy. From the perspective of isoform discovery, these advances suggest that re-basecalling archived datasets with Dorado offers not only an immediate improvement in data quality but also deeper and more reliable insights during downstream analyses.

While advancements in basecalling can mitigate some issues, addressing the quality differences introduced during library preparation (PCR- amplified cDNA, Direct cDNA or Direct RNA) remain pivotal. PCR-amplified cDNA libraries exhibited lower isoform detection accuracy than direct cDNA libraries. This discrepancy likely stems from shorter read lengths and reduced basecalling accuracy in PCR-based data, which obscures splice junctions and impedes full-length transcript capture. Additionally, because shorter reads load and translocate more efficiently on Nanopore flow cells, their predominance in a PCR-amplified cDNA pool causes longer fragments to be sequenced even less frequently, further amplifying the read-length gap in that library (Jain *et al*. 2018).Somewhat ironically, while PCR aims to increase coverage (particularly for low-abundance transcripts), its adverse impact on read quality and length can undercut the benefits for isoform discovery. Furthermore, although higher read depth can increase sensitivity in some scenarios (e.g., lr-kallisto), having sufficiently long, high-fidelity reads often proves more critical for accurate isoform quantification than simply adding more coverage from PCR. By comparison, direct RNA libraries often suffer from low throughput that yields limited isoform recovery, making them generally less suitable for isoform-level analyses. Direct cDNA libraries, on the other hand, offer a more favorable balance between throughput and accuracy, frequently resulting in superior isoform detection. Nonetheless, in instances where only PCR-amplified libraries are available, isoform discovery tools such as Bambu, IsoQuant and lr-kallisto showed relatively robust performance, indicating a level of resilience to amplification- induced biases compared to other tools.

As indicated in previous benchmarking results (Chen et al., 2021; Dong et al., 2023; Su et al., 2024), no single isoform discovery tool consistently outperforms others under every sequencing condition, with the choice depending largely on factors such as read length, accuracy, and overall data complexity. StringTie2, for instance, excels when the dataset comprises abundant, high-fidelity reads but adopts strict alignment thresholds that can hinder recall in lower-quality datasets by discarding uncertain reads. By contrast, Bambu preserves low confidence reads through its dynamic modeling strategy, yielding higher recall but introducing a slight trade-off in precision. IsoQuant and FLAIR balance the retention of uncertain reads for quantification while reserving more definitive data for isoform assembly, allowing them to maintain precision with minimal loss of sensitivity. Notably, IsoQuant remains remarkably robust to variability in sequencing quality, offering a favorable precision-

sensitivity balance in most scenarios. Finally, lr-kallisto's pseudoalignment-based approach proves highly effective with high-accuracy reads, particularly in PCR-amplified libraries and suggests that pseudoalignment methods may become increasingly relevant as ONT technology advances toward routinely producing more accurate long reads. These findings also underscore the importance of tailoring tool choice to specific research objectives. For instance, Bambu and IsoQuant excel at reconstructing transcripts exceeding 10 kilobases, which are notoriously prone to fragmentation during ONT library preparation and PCR amplification (Jain et al., 2018), making these tools well-suited for studying ultra-long transcripts. Given their superior quantification accuracy, lr-kallisto and Bambu may be especially advantageous for large-scale differential transcript usage studies or when rapid throughput is crucial. In contrast, projects aiming to discover novel splice variants or comprehensively profile a single gene might favor tools balancing sensitivity and precision (e.g., IsoQuant). Overall, these observations highlight that the optimal isoform detection strategy depends on the interplay between sequencing conditions and the specific goals of the study.

Building on these findings, both sequence quality and the specific research goal of unravelling how alternative splicing shapes MDA5 and RIG-I isoforms informed the choice of the isoform discovery tool used for this analysis. When focusing on a single gene, high sensitivity is paramount for capturing the full spectrum of possible splice variants. In this study, where known IFN-β–induced isoforms were the primary interest, the priority shifted slightly from precision to ensuring robust detection that could establish a baseline isoform profile. Accurate quantification was likewise crucial for determining precise isoform proportions, which serve as a benchmark for evaluating aberrant splicing in disease contexts. Given these requirements, Bambu emerged as the optimal choice. It not only excelled at transcript quantification but also maintained strong sensitivity for mid-length transcripts (2,000–6,000 bp) under lower sequence accuracies. This capability was essential for reliably distinguishing less-dominant isoforms from the primary isoforms in both MDA5 and RIG-I, thereby providing confidence in the measured proportions.

Both RIG-I and MDA5 play pivotal roles in antiviral defense, and the IFN-β–induced expression of their alternative isoforms suggests a functional impact on inflammatory pathways (Brisse and Ly, 2019; Lad *et al.*, 2008). By quantifying the relative

abundance of each isoform post–IFN-β exposure, this study establishes a baseline for normal RLR regulation through alternative splicing, offering valuable insights into how these receptors fine-tune immune activation. Among these regulatory mechanisms, NMD emerged as particularly noteworthy for MDA5: approximately 25% of its transcripts harbour an exon containing a premature stop codon, triggering nonsense-mediated decay (NMD). While NMD typically acts as a quality-control process to eliminate faulty transcripts, the prevalence of these isoforms implies a regulated function rather than a mere splicing error (Fair *et al.*, 2024). Constitutive expression of NMD-prone MDA5 isoforms may allow healthy cells to temper immune responses by "switching off" a fraction of MDA5 expression. Disruption of this balance has been associated with immune disorders; for instance, truncated MDA5 isoforms can contribute to inflammatory bowel disease and other autoimmune conditions (Cananzi *et al.*, 2021), while dysregulation of MDA5 has been linked to Type 1 diabetes (Zurawek *et al.*, 2015). Similarly, we identified an aberrant RIG-I isoform (RIG-I-204) lacking the critical CARD domain necessary for antiviral signalling. Although it cannot activate downstream pathways, it appears to sequester viral ssRNA, thereby preventing the fully functional, CARD-containing RIG- I isoforms from engaging RNA (Lad *et al.*, 2008). Notably, this regulatory isoform was detected only in lung epithelial cells, highlighting a cell-specific mechanism to limit RIG-I–induced inflammation and providing an additional cytoprotective response to infection in the lungs (Mihaylova *et al.*, 2018).

Overall, these findings underscore the importance of alternative splicing in shaping RLR-mediated immune pathways and establish a benchmark for contrasting normal and diseased states, particularly where IFN-β signaling may be dysregulated. RLR isoforms could serve as a preclinical biomarker for adverse immune reactions and a starting point for immunotherapeutic strategies. Notably, targeting spliceosome function, such as using spliceostatin to disrupt the SF3B complex (Lv *et al.*, 2025), has demonstrated promising anti-cancer effects by countering tumor-promoting splicing events. A deeper understanding of these baseline splicing patterns could thus inform diagnostic approaches and spur development of splicing-focused therapies aimed at restoring or optimizing immune function.

Study Limitations

This study has two primary limitations. First, only one replicate of IFN-β– and mock-

treated samples was analyzed, precluding the possibility of robust differential transcript usage assays. Additionally, although our findings illuminate how factors such as read accuracy, read count, and read length affect isoform detection, newer ONT pore chemistries, offering potentially longer and more accurate reads at higher throughput, remain to be evaluated under similar conditions.

Future Work

Further work is needed to extend and refine these findings. For instance, other isoform discovery tools not included in this analysis, such as Oarfish and Espresso, warrant evaluation to broaden the benchmarking landscape. In the present study, each sample was processed independently to facilitate tool-to-tool comparisons; however, some methods (e.g., TALON) rely on grouped samples to enhance isoform confidence. Future benchmarks could explore these multi-sample approaches, potentially providing more robust insights. Additionally, although each tool employs a distinct algorithmic framework, a deeper assessment could compare tools based on shared methodologies, such as grouping all pseudoalignment tools together, to clarify how specific analytical strategies influence performance. Lastly, this work did not assess the tools' ability to detect novel isoforms. Future analyses could address how sequencing quality and read length impact novel isoform discovery, thereby offering a more comprehensive understanding of each tool's capabilities.

Conclusion

Our findings show that basecalling software strongly affects read length, accuracy, and downstream isoform detection. Each isoform discovery tool has distinct strengths, so tool choice depends on research goals and sequencing characteristics. Among the tools tested, Bambu excelled at handling data with moderate read lengths and lower accuracies. Using these insights, we were able to identify context-specific isoform switching in MDA5 and RIG-I, suggesting a regulatory balance between immune activation and suppression that could be targeted by spliceosome-focused immunotherapies. Overall, our results underscore how basecalling accuracy, library preparation, and isoform discovery methods shape long-read transcriptomic quality, providing a foundation for future work on splicing-related immune dysregulation and novel therapeutic strategies.

REFERENCES

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 Biol. 12, R18. Https://doi.org/10.1186/gb-2011-12-2-r18

Albrecht, T., Slabaugh, G., Alonso, E., Al-Arif, S.M.R., 2017. Deep learning for single-molecule science. Nanotechnology 28, 423001. Https://doi.org/10.1088/1361-6528/aa8334

Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., Gouil, Q., 2020. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 21, 30. Https://doi.org/10.1186/s13059-020-1935-5

Arias, M.A., Ke, S., Chasin, L.A., 2010. Splicing by cell type. Nat. Biotechnol. 28, 686–687. Https://doi.org/10.1038/nbt0710-686

Banday, A.R., Stanifer, M.L., Florez-Vargas, O., Onabajo, O.O., Papenberg, B.W., Zahoor, M.A., Mirabello, L., Ring, T.J., Lee, C.-H., Albert, P.S., Andreakos, E., Arons, E., Barsh, G., Biesecker, L.G., Boyle, D.L., Brahier, M.S., Burnett-Hartman, A., Carrington, M., Chang, E., Choe, P.G., Chisholm, R.L., Colli, L.M., Dalgard, C.L., Dude, C.M., Edberg, J., Erdmann, N., Feigelson, H.S., Fonseca, B.A., Firestein, G.S., Gehring, A.J., Guo, C., Ho, M., Holland, S., Hutchinson, A.A., Im, H., Irby, L., Ison, M.G., Joseph, N.T., Kim, H.B., Kreitman, R.J., Korf, B.R., Lipkin, S.M., Mahgoub, S.M., Mohammed, I., Paschoalini, G.L., Pacheco, J.A., Peluso, M.J., Rader, D.J., Redden, D.T., Ritchie, M.D., Rosenblum, B., Ross, M.E., Anna, H.P.S., Savage, S.A., Sharma, S., Siouti, E., Smith, A.K., Triantafyllia, V., Vargas, J.M., Vargas, J.D., Verma, A., Vij, V., Wesemann, D.R., Yeager, M., Yu, X., Zhang, Y., Boulant, S., Chanock, S.J., Feld, J.J., Prokunina-Olsson, L., 2022. Genetic regulation of OAS1 nonsense-mediated decay underlies association with COVID-19 hospitalization in patients of

European and African ancestries. Nat. Genet. 54, 1103–1116. Https://doi.org/10.1038/s41588-022-01113-z

Black, D.L., 2003. Mechanisms of Alternative Pre-Messenger RNA Splicing. Annu. Rev. Biochem. 72, 291–336. Https://doi.org/10.1146/annurev.biochem.72.121801.161720

Bray, N.L., Pimentel, H., Melsted, P., Pachter, L., 2016. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527. Https://doi.org/10.1038/nbt.3519

BUBER, E., DIRI, B., 2018. Performance Analysis and CPU vs GPU Comparison for Deep Learning, in: 2018 6th International Conference on Control Engineering & Information Technology (CEIT). Presented at the 2018 6th International Conference on Control Engineering & Information Technology (CEIT), pp. 1–6. Https://doi.org/10.1109/CEIT.2018.8751930

Cárdenas, W.B., Loo, Y.-M., Gale, M., Hartman, A.L., Kimberlin, C.R., Martínez-Sobrido, L., Saphire, E.O., Basler, C.F., 2006. Ebola Virus VP35 Protein Binds Double-Stranded RNA and Inhibits Alpha/Beta Interferon Production Induced by RIG-I Signaling. J. Virol. 80, 5168–5178. Https://doi.org/10.1128/jvi.02199-05

Castle, J.C., Zhang, C., Shah, J.K., Kulkarni, A.V., Kalsotra, A., Cooper, T.A., Johnson, J.M., 2008. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat. Genet. 40, 1416–1425. Https://doi.org/10.1038/ng.264

Chen, Y., Davidson, N.M., Wan, Y.K., Patel, H., Yao, F., Low, H.M., Hendra, C., Watten, L., Sim, A., Sawyer, C., Iakovleva, V., Lee, P.L., Xin, L., Ng, H.E.V., Loo, J.M., Ong, X., Ng, H.Q.A., Wang, J., Koh, W.Q.C., Poon, S.Y.P., Stanojevic, D., Tran, H.-D., Lim, K.H.E., Toh, S.Y., Ewels, P.A., Ng, H.-H., Iyer, N.G., Thiery, A., Chng, W.J., Chen, L., dasgupta, R., Sikic, M., Chan, Y.-S., Tan, B.O.P., Wan, Y., Tam, W.L., Yu, Q., Khor, C.C., Wüstefeld, T.,

Pratanwanich, P.N., Love, M.I., Goh, W.S.S., Ng, S.B., Oshlack, A., Göke, J., Consortium, S.-Ne., 2021a. A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines. Https://doi.org/10.1101/2021.04.21.440736

Chen, Y., Davidson, N.M., Wan, Y.K., Patel, H., Yao, F., Low, H.M., Hendra, C., Watten, L., Sim, A., Sawyer, C., Iakovleva, V., Lee, P.L., Xin, L., Ng, H.E.V., Loo, J.M., Ong, X., Ng, H.Q.A., Wang, J., Koh, W.Q.C., Poon, S.Y.P., Stanojevic, D., Tran, H.-D., Lim, K.H.E., Toh, S.Y., Ewels, P.A., Ng, H.-H., Iyer, N.G., Thiery, A., Chng, W.J., Chen, L., dasgupta, R., Sikic, M., Chan, Y.-S., Tan, B.O.P., Wan, Y., Tam, W.L., Yu, Q., Khor, C.C., Wüstefeld, T., Pratanwanich, P.N., Love, M.I., Goh, W.S.S., Ng, S.B., Oshlack, A., Göke, J., Consortium, S.-Ne., 2021b. A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines. Https://doi.org/10.1101/2021.04.21.440736

Chen, Y., Sim, A., Wan, Y.K., Yeo, K., Lee, J.J.X., Ling, M.H., Love, M.I., Göke, J., 2023. Context-aware transcript quantification from long-read RNA-seq data with Bambu. Nat. Methods 20, 1187–1195. Https://doi.org/10.1038/s41592-023-01908-w

Dale, R., 2025. Daler/gffutils.

De Arras, L., Alper, S., 2013. Limiting of the innate immune response by SF3A-dependent control of myd88 alternative mrna splicing. Plos Genet. 9, e1003855. Https://doi.org/10.1371/journal.pgen.1003855

Deamer, D., Akeson, M., Branton, D., 2016. Three decades of nanopore sequencing. Nat. Biotechnol. 34, 518–524. Https://doi.org/10.1038/nbt.3423

Espinosa, E., Bautista, R., Larrosa, R., Plata, O., 2024. Advancements in long-read genome sequencing technologies and algorithms. Genomics 116, 110842. Https://doi.org/10.1016/j.ygeno.2024.110842

Fair, B., Buen Abad Najar, C.F., Zhao, J., Lozano, S., Reilly, A., Mossian, G., Staley, J.P., Wang, J., Li, Y.I., 2024. Global impact of unproductive splicing on human gene expression. Nat. Genet. 56, 1851–1861. Https://doi.org/10.1038/s41588-024-01872-x

Gack, M.U., Kirchhofer, A., Shin, Y.C., Inn, K.-S., Liang, C., Cui, S., Myong, S., Ha, T., Hopfner, K.-P., Jung, J.U., 2008. Roles of RIG-I N-terminal tandem CARD and splice variant in TRIM25-mediated antiviral signal transduction. Proc. Natl. Acad. Sci. U. S. A. 105, 16743–16748. Https://doi.org/10.1073/pnas.0804947105

Gack, M.U., Shin, Y.C., Joo, C.-H., Urano, T., Liang, C., Sun, L., Takeuchi, O., Akira, S., Chen, Z., Inoue, S., Jung, J.U., 2007. TRIM25 RING-finger E3 ubiquitin ligase is essential for RIG-I-mediated antiviral activity. Nature 446, 916–920. Https://doi.org/10.1038/nature05732

Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., Jordan, M., Ciccone, J., Serra, S., Keenan, J., Martin, S., mcneill, L., Wallace, E.J., Jayasinghe, L., Wright, C., Blasco, J., Young, S., Brocklebank, D., Juul, S., Clarke, J., Heron, A.J., Turner, D.J., 2018. Highly parallel direct RNA sequencing on an array of nanopores. Nat. Methods 15, 201–206. Https://doi.org/10.1038/nmeth.4577

Goodwin, S., mcpherson, J.D., mccombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333–351. Https://doi.org/10.1038/nrg.2016.49

Huang, N., Nie, F., Ni, P., Luo, F., Wang, J., 2022. Sacall: A Neural Network Basecaller for Oxford Nanopore Sequencing Data Based on Self-Attention Mechanism. IEEE/ACM Trans. Comput. Biol. Bioinform. 19, 614–623. Https://doi.org/10.1109/TCBB.2020.3039244

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H.E., Pedersen, B.S., Rhie, A., Richardson, H., Quinlan, A.R., Snutch, T.P., Tee, L., Paten, B., Phillippy, A.M., Simpson, J.T., Loman, N.J., Loose, M., 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. 36, 338–345. Https://doi.org/10.1038/nbt.4060

Kasianowicz, J.J., Brandin, E., Branton, D., Deamer, D.W., 1996. Characterization of individual polynucleotide molecules using a membrane channel. Proc. Natl. Acad. Sci. U. S. A. 93, 13770–13773. Https://doi.org/10.1073/pnas.93.24.13770

Kato, H., Fujita, T., 2015. RIG-I-like receptors and autoimmune diseases. Curr. Opin. Immunol. 37, 40–45. Https://doi.org/10.1016/j.coi.2015.10.002

Kell, A.M., Gale, M., 2015. RIG-I in RNA virus recognition. Virology 479, 110–121. Https://doi.org/10.1016/j.virol.2015.02.017

Kim, H.K., Pham, M.H.C., Ko, K.S., Rhee, B.D., Han, J., 2018. Alternative splicing isoforms in health and disease. Pflüg. Arch. - Eur. J. Physiol. 470, 995–1016. Https://doi.org/10.1007/s00424-018-2136-x

Kovaka, S., Zimin, A.V., Pertea, G.M., Razaghi, R., Salzberg, S.L., Pertea, M., 2019. Transcriptome assembly from long-read RNA-seq alignments with stringtie2. Genome Biol. 20, 278. Https://doi.org/10.1186/s13059-019-1910-1

Lai, H.-C., Ho, U.Y., James, A., De Souza, P., Roberts, T.L., 2021. RNA metabolism and links to inflammatory regulation and disease. Cell. Mol. Life Sci. CMLS 79, 21. Https://doi.org/10.1007/s00018-021-04073-5

Lee, J.Y., Kong, M., Oh, J., Lim, J., Chung, S.H., Kim, J.-M., Kim, J.-S., Kim, K.-H., Yoo, J.-C., Kwak, W., 2021. Comparative evaluation of Nanopore polishing tools for microbial genome assembly and polishing strategies for downstream analysis. Sci. Rep. 11, 20740. Https://doi.org/10.1038/s41598-021-00178-w

Li, H., 2021. New strategies to improve minimap2 alignment accuracy. Bioinformatics 37, 4572–4574. Https://doi.org/10.1093/bioinformatics/btab705

Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinforma. Oxf. Engl. 34, 3094–3100. Https://doi.org/10.1093/bioinformatics/bty191

Liao, K.-C., Garcia-Blanco, M.A., 2021. Role of Alternative Splicing in Regulating Host Response to Viral Infection. Cells 10, 1720. Https://doi.org/10.3390/cells10071720

Linehan, M.M., Dickey, T.H., Molinari, E.S., Fitzgerald, M.E., Potapova, O., Iwasaki, A., Pyle, A.M., 2018. A minimal RNA ligand for potent RIG-I activation in living mice. Sci. Adv. 4, e1701854. Https://doi.org/10.1126/sciadv.1701854

Liu, Y., Gonzàlez-Porta, M., Santos, S., Brazma, A., Marioni, J.C., Aebersold, R., Venkitaraman, A.R., Wickramasinghe, V.O., 2017. Impact of Alternative Splicing on the Human Proteome. Cell Rep. 20, 1229–1241. Https://doi.org/10.1016/j.celrep.2017.07.025

Loo, Y.-M., Gale, M., 2011. Immune signaling by RIG-I-like receptors. Immunity 34, 680–692. Https://doi.org/10.1016/j.immuni.2011.05.003

Loving, R.K., Sullivan, D.K., Booeshagi, A.S., Reese, F., Rebboah, E., Sakr, J., Rezaie, N., Liang, H.Y., Filimban, G., Kawauchi, S., Oakes, C., Trout, D., Williams, B.A., macgregor, G., Wold, B.J., Mortazavi, A., Pachter, L., 2025. Long-read sequencing transcriptome quantification with lr-kallisto. Https://doi.org/10.1101/2024.07.19.604364

Manuel, J.M., Guilloy, N., Khatir, I., Roucou, X., Laurent, B., 2023. Re-evaluating the impact of alternative RNA splicing on proteomic diversity. Front. Genet. 14, 1089053. Https://doi.org/10.3389/fgene.2023.1089053

Matsumiya, T., Stafforini, D.M., 2010. Function and Regulation of Retinoic Acid-Inducible Gene-I. Crit. Rev. Immunol. 30, 489–513.

Mcnab, F., Mayer-Barber, K., Sher, A., Wack, A., O'Garra, A., 2015. Type I interferons in infectious disease. Nat. Rev. Immunol. 15, 87–103. Https://doi.org/10.1038/nri3787

Mihaylova, V.T., Kong, Y., Fedorova, O., Sharma, L., Dela Cruz, C.S., Pyle, A.M., Iwasaki, A., Foxman, E.F., 2018. Regional Differences in Airway Epithelial Cells Reveal Tradeoff between Defense against Oxidative Stress and Defense against Rhinovirus. Cell Rep. 24, 3000-3007.e3. Https://doi.org/10.1016/j.celrep.2018.08.033

Nanoporetech/medaka, 2025.

Nilsen, T.W., Graveley, B.R., 2010. Expansion of the eukaryotic proteome by alternative splicing. Nature 463, 457–463. Https://doi.org/10.1038/nature08909

Oikonomopoulos, S., Wang, Y.C., Djambazian, H., Badescu, D., Ragoussis, J., 2016. Benchmarking of the Oxford Nanopore minion sequencing for quantitative and qualitative assessment of cdna populations. Sci. Rep. 6, 31602. Https://doi.org/10.1038/srep31602

Ono, Y., Hamada, M., Asai, K., 2022. PBSIM3: a simulator for all types of pacbio and ONT long reads. NAR Genomics Bioinforma. 4, lqac092. Https://doi.org/10.1093/nargab/lqac092

Pagès-Gallego, M., de Ridder, J., 2023. Comprehensive benchmark and architectural analysis of deep learning models for nanopore sequencing basecalling. Genome Biol. 24, 71. Https://doi.org/10.1186/s13059-023-02903-2

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., Salzberg, S.L., 2015. Stringtie enables improved reconstruction of a transcriptome from

RNA-seq reads. Nat. Biotechnol. 33, 290–295. Https://doi.org/10.1038/nbt.3122

Prjibelski, A.D., Mikheenko, A., Joglekar, A., Smetanin, A., Jarroux, J., Lapidus, A.L., Tilgner, H.U., 2023. Accurate isoform discovery with isoquant using long reads. Nat. Biotechnol. 1–4. Https://doi.org/10.1038/s41587-022-01565-y

Quinlan, A.R., Hall, I.M., 2010. Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. Https://doi.org/10.1093/bioinformatics/btq033

Rang, F.J., Kloosterman, W.P., de Ridder, J., 2018. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 19, 90. Https://doi.org/10.1186/s13059-018-1462-9

Rehwinkel, J., Gack, M.U., 2020. RIG-I-like receptors: their regulation and roles in RNA sensing. Nat. Rev. Immunol. 20, 537–551. Https://doi.org/10.1038/s41577-020-0288-3

Ren, X., Linehan, M.M., Iwasaki, A., Pyle, A.M., 2019. RIG-I Selectively Discriminates against 5′-Monophosphate RNA. Cell Rep. 26, 2019-2027.e4. Https://doi.org/10.1016/j.celrep.2019.01.107

Salmela, L., Rivals, E., 2014. Lordec: accurate and efficient long read error correction. Bioinformatics 30, 3506–3514. Https://doi.org/10.1093/bioinformatics/btu538

Savarese, M., Jonson, P.H., Huovinen, S., Paulin, L., Auvinen, P., Udd, B., Hackman, P., 2018. The complexity of titin splicing pattern in human adult skeletal muscles. Skelet. Muscle 8, 11. Https://doi.org/10.1186/s13395-018-0156-z

Schaub, A., Glasmacher, E., 2017. Splicing in immune cells—mechanistic insights and emerging topics. Int. Immunol. 29, 173–181. Https://doi.org/10.1093/intimm/dxx026

Slaff, B., Radens, C.M., Jewell, P., Jha, A., Lahens, N.F., Grant, G.R., Thomas-Tikhonenko, A., Lynch, K.W., Barash, Y., 2021. MOCCASIN: a method for correcting for known and unknown confounders in RNA splicing analysis. Nat. Commun. 12, 3353. Https://doi.org/10.1038/s41467-021-23608-9

Stanifer, M.L., Guo, C., Doldan, P., Boulant, S., 2020. Importance of Type I and III Interferons at Respiratory and Intestinal Barrier Surfaces. Front. Immunol. 11.

Su, C.-H., D, D., Tarn, W.-Y., 2018. Alternative Splicing in Neurogenesis and Brain Development. Front. Mol. Biosci. 5, 12. Https://doi.org/10.3389/fmolb.2018.00012

Takeuchi, O., Akira, S., 2010. Pattern Recognition Receptors and Inflammation. Cell 140, 805–820. Https://doi.org/10.1016/j.cell.2010.01.022

Tang, A.D., Soulette, C.M., van Baren, M.J., Hart, K., Hrabeta-Robinson, E., Wu, C.J., Brooks, A.N., 2020. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat. Commun. 11, 1438. Https://doi.org/10.1038/s41467-020-15171-6

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515. Https://doi.org/10.1038/nbt.1621

Udaondo, Z., Sittikankaew, K., Uengwetwanit, T., Wongsurawat, T., Sonthirod, C., Jenjaroenpun, P., Pootakham, W., Karoonuthaisiri, N., Nookaew, I., 2021. Comparative Analysis of pacbio and Oxford Nanopore Sequencing Technologies for Transcriptomic Landscape Identification of Penaeus monodon. Life 11, 862. Https://doi.org/10.3390/life11080862

Vaser, R., Sović, I., Nagarajan, N., Šikić, M., 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27, 737–746. Https://doi.org/10.1101/gr.214270.116

Volden, R., Schimke, K., Byrne, A., Dubocanin, D., Adams, M., Vollmers, C., 2022. Identifying and quantifying isoforms from accurate full-length transcriptome sequencing reads with Mandalorion. Genome Biology volume 24, Article number: 167 https://doi.org/10.1101/2022.06.29.498139

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. Nat. Commun. 15, 7148. Https://doi.org/10.1038/s41467-024-51639-5

Wen, C., Dematties, D., Zhang, S.-L., 2021. A Guide to Signal Processing Algorithms for Nanopore Sensors. ACS Sens. 6, 3536–3555. Https://doi.org/10.1021/acssensors.1c01618

Wick, R.R., Judd, L.M., Holt, K.E., 2019. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 20, 129. Https://doi.org/10.1186/s13059-019-1727-y

Will, C.L., Lührmann, R., 2011. Spliceosome Structure and Function. Cold Spring Harb. Perspect. Biol. 3, a003707. Https://doi.org/10.1101/cshperspect.a003707

Workman, R.E., Tang, A.D., Tang, P.S., Jain, M., Tyson, J.R., Razaghi, R., Zuzarte, P.C., Gilpatrick, T., Payne, A., Quick, J., Sadowski, N., Holmes, N., de Jesus, J.G., Jones, K.L., Soulette, C.M., Snutch, T.P., Loman, N., Paten, B., Loose, M., Simpson, J.T., Olsen, H.E., Brooks, A.N., Akeson, M., Timp, W., 2019. Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat. Methods 16, 1297–1305. Https://doi.org/10.1038/s41592-019-0617-2

Xu, Z., Mai, Y., Liu, D., He, W., Lin, X., Xu, C., Zhang, L., Meng, X., Mafofo, J., Zaher, W.A., Koshy, A., Li, Y., Qiao, N., 2021. Fast-bonito: A faster deep learning based basecaller for nanopore sequencing. Artif. Intell. Life Sci. 1, 100011. Https://doi.org/10.1016/j.ailsci.2021.100011

Yanai, H., Chiba, S., Hangai, S., Kometani, K., Inoue, A., Kimura, Y., Abe, T., Kiyonari, H., Nishio, J., Taguchi-Atarashi, N., Mizushima, Y., Negishi, H., Grosschedl, R., Taniguchi, T., 2018. Revisiting the role of IRF3 in inflammation and immunity by conditional and specifically targeted gene ablation in mice. Proc. Natl. Acad. Sci. 115, 5253–5258. Https://doi.org/10.1073/pnas.1803936115

Zhang, Q., Wan, H., Huang, S., Zhang, Y., Wang, Y., Guo, X., He, P., Zhou, M., 2016. Critical role of RIG-I-like receptors in inflammation in chronic obstructive pulmonary disease. Clin. Respir. J. 10, 22–31. Https://doi.org/10.1111/crj.12177

Zikherman, J., Weiss, A., 2008. Alternative Splicing of CD45: The Tip of the Iceberg. Immunity 29, 839–841. Https://doi.org/10.1016/j.immuni.2008.12.005

Supplementary Figures

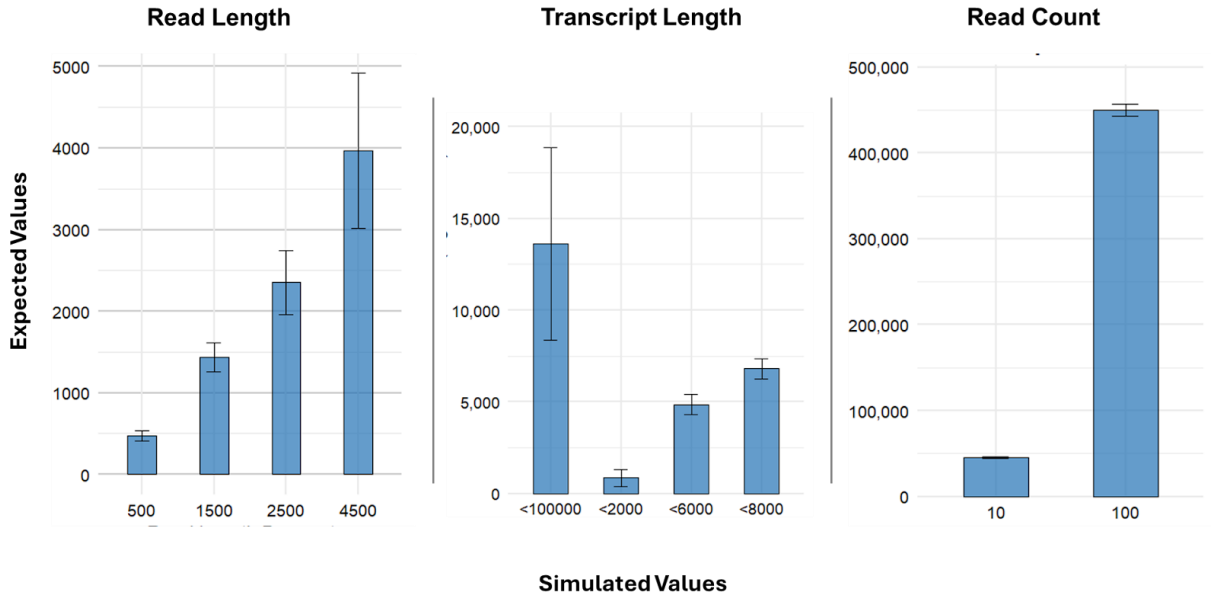


Figure S1: Characteristics of sequencing data from simulated values. The simulated datasets were characterized using NanoPlot, which confirmed that the observed read length, transcript length, and read accuracy distributions aligned with the specified parameters (Figure 3). Notably, simulations with longer read lengths (e.g., 4,500 base pairs) exhibited higher standard deviations, reflecting the distribution model in PBSIM. Transcript lengths ranging from 10,000 to 100,000 base pairs introduced significant variability, while read counts scaled by a factor of ten between simulation conditions

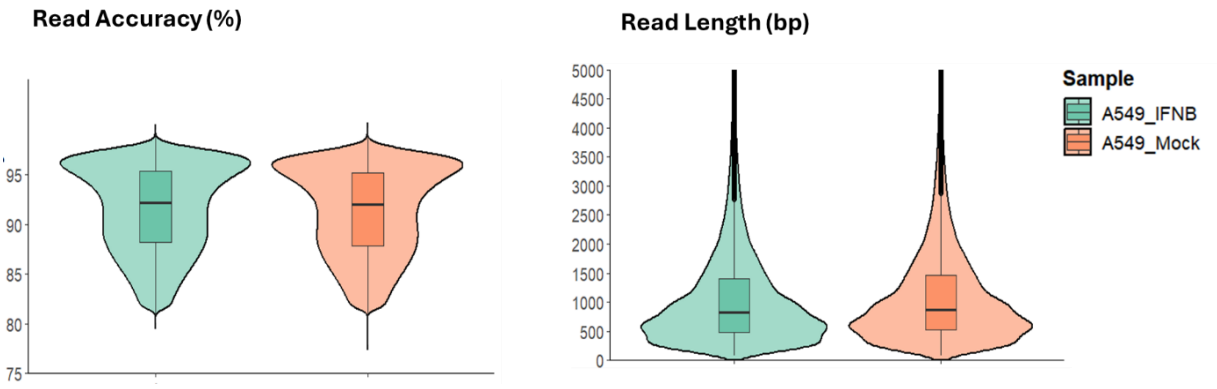


Figure S2: Distribution of Read Lengths and Read Accuracies in A549 Samples.

Violin plots and box and whiskers plots illustrate the distribution of read lengths and accuracies in A549 cells under mock treatment (orange) and IFN-β treatment (green) conditions. The violin plots show the density of data points, with the width of each plot indicating the data distribution at different values. The box and whiskers plots represent the interquartile range (IQR), where the top and bottom edges of the box correspond to the 75th and 25th percentiles, respectively, and the line inside the box indicates the median value. The whiskers extend from the edges of the box to the maximum and minimum values within 1.5 times the IQR, with data points outside this range shown as individual outliers. This combined visualization provides insight into the spread, central tendency, and density of the data across both treatment conditions.