BMC Infectious Diseases

RESEARCH

Open Access

Utilizing Nanopore direct RNA sequencing of blood from patients with sepsis for discovery of co- and post-transcriptional disease biomarkers

Jingni He[1†], Devika Ganesamoorthy[2,3†], Jessie J.-Y. Chang[4†], Jianshu Zhang[4], Sharon L. Trevor[4], Kristen S. Gibbons[3], Stephen J. McPherson[5], Jessica C. Kling[5], Luregn J. Schlapbach[3,6], Antje Blumenthal[5], Lachlan J. M. Coin[1,2,4,7*] and RAPIDS Study Group

Abstract

Background RNA sequencing of whole blood has been increasingly employed to find transcriptomic signatures of disease states. These studies traditionally utilize short-read sequencing of cDNA, missing important aspects of RNA expression such as differential isoform abundance and poly(A) tail length variation.

Methods We used Oxford Nanopore Technologies sequencing to sequence native mRNA extracted from whole blood from 12 patients with definite bacterial and viral sepsis and compared with results from matching Illumina short-read cDNA sequencing data. Additionally, we explored poly(A) tail length variation, novel transcript identification, and differential transcript usage.

Results The correlation of gene count data between Illumina cDNA- and Nanopore RNA-sequencing strongly depended on the choice of analysis pipeline; *NanoCount* for Nanopore and *Kallisto* for Illumina data yielded the highest mean Pearson's correlation of 0.927 at the gene level and 0.736 at the transcript isoform level. We identified 2 genes with differential polyadenylation, 9 genes with differential expression and 4 genes with differential transcript usage between bacterial and viral infection. Gene ontology gene set enrichment analysis of poly(A) tail length revealed enrichment of long tails in mRNA of genes involved in signaling and short tails in oxidoreductase molecular functions. Additionally, we detected 240 non-artifactual novel transcript isoforms.

Conclusions Nanopore RNA- and Illumina cDNA-gene counts are strongly correlated, indicating that both platforms are suitable for discovery and validation of gene count biomarkers. Nanopore direct RNA-seq provides additional advantages by uncovering additional post- and co-transcriptional biomarkers, such as poly(A) tail length variation and transcript isoform usage.

Keywords Direct RNA-sequencing, Oxford Nanopore Technologies, Polyadenylation, Long-read sequencing, Differential transcript usage, Novel isoform detection, Disease biomarkers

[†]Jingni He, Devika Ganesamoorthy and Jessie J.-Y. Chang contributed equally to this work.

*Correspondence:
Lachlan J. M. Coin
lachlan.coin@unimelb.edu.au
Full list of author information is available at the end of the article

Background

Transcriptomics provides a time- and cost-effective method of understanding disease status of the patient and enables an avenue to develop targeted prophylactic, diagnostic, and therapeutic strategies. Studies investigating host transcriptional response typically employ high-throughput short-read sequencing, such as Illumina sequencing, to identify gene-count biomarkers of disease [1–5]. These platforms provide highly accurate sequence data with high coverage [6]. However, short-read approaches rely on converting to complementary DNA (cDNA) followed by cDNA amplification using polymerase chain reaction (PCR), both of which may introduce biases that interfere with the accurate quantification of transcripts [7]. Moreover, short-read sequencing has transcript/gene length-dependent expression bias towards longer transcripts/genes [8], as well as complex compositional biases such as with guanine-cytosine (GC) content [9]. Additionally, the short read lengths limit the resolution of transcript isoforms, leading to challenges in accurately quantifying the expression of different transcripts and interrogating alternative splicing patterns and differential isoform expression [10].

Biomarker discovery within the transcriptome can be extended beyond expression levels to include the detection of co-/post-transcriptional modifications such as 3' end modification by addition of a polyadenine (poly(A)) tail facilitated by poly(A) polymerases [11]. RNA Poly(A) tails play a role in post-transcriptional regulation, including mRNA stability and translational efficiency [12], where the length has been shown to be important in translation stimulation via poly(A) binding protein (PABP) [13]. Furthermore, highly expressed transcripts have been shown to harbor shorter poly(A) tails [14]. While poly(A) tail lengths have been investigated via head-to-tail ligation PCR [15] or alternative short-read sequencing techniques (e.g. PAL-seq [16] and TAIL-seq [17]), these homopolymers can extend to several hundred nucleotides (nt), which therefore poses limitations with short-read sequencing technologies [17]. By design, short-read RNA-seq typically uses anchored oligo-dT priming for reverse transcription, which prohibits the capture of the full poly(A) tail length, failing to capture the full range of poly(A) tail lengths. Therefore, most biomarker discovery projects are unable explore co-/post-transcriptional modifications as potential biomarkers.

To overcome these challenges, an alternative strategy for RNA-sequencing (RNA-seq) has emerged, using direct or native RNA-seq) on an array of nanopores by Oxford Nanopore Technologies (ONT) [18–20]. This advancement facilitates the direct analysis of RNA transcripts, minimizing potential errors and bias associated with cDNA synthesis and amplification, detection of polyadenylation length as well as the acquisition of long read data, which allows the identification of splice variants [19, 21], thus providing a more comprehensive view of the transcriptome. The additional information gained from this platform provides alternative methods of disease biomarker detection.

While the gene expression biases of Illumina cDNA sequencing have been widely studied [22], it remains unclear which biases are present in quantification of Nanopore direct RNA-seq, and whether Nanopore direct RNA-seq can be used in place of Illumina short-read sequencing in transcriptional biomarker discovery and validation studies [18–20]. We therefore set out to compare blood mRNA data derived from patients with definite viral or bacterial sepsis in previously published Illumina cDNA [23] with Nanopore direct RNA-seq data to understand the gene expression correlation between the two platforms. We also set out to investigate which additional information for biomarker studies could be obtained from Nanopore direct RNA-seq.

Methods

Study design and participants

The samples in this study were selected from RNA collected for a larger study of 907 children evaluated for sepsis [23]. The institutional Human Research Ethics Committee approved the study on June 9, 2017 (HREC/17/QRCH/85). Written informed consent or permission to proceed was obtained from the parents or caregivers of all participants. Bacterial infections were confirmed by cultures of sterile sites by standard pathology services which must be compatible with the clinical presentation. Confirmed viral infection were based on routine diagnostics (influenza A and B, respiratory syncytial Virus (RSV), parainfluenza 1–3, human metapneumovirus (hMPV), adenovirus, enterovirus) and add-on viral diagnostics of specimens as clinically indicated (such as Enterovirus-PCR in infants with suspected sepsis or central nervous system infection). Out of the 907 children, 235 (~25.9%) and 210 (~23.2%) had definite bacterial or viral infections, respectively. Out of the children with definite bacterial or viral infections, 12 samples (6×definite bacterial and 6×definite viral) were chosen for this study based on samples with the most abundant RNA remaining after the original study (Table 1) [23].

Sample collection and processing

Blood samples were collected from children patients evaluated for sepsis. 2.5 mL of blood was collected in PAXgene Blood RNA tubes (PreAnalytix) and total RNA was extracted using the PAXgene Blood miRNA Kit (PreAnalytix).

Table 1 Clinical, microbiological, and severity characteristics of cohort

Characteristic	Category	Cohort ***N*** **= 12**
Gender *n (%)*	Female	6 (50)
Age *n (%)*	< 1 year	10 (83)
	1–5 years	1 (8)
	5–10 years	0 (-)
	10–18 years	1 (8)
Age (years) *median (IQR)*		0.6 (0.4, 0.8)
Chronic condition *n (%)*	No	9 (75)
	Yes	3 (25)
Symptoms at presentation *n (%)*	Fever	8 (67)
	Rash	2 (17)
	Altered level of consciousness	2 (17)
	Irritability	3 (25)
	Seizures	1 (8)
	Pain	1 (8)
	Nausea/Vomiting	4 (33)
	Diarrhoea	2 (17)
	Respiratory distress/apnoea	4 (33)
	Cough	6 (50)
	Pale/cyanotic episode	2 (17)
	Cold extremities	1 (8)
	Skin / wound infection	0 (-)
	Other	1 (8)
Primary clinical focus *n (%)*	Sepsis without a source	3 (25)
	Lower respiratory infection	2 (17)
	Upper respiratory infection	3 (25)
	ENT infection/abscess	1 (8)
	Other	3 (25)
Time from hospital admission to sampling (hours) *median (IQR)*		4.1 (2.9, 16.7)
Admission to PICU *n (%)*	Yes	7 (58)
Laboratory characteristics at baseline *median (IQR)*	Lactate [mmol/l]	1.5 (1.3, 1.8) *(N = 9)*
	Creatinine [µmol/l]	29 (29, 30) *(N = 11)*
	Bilirubin [µmol/l]	10 (6, 22) *(N = 11)*
	Platelets [$*10^3$/µL]	310 (183, 406) *(N = 11)*
	White Cell Count [$*10^3$/µL]	19 (11.6, 20.1) *(N = 11)*
	C-reactive protein [mg/L]	37.5 (29, 120) *(N = 10)*
Infection Type *n (%)*	Definite Bacterial	6 (50)
	Definite Viral	6 (50)
At least one organ dysfunction *n (%)*	Baseline	6 (50)
	24 h	3 (25)
Organ dysfunction remote from the primary site of infection *n (%)*	Baseline	6 (50)
	24 h	3 (25)
Any organ support *n (%)*	Baseline	5 (42)
	24 h	3 (25)
Any Inotropes *n (%)*	Baseline	2 (17)

Table 1 (continued)

Characteristic	Category	Cohort *N* = 12
	24 h	2 (17)
Multi-organ dysfunction *n (%)*	Baseline	5 (42)
	24 h	3 (25)

RNA QC and quantification

RNA samples were quantified using the Qubit™ RNA Broad Range Assay Kit (Invitrogen) and QC was performed using the Agilent RNA assay (#5067–5576) on the TapeStation 4200 (Agilent # G2991AA) as per the manufacturer's protocol.

GLOBINclear™-Globin mRNA depletion

1–4 µg of total RNA in a maximum volume of 14 µL was used to remove globin mRNA using the GLOBINclear™-Human Kit, for globin mRNA depletion (Invitrogen #AM1980), as per the manufacturer's protocol. On completion of the mRNA depletion protocol, each RNA was quantified, and QC was performed using the Qubit™ RNA Broad Range Assay Kit (Invitrogen) and the Agilent RNA assay (#5067–5576) on the TapeStation 4200 (Agilent #G2991AA) as per the manufacturer's protocol.

ONT library preparation and sequencing

Libraries were prepared following the Direct RNA Sequencing protocol (ONT, #SQK-RNA002) as per the manufacturer's instructions including the RCS, with only modifications to the amount of input RNA (500 – 700 ng of globin-depleted total RNA), to take into account the variability in mRNA content (~1–5%) within total RNA and maximize our output. For samples with RNA concentration lower than 50 ng/µL, a maximum input volume of 9 µL was used to prepare the libraries. On completion of the library prep, the reversed-transcribed and adapted RNA was sequenced on a MinION Mk1B (Oxford Nanopore) using a R9.4.1 flow cell using *MinKNOW* v22.12.7 with the default settings when the flow cells were used once, and v20.06.18 with the default settings for a total of 24 h if the flow cell was washed and reused. On completion of the first round of sequencing, a flow cell wash was performed using a Flow Cell Wash Kit (ONT, #EXP-WSH004) as per the manufacturer's protocol. Once the flow cell was washed and pore QC checked, a second library was loaded and sequenced according to the same settings that was mentioned previously.

Illumina cDNA-Sequencing

Data from Illumina cDNA-sequencing (cDNA-seq) was derived from our previous work [23]. Briefly, libraries were prepared from total RNA using the TruSeq Stranded Total RNA (Ribo-Zero GOLD) Library Preparation kit (Illumina). Strand-specific libraries were sequenced using the Illumina NextSeq 75 cycle (1×75 bp) High Output Run.

Basecalling and alignment

For the 12 Nanopore sequencing datasets, *Dorado* v5.3 was used for basecalling while the model was set as rna002_70bps_hac@v3 and the "–estimate-poly-a" parameter was applied. Fast5 files were converted to Pod5 format before inputting to *Dorado* software using "pod5 convert fast5" as per recommended in the *Dorado* user manual. Only passed reads were kept for analysis. The output format for *Dorado* was set to bam files to keep more information including poly(A) tail length. By using "samtools bam2fq", bam output files were converted to fastq format for mapping purposes. For 12 Illumina sequence datasets, fastq files were demultiplexed on the sequencing machine. Gencode GRCh38 v35 genome and transcriptome human references were used as the references. *Minimap2* v2.24 was used for mapping, with the command "minimap2 -t 20 -ax splice -uf -k14 -L ref.fa sample.fastq" for ONT reads and default *Minimap2* short read parameters for Illumina datasets. *Samtools* v1.16.1 was used to sort and index the bam file created from mapping process. Mapping statistics results were calculated using the *Samtools* "flagstats" function.

Pearson correlation analysis of nanopore and illumina sequencing data

To evaluate the correlation between Nanopore direct RNA-seq data and Illumina cDNA sequencing data, we conducted Pearson correlation analysis using *R*. Nanopore direct RNA-seq data underwent processing with various software packages, including *NanoCount* [24], *IsoQuant* [25], *HTSeq* [26], and *Bambu* [27] for Nanopore RNA-seq, a while Illumina cDNA-seq data were processed with *Kallisto* [28] and *HTSeq* [26]. We calculated

the fishers-z transformation on the Nanocount-Kallisto Pearson correlation coefficient and Isoquant-Kallisto Pearson correlation coefficient. By conducting a z-test between corresponding samples' transformed values, we calculated the p-value to decide the significance of difference.

Python3 and *R* scripts were developed to standardize transcript IDs and gene names across different software. Detailed instructions for using each software and their respective scripts can be found in their software documentation. Transcript isoforms were grouped into genes using established gene annotation databases—Ensembl and GENCODE [29, 30].

For each combination of sequencing platform (Nanopore or Illumina) and processing software, raw count data or transcript-level abundance estimates were obtained. Pearson correlation coefficients were then computed between corresponding gene-to-gene expression values across samples for the mapped data. Transcript-level analyses were carried out without mapping to genes, but via calculating Pearson correlations directly. All correlation analyses were conducted in *R* v4.3.1, utilizing built-in functions for calculating Pearson correlation coefficients.

Poly(A) tail length analysis

When using the *Dorado* basecaller [31] with the parameter "–estimate-poly-a", the output bam file will contain an extra tag to record the poly(A) tail length in each read. A summary on read length and poly(A) tail length was created with the command "samtools view basecalling.bam | awk '/pt:i/{print $1,length($10),$NF}' | sed 's/pt:i://g'".

Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) was conducted to identify significantly enriched pathways and biological processes associated with the experimental conditions. We utilized pre-ranked GSEA with the GSEA *R* packages, focusing on coding genes, excluding mitochondrial transcripts. Typically, GSEA is employed to analyze genes based on their differential expression ranks or other relevant scores. In our study, genes were ranked according to their poly(A) tail lengths, from longest to shortest.

For the analysis, we utilized the *clusterProfiler* package from Bioconductor. Specifically, the "ridgeplot" function within *clusterProfiler* was used to perform the GSEA targeting the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and Gene Ontology (GO) terms databases. For KEGG pathway enrichment analysis, the "enrichKEGG" function was utilized to identify significantly enriched pathways. Similarly, for GO term enrichment analysis, the "enrichGO" function was used to determine significantly enriched molecular functions (MF) and cellular components (CC). Enrichment scores and significance levels were computed using permutation testing, with a False Discovery Rate (FDR) threshold set at 0.05 to determine statistically significant enrichment. All analyses were conducted in *R* v4.3.1 with *clusterProfiler* v4.12.0 [32], ensuring reproducibility and robustness of the results.

Differential expression analysis

DESeq2 v1.42.0 was used to identify differentially expressed genes from direct RNA-seq data. A minimum expression threshold of 10 reads per gene across all samples was applied. Comparisons between viral and bacterial infection samples were conducted using the standard pipeline. Genes with an adjusted P-value < 0.05 and $|\log_2 FC| \geq 1$ were considered significantly differentially expressed. Volcano plots were generated using the *EnhancedVolcano* v1.20.0 package in R.

Differential polyadenylation analysis

The differential polyadenylation analysis aimed to identify variations in poly(A) tail lengths across different experimental conditions, specifically comparing viral and bacterial infection samples. This analysis sought to elucidate how changes in polyadenylation patterns might correlate with gene expression and functional outcomes.

Poly(A) tail length measurements were obtained from Nanopore RNA-seq data, providing high-resolution insights into polyadenylation dynamics. The raw poly(A) lengths were log-transformed due to their right-skewed distribution. Subsequently, the package *lmerTest* v3.1.3 [33] was employed to perform a linear mixed-effects regression (*lmer*), where the log-transformed poly(A) length for all reads mapped to one gene served as the response variable, the infection type (viral or bacterial) as the fixed effect, and the sample batch as the random effect. Per-gene P-values were generated and adjusted using the Benjamini-Hochberg (BH) method with the 'p.adjust' function in *R*. Genes exhibiting differential polyadenylation were identified using cutoffs of an adjusted P-value < 0.05 and $|\log_2 FC| \geq 0.5$.

Raincloud plots were generated for raw poly(A) tail lengths of all reads that mapped to each differentially polyadenylated gene (DPG) under both conditions (viral and bacterial infections) using *ggplot2* v3.5.1, replicating the raincloud plots generated by the *raincloudplots* v0.2.0 package in *R* [34]. To perform the sensitivity analysis, the bootstrapping method was applied. For each DPG, reads assigned to the gene were resampled with replacement

and *lmerTest* was applied 100 times. A comparison between the adjusted *P*-value from 100 experiments and the set threshold (adjusted *P*-value < 0.05) was performed to test the hypothesis of significant robust difference for poly(A) length on DPGs between viral and bacterial samples. Principal component analysis (PCA) was conducted using both gene-level abundance from *NanoCount* and average poly(A) length of genes across all ONT samples estimated by *Dorado* using 'procomp' and 'ggplot' via *R*.

Novel isoform identification

Due to low sequence coverage per sample, we aggregated the direct Nanopore RNA-seq data from all samples to detect novel isoforms with *IsoQuant* v3.3.1 using fastq files as inputs. *IsoQuant* utilizes the input annotation file (hg38 GFF3), and matches reads against known transcripts. Next, it performs splice site correction, intron graph construction and transcript discovery. The counts file derived from this step were used for downstream analyses (e.g. differential expression analysis).

To identify any artifacts in our list of detected novel isoforms, we applied *SQANTI3* v5.1.2 [35] to the *IsoQuant* output GTF file containing the entire reference annotation plus all discovered novel transcripts to check the quality of the detected transcripts and filter for true isoforms. Quality control was carried out using the 'qc' function and cross-validated by publicly available datasets such as human refTSS (v3.1.hg38), and poly(A) motifs. *SQANTI3* uses a random forest classifier to filter out artifacts by learning high and low-quality attributes from a True Positive (TP) and True Negative (TN) transcript set, building a model to distinguish artifacts and isoforms based on TN and TP properties. A random forest probability filter of greater than or equal to 0.7 was utilized for this step.

Furthermore, we extended the annotation with new novel isoforms discovered from *IsoQuant* and *SQANTI3* and ran the *Featurecounts* tool to confirm the existence of the novel isoforms.

Differential transcript usage analysis

We integrated the identified novel transcripts into the input annotation file and subsequently re-ran *IsoQuant*. Counts (TPM) derived from *IsoQuant* utilizing transcriptome-mapped BAM files were used to quantify the differential transcript usage between bacterial and viral samples. For differential transcript usage analysis, the quantified counts were input into *DRIMSeq* v1.14.0 [36], a tool designed to detect differences in transcript isoform usage. Prior to analysis, the counts underwent filtering based on specific conditions to ensure robustness and reliability. Specifically, parameters including *min_samps_gene_expr, min_samps_feature_expr, min_gene_expr, and min_feature_exp*r were set to 12, 4, 10, and 10, respectively. As there are issues with utilizing solely the outputs of *DRIMSeq* due to the lack of an appropriate FDR control, we applied the recommended stage-wise testing to alleviate this issue via *StageR v1.26.0* [37]. In this approach, the first stage involves filtering genes based on BH-adjusted p-values at the gene level. Genes that pass this stage proceed to the second stage, where transcript-level *P* -values are adjusted for each gene to control both Family-Wise Error Rate (FWER) and BH-adjusted p-values. The threshold utilized was padj < 0.05.

Results

Comparison of gene and transcript expression quantification between direct RNA-sequencing and short-read Illumina cDNA-sequencing

To make comparisons between Illumina and Nanopore direct RNA-seq data, we sequenced RNA samples derived from whole blood of 12 patients with sepsis with Nanopore direct RNA-seq and compared the data to Illumina sequencing data (described in our previous work [23]). Nanopore sequencing yielded an average of 1,279,075 reads per sample (Supplementary Table 1). The aligned read lengths had a median of 971 nucleotides (Supplementary Table 1).

We evaluated the Pearson correlation in read counts per coding gene across different sequencing methods and all 12 samples. Figure 1A illustrates the correlations between Nanopore RNA-seq and Illumina cDNA-seq for all samples using widely-used RNA-seq quantification tools, including *NanoCount* [24], *IsoQuant* [25], *HTSeq* [26], and *Bambu* [27] for Nanopore RNA-seq, and *Kallisto* [28] and *HTSeq* [26] for Illumina RNA-seq. For the majority of individual samples,

(See figure on next page.)

Fig. 1 Gene-to-gene comparison with direct RNA-seq and Illumina cDNA-seq using different pipelines. **A** Pearson correlations between Nanopore RNA-seq and Illumina RNA-seq for all samples using different quantification tools, including *NanoCount*, *IsoQuant*, *HTSeq*, *Bambu*, and *Kallisto*. The order of the keys on the X-axis is ONT_Illumina, for example, HTSeq_Kallisto represents *HTSeq* for ONT correlated with *Kallisto* for Illunima. **B** JSD (Jensen–Shannon Divergence) between Nanopore RNA-seq and Illumina cDNA-seq for all samples using different RNA-seq quantification tools, including *NanoCount*, *IsoQuant*, *HTSeq*, *Bambu*, and *Kallisto*. **C** The heatmap of Pearson correlations on coding genes across all 12 samples using *NanoCount* for Nanopore and *Kallisto* for Illumina RNA-seq

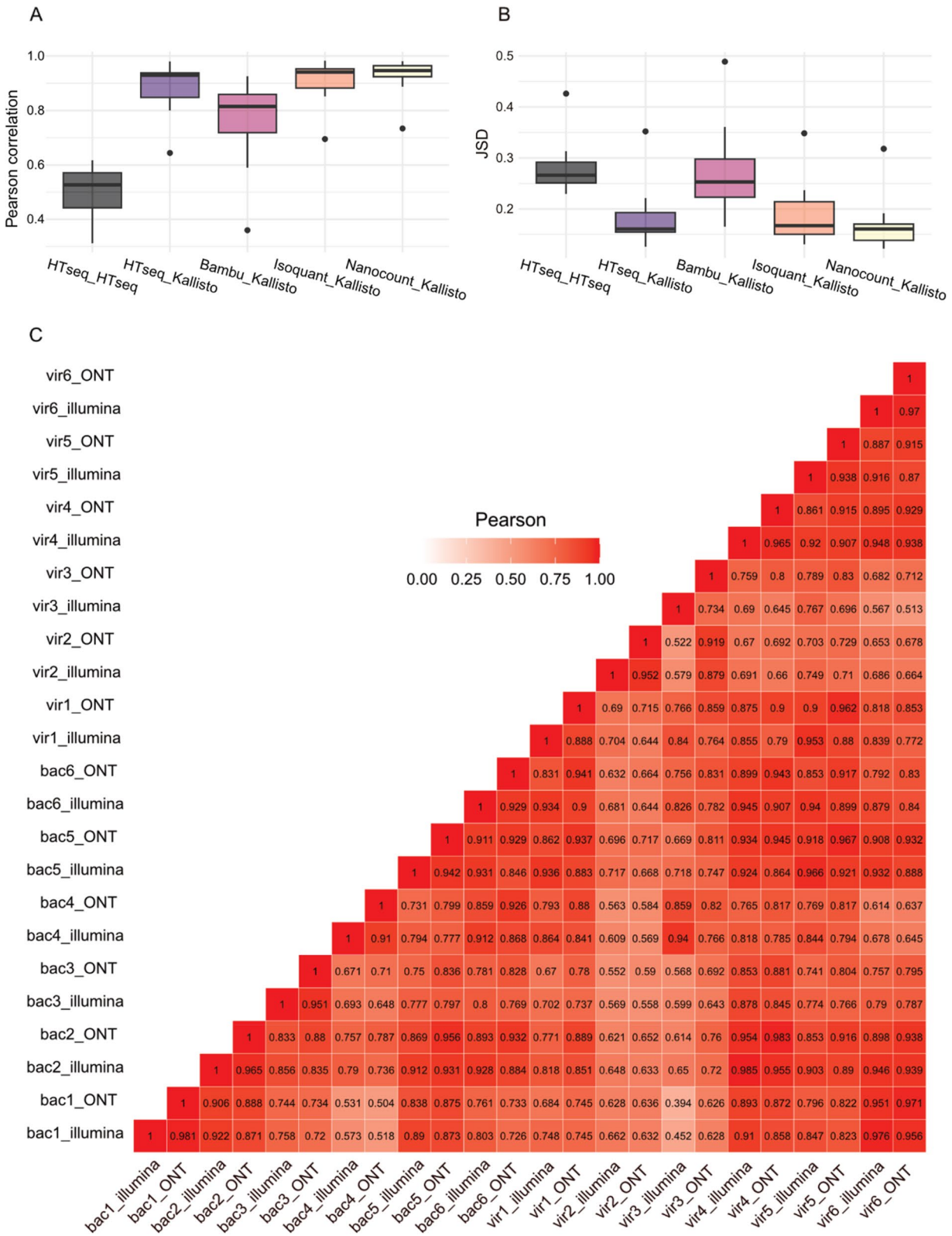

Fig. 1 (See legend on previous page.)

high correlations were observed between Nanopore and Illumina RNA-seq, with the highest correlations found between *NanoCount* and *Kallisto* (r = 0.734–0.981, mean = 0.927), followed by *IsoQuant* and *Kallisto* (r = 0.695–0.983, = mean 0.910), *HTSeq* and *Kallisto* (r = 0.644–0.980, mean = 0.885), *Bambu* and *Kallisto* (r = 0.360–0.926, mean = 0.760), and *HTSeq* and *HTSeq* (r = 0.312–0.617, mean = 0.500) (Fig. 1A). Overall, we observed better consistency between *Kallisto* and other Nanopore RNA-seq tools compared to using *HTSeq* for both Nanopore and Illumina RNA-seq (Supplementary Fig. 1), which suggested that *Kallisto* performed better than *HTSeq* for short-read sequencing performed on the Illumina platform. The correlations for each sample between Isoquant-Kallisto and Nanocount-Kallisto were found to be significantly different in 8 out of 12 samples (P-value < 0.01, Supplementary Table 2). Additionally, we used Jensen–Shannon Divergence (JSD) to measure the similarity between the distributions of Nanopore RNA-seq and Illumina RNA-seq data for all samples using various RNA-seq quantification tools, where two identical distributions have JSD = 0 (the smaller, the better). *NanoCount* and *Kallisto* outperformed the alternatives, not only in terms of mean JSD values (mean 0.168) but also in their variances (Fig. 1B). We further evaluated gene-to-gene correlations between Nanopore RNA-seq and Illumina RNA-seq and observed that the number of highly correlated genes increased as we excluded genes with low expression levels (Supplementary Fig. 2A). A similar trend was noted for transcript-to-transcript correlations (Supplementary Fig. 2B). Correlations between Nanopore RNA-seq (*Nanocount*) and Illumina RNA-seq (*Kallisto*) were lower (r = 0.435–0.885, mean = 0.736) compared to those observed at the gene level (Supplementary Fig. 3).

Interestingly, we noted that when analyzing Illumina data with *Kallisto* [28], the pipeline mitigated biases introduced by gene lengths by utilizing the Transcripts Per Million (TPM) metric, with normalization accounting for gene length (p > 0.37) (Supplementary Fig. 4A). However, we observed a length bias towards shorter genes in Nanopore data with *NanoCount*, when using the TPM metric, without normalization accounting for gene length (p < 0.00001) (Supplementary Fig. 4B). Furthermore, GC content impacted both *Kallisto* and *NanoCount* (p < 0.005) (Supplementary Figs. 4C-D).

Collectively, our results highlight that gene expression estimates from Illumina and Nanopore platforms are highly correlated with certain combinations of pipelines, especially when using *NanoCount* for Nanopore direct RNA-seq and *Kallisto* for Illumina sequencing. Furthermore, length-dependent biases are more prevalent in Nanopore sequencing and GC content biases are present in both sequencing platforms.

Poly(A) tail lengths of mitochondrial vs non-mitochondrial transcripts in human blood mRNA

From the results above, it was apparent that since we obtained similar quantification outputs to Illumina cDNA-seq with Nanopore RNA-seq, the two platforms may be considered equivalent terms of expression estimations, noting that Illumina cDNA-seq still remains more cost-effective. However, as mentioned previously, Nanopore direct RNA-seq provides additional advantages with its long-read capability, such as poly(A) tail length detection, although it remains unclear whether these features are important for biomarker discovery.

We therefore estimated the length of poly(A) tails at the 3' end of transcripts using the built-in function of the ONT *Dorado* basecaller [31]. For mitochondrial transcripts, the overall distribution of poly(A) lengths was centred at ~ 45 nt, and few poly(A) tails exceeded 70 nt in length (Fig. 2A). These findings are consistent with previous studies on mitochondrial poly(A) RNA in human cell lines [38, 39]. In contrast, nuclear transcripts exhibited a wider length distribution across the 12 samples, with a peak around ~ 80 nt, and an average of 0.21% of poly(A) tails of transcripts across the samples were longer than 350 nt (Fig. 2B). This highlighted the capability of long-read sequencing for transcriptome-wide poly(A) length estimations.

GSEA of genes ranked by poly(A) tail lengths highlights molecular pathways enriched in genes with short and long poly(A) tails

Whether poly(A) tail lengths are randomly distributed or specific to functional units of cellular pathways is yet to be fully understood. Gene Set Enrichment Analysis (GSEA) identifies pathways where genes are enriched at

(See figure on next page.)

Fig. 2 Poly(A) length distribution and Gene Set Enrichment Analysis (GSEA) using genes ranked by poly(A) tail lengths. **A** Poly(A) length distribution in mitochondrial transcripts. **B** Poly(A) length distribution in nuclear transcripts. **C-E** Ridgeplots from the clusterProfiler package with the X-axis indicating the poly(A) lengths and Y-axis indicating the GO term or KEGG pathway. The distribution is the distribution of poly(A) length of those genes that enriched in the corresponding GO enrichment analysis **C** molecular function **D** cellular component and **E** KEGG pathway, and the colour indicates the significance, with adjusted P-value < 0.05 deemed as significant (the full list of significant pathways can be viewed in Supplementary Tables 2–4). The mitochondrial transcripts are excluded and numbers on the plots indicate the number of genes relevant to the GO term/pathway

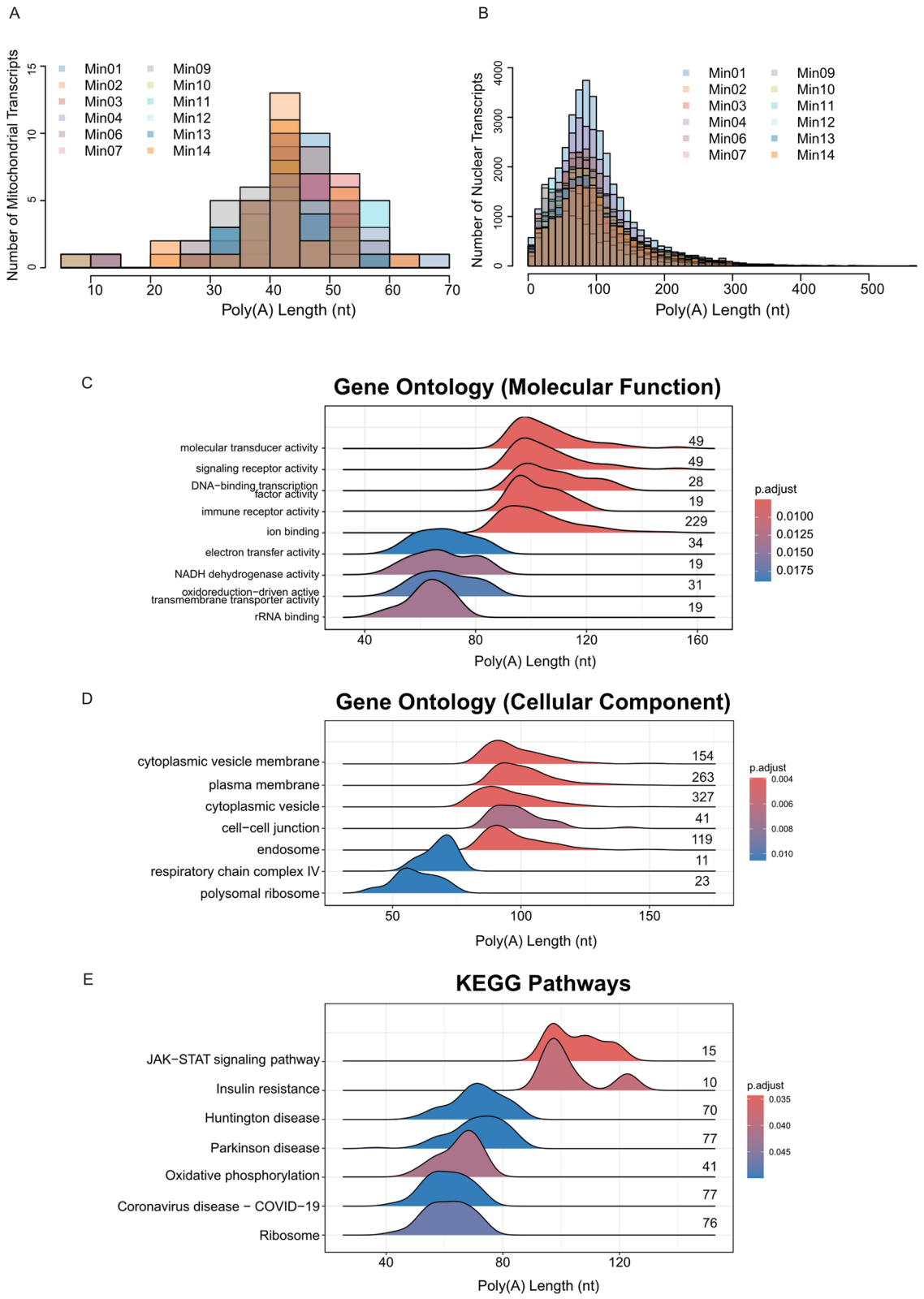

Fig. 2 (See legend on previous page.)

the extremes of the ranked gene list, more than would be expected by chance alone. Traditionally, GSEA has found widespread application in the analysis of genes based on their differential expression rank or other scores [40–42]. Here, we employed pre-ranked GSEA using the GSEA *R* packages on 1,520 coding genes, excluding mitochondrial transcripts [32, 43]. In our study, genes were ranked according to their median poly(A) tail lengths, from longest to shortest. The median poly(A) tail lengths for the coding genes ranged from 26 to 147 nt, with a mean of 83 nt. We conducted GSEA to explore the GO terms (Fig. 2C-D Supplementary Tables 3–4) and KEGG pathway databases (Fig. 2E & Supplementary Table 5) and identified pathways significantly associated with longer or shorter poly(A) tails.

The GO term analysis revealed that genes with shorter poly(A) tails exhibited significant enrichment in functions related to energy production and protein synthesis such as *NADH dehydrogenase activity, electron transfer activity, oxidoreduction-driven active transmembrane transporter activity* and *rRNA binding* (Fig. 2C). The presence of shorter poly(A) tails in these pathways suggested that stability of mRNA derived from genes in these pathways may be reduced compared to genes belonging to other cellular pathways [14]. In contrast, the recent evidence regarding abundant and efficiently translated mRNAs across eukaryotes having shorter poly(A) tail lengths may suggest that the genes involved in these pathways may have higher abundance and/or efficient translation [14].

Genes with longer poly(A) tails were significantly enriched in functional categories pivotal for more specialized and regulated cellular processes. These functions are predominantly related to signal transduction, including *signaling receptor activity, molecular transducer activity* as well as *ion binding* (Fig. 2C). Other enriched functions include *DNA-binding transcription factor activity* involved in transcriptional regulation and *immune receptor activity* related to an immune response. The longer poly(A) tails in these genes may enhance mRNA stability and translation efficiency, ensuring robust and sustained production of proteins involved in these complex and highly regulated pathways [44]. These results were reciprocated in cellular component GO enrichment analysis (Fig. 2D). In addition, we observed that the poly(A) distributions for each molecular functional pathway revealed a high degree of consistency across different samples. This consistency underscores the robustness of the poly(A) distribution patterns within each pathway, indicating that these distributions are maintained irrespective of sample variability (Supplementary Fig. 5).

KEGG pathways belonging to infection, disease-related, ribosome and oxidative phosphorylation pathways comprised transcripts with shorter poly(A) lengths and immunity-related pathways showed longer poly(A) lengths overall (Fig. 2E). This result suggests the potential stronger stability of immunity-related transcripts and high turnover of ribosomal and disease-related transcripts in patients experiencing an acute bacterial or viral infection Furthermore, we observed a bimodal distribution within one of the significant pathways—*insulin resistance* (Fig. 2E). Upon investigating further, the first peak was enriched with a set of genes, including *SOCS3, TNFRSF1A, RPS6KA1, CD36, STAT3, PTEN, MLX*, and *PRKCB*. In contrast, the second peak notably included *PYGL* and *PPP1CB*, both exhibiting relatively longer poly(A) tails. *PYGL* and *PPP1CB* encode proteins that function as phosphatases, playing critical roles in metabolic regulation [45–48]. Most genes in the first peak are actively involved in inflammatory processes and immune response, including *CD36* [49], *SOCS3* [50], *TNFRSF1A* [51], *STAT3* [52], *PTEN* [53], and *PRKCB* [54]. These results highlight the importance of visualizing poly(A) tail lengths in RNA-seq data, as they may underlie diverse regulatory mechanisms and functional outcomes.

Direct RNA sequencing uncovers hundreds of novel mRNA isoforms expressed in whole blood of patients with sepsis

Another advantageous feature of Nanopore sequencing is the ability to accurately determine novel transcript isoforms [55]. Therefore, we explored novel isoform detection in our datasets. *IsoQuant* [25] has proven to be an effective tool for transcript discovery and quantification using long RNA reads, which showed correlation with Illumina cDNA sequencing comparable to *NanoCount* (Fig. 1). We detected a total of 159,824 transcripts, of which 958 were considered novel isoforms by *IsoQuant*, with 240 non-artifact novel isoforms detected by *SQANTI3* after machine learning filtering (Supplementary Table 5). The majority of identified 958 novel isoforms fell into the categories "Novel In Catalog", "Incomplete-splice match" and "Novel Not in Catalog" (Fig. 3A). Of the 240 true isoforms, most isoforms were of the class "Combination of Known Junctions" (~ 41.7%, Fig. 3B). Overall, the set of novel transcript isoforms identified in Nanopore sequence data exhibited a wide range of inferred transcript lengths, from 331 to 8,495 nt, with a mean length of ~ 2,142 nt across all categories (Fig. 3C) and spanning all chromosomes (Supplementary Fig. 6) with a peak on chromosome 1. Consistent with previous literature [56], the identified novel isoforms were often multi-exonic, with a mean exon count of 7.9 (Fig. 3D). These results highlight the potential to discover

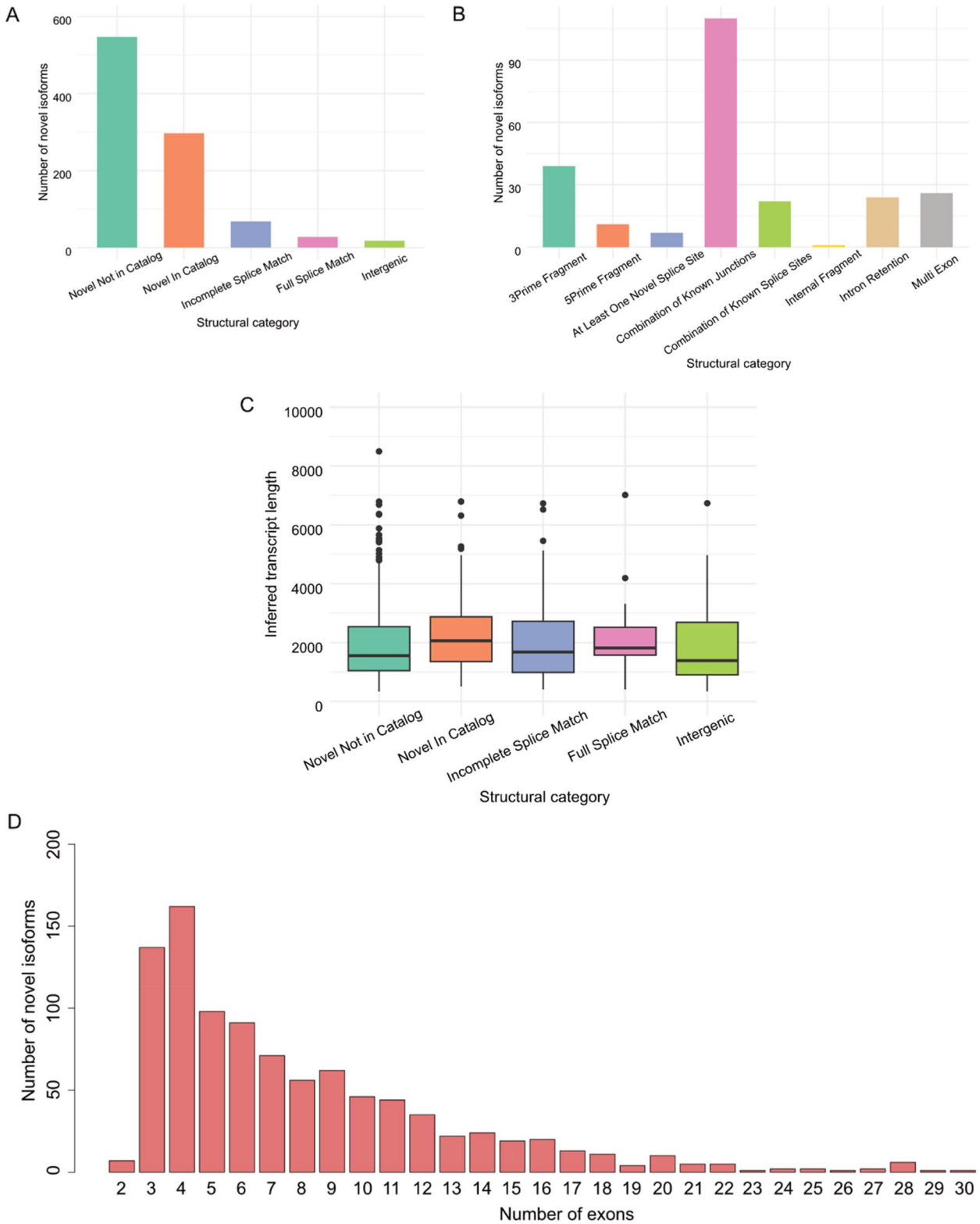

Fig. 3 Characterization of novel isoforms identified by *IsoQuant*. **A** Structural category distribution for detected novel isoforms. The structural category for an isoform indicates its relation to the closest annotated transcript. **B** Structural subcategory distribution for detected novel isoforms. **C** The length distribution of transcripts, stratified by the relation to the annotated transcripts (represented by the assigned structural category). The center line represents the median; hinges represent first and third quartiles; whiskers the most extreme values within 1.5 interquartile range from the box. **D** The exon number distribution for identified isoforms

novel isoforms using Nanopore direct RNA-seq on primary samples.

Investigating differential expression and polyadenylation between bacterial and viral infection

The samples we have studied here were a selected small subset of a larger study of 907 patients investigated via Illumina cDNA-seq for differences in host transcriptional response associated with confirmed bacterial or viral infection [23]. The bacterial and viral pathogens detected in these samples is shown in Supplementary Table 7. It was of interest to see whether we could recapitulate the major differentially expressed genes identified in this larger comparison using Nanopore direct RNA-seq. To this end, we carried out a differential gene expression analysis between Nanopore direct RNA-seq data on 6 patients with definite bacterial infection and 6 patients with definite viral infection. A total of 9 significant differentially expressed genes (DEGs) were identified when applying thresholds of adjusted *P*-value < 0.05 and $|\log_2 FC| \geq 1$. Of these, 8 DEGs were more highly expressed in patients with viral infection, while 1 was more highly expressed in patients with bacterial infection (Fig. 4A). Notably, all these 9 DEGs were consistent with DEG results obtained from Illumina cDNA-seq, in our previous work [23]. This consistency underscores the reliability and validity of our findings across different sequencing platforms.

Following this, we focused on differential polyadenylation (DP) analysis using linear mixed-effects regression (*lmer*) [33]. Through the differential polyadenylation analysis of blood from 6 patients with viral

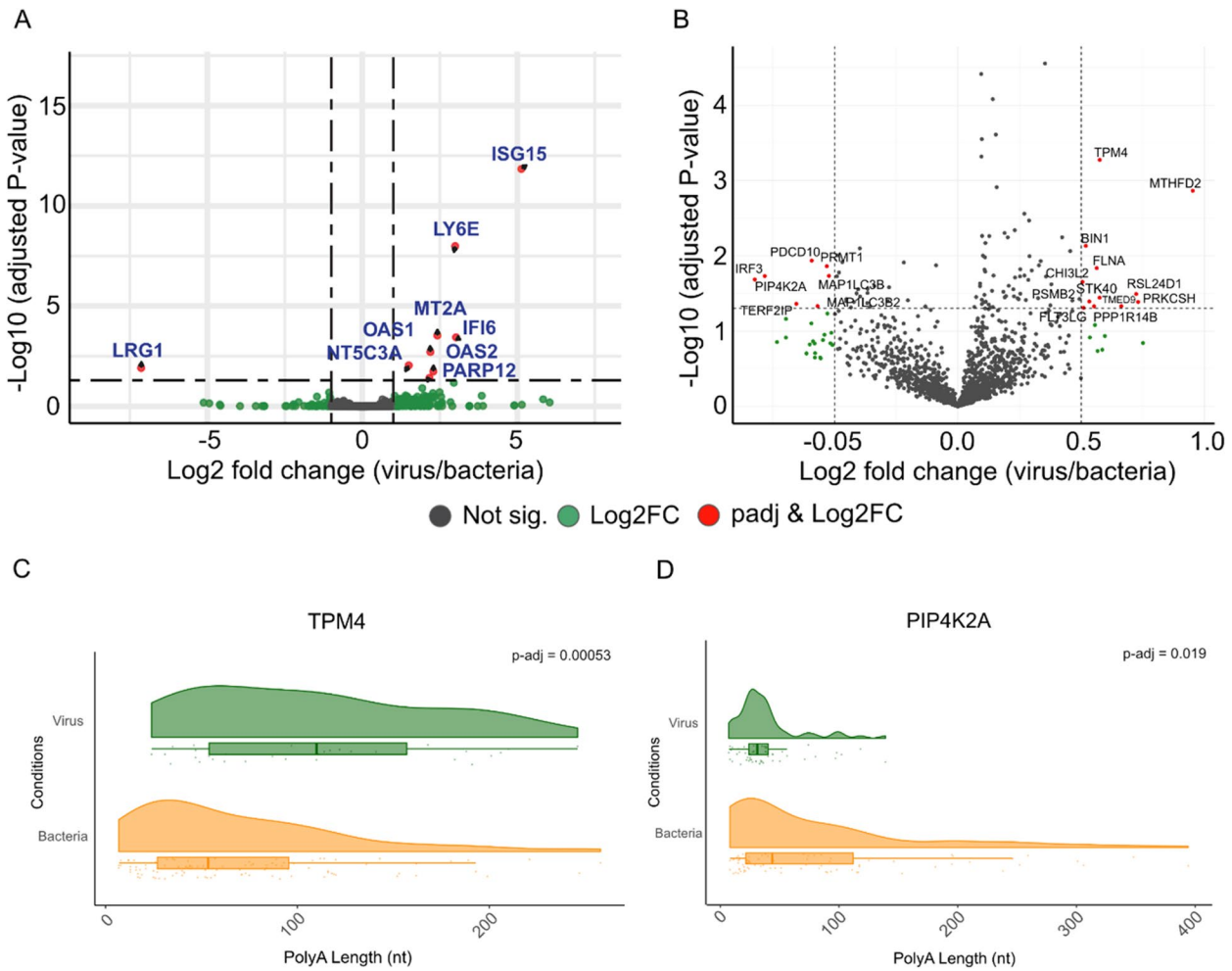

Fig. 4 Differential expression and polyadenylation differences between bacterial vs viral infection. **A** Volcano plot of viral vs bacterial differential expression from Nanopore direct RNA-seq datasets. Red dots indicate differentially expressed genes (DEGs) using adjusted *P*-value < 0.05 and $|\log_2 FC| \geq 1$ as cutoffs. **B** Volcano plot of differential polyadenylation results from linear mixed-effects regression (lmer). Red dots indicate DPGs using adjusted *P*-value < 0.05 and $|\log_2 FC| \geq 0.5$ as cutoffs. **C**-**D** Raincloud plots showing read-level polyadenylation estimates for top significantly differentially polyadenylated genes for **C** *TPM4* (adjusted *P*-value = 0.00053), **D** *PIP4K2A* (adjusted *P*-value = 0.019). Each point corresponds to a single read

infection and 6 patients with bacterial infection, using thresholds of adjusted *P*-value < 0.05 and $|\log_2 FC| \geq 0.5$, we identified 19 differentially polyadenylated genes (DPGs). Among these, 12 DPGs (*BIN1, CHI3L2, FLNA, FLT3LG, MTHFD2, PPP1R14B, PRKCSH, PSMB2, RSL24D1, STK40, TMED9, TPM4*) exhibited increased polyadenylation, and 7 (*IRF3, MAP1LC3B, MAP1LC3B2, PDCD10, PIP4K2A, PRMT1, TERF2IP*) exhibited decreased polyadenylation in the samples from patients with viral compared to bacterial infection (Fig. 4B, Supplementary Table 8).

These observed differences showed more genes with DP than differential expression (DE), although with smaller effect sizes (Fig. 4A-B). Therefore, we applied a bootstrapping method to check the sensitivity of our *lmerTest* approach. We found only 2 out of 19 genes being considered robustly DP (*TPM4* and *PIP4K2A*) (Fig. 4C-D, Supplementary Fig. 7). Therefore, secondary review of DPG's is required. PCA plots based on gene expression and average poly(A) tail lengths did not show clear separation between the viral and bacterial samples, which may in part explain the lack of significant DEGs or DPGs between these datasets (Supplementary Figs. 8A-B). Nevertheless, these results show some significant variations in the dynamic regulation of gene expression at the post-transcriptional level between viral and bacterial infections, and therefore, suggests the potential utility of polyadenylation as a disease biomarker.

Investigating differential transcript usage between patients with confirmed bacterial and viral infection

Next, we explored differential transcript usage (DTU)—the variation in the proportion of different transcript isoforms per gene across different conditions—between blood samples from patients with viral and bacterial sepsis. Using *DRIMSeq* [36] and *StageR* [37], we observed significant DTU between viral and bacterial infection samples (Supplementary Tables 9–10). In total, four genes, *SOD2, RPS21, CD36*, and *RPL37*, showed significant DTU with adjusted *P*-value < 0.05 (Supplementary Table 10). For the gene *SOD2* (*ENSG00000112096.19*), transcript *ENST00000367055.8* (adjusted *P*-value = 0.029) showed reduced usage, whereas transcript *ENST00000538183.7* (adjusted *P*-value = 0.003)

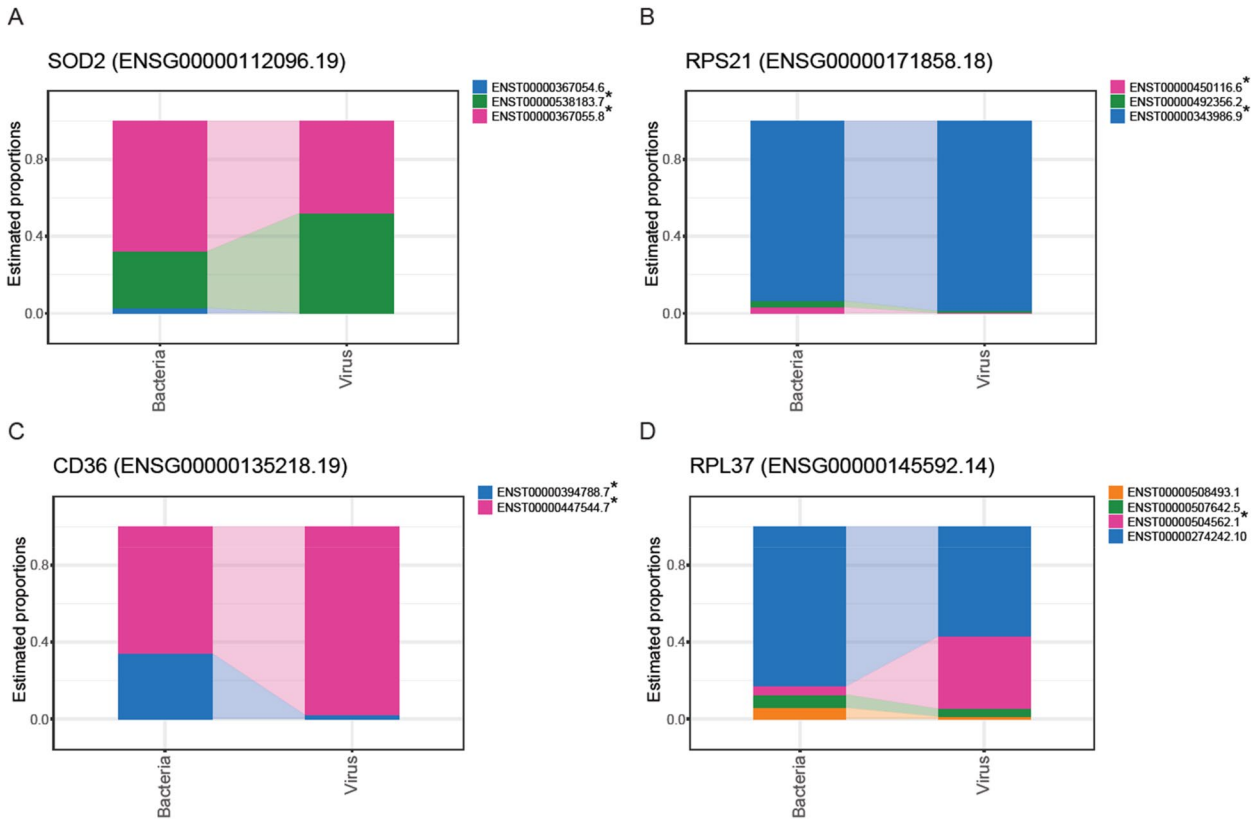

Fig. 5 Differential transcript usage occurs between bacterial and viral samples. **A**-**D** Differential estimated proportions of transcripts of genes for **A**) *SOD2* (*ENSG00000112096.19*), **B** *RPS21* (*ENSG00000171858.18*), **C** *CD36* (*ENSG00000135218.19*), and **D** *RPL37* (*ENSG00000145592.14*), with adjusted *P*-values < 0.05. Asterisks indicate transcripts which meet the adjusted *P*-value threshold of < 0.05

exhibited increased usage in samples from patients with viral compared to bacterial infection (Fig. 5A; Supplementary Fig. 9). Similar patterns of differential transcript usage were identified for the genes *RPS21* (*ENSG00000171858.18*), *CD36* (*ENSG00000135218.19*) and *RPL37* (*ENSG00000145592.14*). These genes were of interest as, *RPS21* and *RPL37* are both genes encoding ribosomal proteins [57], indicating the essential role of protein synthesis. *SOD2* is a critical regulator of antiviral signaling [58], while *CD36* is known to promote inflammatory responses and phagocytosis, processes involved in the host response to both viral and bacterial infections [49, 59, 60]. For *RPS21*, transcript *ENST00000343986.9* (adjusted *P*-value = 0.010) showed increased usage, while *ENST00000450116.6* (adjusted *P*-value = 0.002) showed reduced usage (Fig. 5B). For *CD36*, both transcripts, *ENST00000394788.7* and *ENST00000447544.7* (adjusted *P*-values = 0.000 for both), showed significant changes in usage (Fig. 5C). Lastly, for *RPL37* (*ENSG00000145592.14*), only transcript *ENST00000504562.1* (adjusted P-value = 0.003) showed increased usage (Fig. 5D;Supplementary Fig. 1).

These findings highlight the utility of Nanopore RNA-seq in uncovering differences in the host response to bacterial and viral infection. By identifying both known and novel transcripts, this technology provides critical insights into pathogen-specific gene expression, which could be pivotal for understanding the molecular mechanisms underlying viral and bacterial infections.

Discussion

Nanopore direct RNA-seq has several advantages over other RNA sequencing approaches; 1) the real-time nature of Nanopore sequencing expedites data acquisition and analysis; 2) direct analysis of RNA molecules removes the need for cDNA sequencing, hence eliminates the bias introduced by cDNA preparation; 3) it also enables continuous reads spanning many thousands of nucleotides, facilitating the identification of splice variants and novel transcript isoforms [61]; and 4) the unique 3' priming method allows the full length detection of poly(A) tails on mRNA transcripts. While each of these features holds individual utility, their combination is unparalleled and promises to yield novel insights into RNA biology.

We first underscored a high level of agreement between Nanopore direct RNA-seq and Illumina cDNA-seq of mRNA levels within our blood mRNA samples, especially with the combination of *NanoCount* for Nanopore and *Kallisto* for Illumina sequencing (Fig. 1A). Correlation analyses revealed concordance at the gene-to-gene levels (Fig. 1A-C), indicative of the reliability and consistency of both technologies in capturing gene expression profiles. In short-read sequencing, the reads are often shorter than the transcripts they originate from, leading to multiple reads aligning consecutively to the gene locus in the reference genome. This can introduce a bias in measuring expression levels, as shorter transcripts may appear to be less expressed [62]. Therefore, the high agreement levels at the gene-to-gene level were surprising. However, the transcript-level analysis showed that the correlations were lower (Supplementary Fig. 3). This is more in line with our understanding that Illumina sequencing, with its shorter read lengths, is less effective at accurately capturing isoform level information, in which its biases and lack of correct transcript assignment would be exacerbated at the transcript level. We note that the common normalized count metric for short-reads is Transcripts Per Million (TPM) which accounts for gene length and while the same TPM metric is still used widely for long-read sequencing outputs, this usually does not include gene-length normalization, and functions more like Counts Per Million (CPM). When we explored gene length-dependent bias, Nanopore data analyzed with *NanoCount* showed evidence of gene-length bias towards shorter genes using the TPM metric without accounting for gene-lengths ($p < 0.00001$) (Supplementary Fig. 4B). Long-read sequencing theoretically should reduce such biases, as a single long read can cover most of the transcript. This discrepancy may be due to Nanopore sequencing potentially overcounting shorter genes, as they pass through the Nanopore more quickly per read. It is common to find read-length distributions to be right skewed in ONT RNA-seq data, which may contribute to this phenomenon. Therefore, it may be beneficial to apply a gene-length-based normalization approach for Nanopore data like for short-read sequencing and clarifying the definition of TPM in future studies. However, despite the significant correlation between gene-length and TPM, we note that the R^2 value was low ($R^2 = 0.003$). Therefore, improved RNA-seq quantification tools and appropriate normalization protocols are needed to thoroughly address this issue and enhance the correlation between these two platforms.

Furthermore, we note that we observed better consistency between *Kallisto* and other Nanopore RNA-seq tools compared to *HTSeq* for Illumina data analysis (Supplementary Fig. 1), and this is partly because *HTSeq* does not use a probabilistic model for ambiguous reads. Given the high rates of multi-mapping in RNA-seq data, the use of probabilistic models is crucial for achieving precise abundance estimates [63]. Overall, further comparative assessments between Nanopore and Illumina RNA-seq expressions should be carried out to further examine these correlations with synthetic RNA with known concentrations, such as Sequins [64].

Variation in results from the same RNA-seq tool across different samples may arise from biological differences, such as varying gene expression levels or RNA degradation, as well as technical factors like sequencing depth or RNA quality. Sample complexity, including isoform diversity, can also contribute to variability in quantification. These factors can affect tool performance, leading to differences in transcript detection and abundance across samples [2, 65].

We utilized the unique capability of Nanopore RNA sequencing to explore polyadenylation in blood from patients with sepsis. The traditional understanding is that the average length of the poly(A) tail in mammalian mRNA is ~ 100–250 nt, at the initial synthesis stage within the nucleus. However, upon length regulation of the poly(A) tail in the cytoplasm, the steady state length of the mRNA poly(A) tail has been identified to be shorter ~ 50–100 nt [17, 66]. In our study, our results agree with the idea that the average poly(A) tail length of non-mitochondrial transcripts in human blood mRNA is closer to ~ 80 nt (Fig. 2B). Furthermore, through subsequent GSEA based on poly(A) tail lengths, we identified specific pathways enriched with variations in polyadenylation. Interestingly, infection-, disease-related, ribosome- and oxidative phosphorylation-related pathways revealed shorter poly(A) lengths and immunity-related pathways such as *JAK-STAT signalling pathway* showed longer poly(A) lengths overall (Fig. 2E). We note that the group of KEGG pathways with shorter poly(A) tails such as *Parkinsons Disease*, *Huntington Disease*, *Oxidative Phosphorylation*, *Ribosome*, *Coronavirus disease—COVID-19*, are commonly enriched together in viral infections, such as SARS-CoV-2 infections [67–70]. Considering our results were derived from patients with definite bacterial and viral infections, these findings shed light on the functional implications of altering poly(A) tail length in cellular functions, and the differential enrichment of poly(A) tail lengths across various biological pathways. Previously, transcripts with shorter poly(A) tails were shown to undergo faster rates of decay [66], which suggests the rapid regulation of these genes involved in the aforementioned pathways. Although it has been universally understood that longer poly(A) tails may lead to increased translation efficiency, a recent report suggests otherwise, where highly expressed and translated transcripts contained a shorter poly(A) tail [14]. As it stands, the relationship between poly(A) length, expression and translation are still unclear and will need further investigations. Furthermore, the number of DPGs outweighed the number of DEGs between viral and bacterial samples (Fig. 4A-B). Through this result, we highlight the potential of polyadenylation as a plausible method of biomarker discovery for disease.

Our study also revealed numerous novel isoforms through Nanopore direct RNA-seq (Supplementary Table 6), highlighting the utility of long-read sequencing in discovering novel transcripts. The identified isoforms exhibited a diverse array of characteristics and were associated with various biological processes, underscoring the complexity and heterogeneity inherent in the transcriptome. Continued efforts to understand the diversity of the transcriptome is crucial in identifying causes and treatment options for disease, and novel isoform discovery is one promising and important method of improving our understanding. As only long-read sequencing can capture the full lengths of transcripts, and therefore identify splicing patterns accurately within isoforms, we expect that Nanopore or Pacific Bioscience (PacBio) will continue to be utilized as gold standards for transcript isoform discovery in the near future.

Lastly, we identified significant differential transcript usage (DTU) for several genes between viral and bacterial samples using both known and novel transcripts from Nanopore RNA-seq (Fig. 5). While differential gene expression is widely used in RNA-seq studies, DTU explores the transcriptome at the transcript/isoform-level and is less frequently studied. This approach provides crucial insights into pathogen-specific gene expression, which are essential for understanding the molecular mechanisms underlying viral and bacterial infections. For instance, our data analysis revealed only 9 significant DEGs (Fig. 4A), but we were able to further interrogate the transcriptomic changes by visualizing the DTU at the gene level and isoform level (Fig. 5), which also highlights the potential of DTU being used for biomarker detection for disease states. However, it is important to note that not all isoforms give rise to functional proteins and their presence could be a regulatory mechanism at the post-transcriptional level for a given gene. Therefore, their direct relationship to disease states can be difficult to ascertain. Despite this, this information is still useful for biomarker discovery.

There are, however, some shortcomings associated with Nanopore direct RNA-seq [28, 71, 72] in comparison with Illumina cDNA-seq. The throughput of Nanopore direct RNA-seq remains lower than that of other high-throughput sequencing platforms, such as Illumina cDNA-seq, potentially limiting its use in large-scale studies [71, 72]. Also, most available and established pipelines have been designed and tested for Illumina cDNA-seq, whereas most Nanopore RNA pipelines are newly developed by the user community and are less maintained and kept up to date in comparison. Furthermore, input requirements for Nanopore sequencing is much higher than that of Illumina cDNA-seq, especially with direct RNA-seq protocols. Although recent developments in

direct RNA-seq have allowed for lower input requirements, the lack of a PCR step in the protocol means that for precious or low-yield samples, e.g. clinical samples, ONT direct RNA-seq may not be feasible.

This current study has various limitations. We have explored a small number of samples (6 in each condition – bacterial vs viral infection), and the results of our statistical analyses will be enhanced by incorporating an increased number of samples. Furthermore, a major advantage of Nanopore direct RNA-seq is the ability to direct post-transcriptional modifications such as nucleotide modifications [20], which we did not explore within this study. RNA modification analysis tools are rapidly evolving and being developed at unprecedented rates, with many variations in outcomes and there is currently no gold standard method for understanding RNA modifications with direct RNA-seq. Currently, the newest versions of the ONT basecaller *Dorado* can detect RNA modifications during the basecalling for direct RNA-seq datasets, which has exponentially increased the ease of analyzing the modifications. We expect that with further improvements to the *Dorado* algorithm, accurate and rapid detection of modifications will be possible, which would lead to the potential use of this technique for biomarker detection, as we have discussed with polyadenylation and DTU. Finally, further experimental validation is required for understanding how poly(A) length and DTU variations could contribute to disease mechanisms, such as via RT-qPCR and ribosomal profiling.

Conclusions

Our comparison of the two sequencing technologies—ONT direct RNA-seq and Illumina cDNA-seq—demonstrates that, with the application of a well-optimized analysis pipeline, there is a strong correlation between gene expression estimates derived from both Illumina and Nanopore platforms. While there is evidence for slight gene-length bias towards shorter genes in ONT direct RNA-seq, the method offers unique advantages not provided by Illumina cDNA-sequencing. Notably, Nanopore RNA-seq reveals critical aspects of RNA regulation, such as variations in poly(A) tail length and the discovery of novel isoforms, which are not easily detectable through Illumina cDNA-sequencing. We visualized the GO term-/KEGG pathway-specific poly(A) length distribution of human blood mRNA using ONT direct RNA-seq for the first time, to our knowledge. Additionally, our analysis identifies significant variations in poly(A) tail length that are closely related to molecular functions, offering a deeper understanding of gene expression and its regulatory mechanisms. Our results suggest that integrating Nanopore direct RNA sequencing into research workflows could significantly enhance insights into RNA regulation and gene expression, providing valuable contributions to understanding disease mechanisms.

Abbreviations

BH	Benjamini-Hochberg
BP	Biological processes
CC	Cellular components
cDNA-seq	cDNA-sequencing
CPM	Counts per million
DE	Differential expression
DEG	Differentially expressed gene
DP	Differential polyadenylation
DPG	Differentially polyadenylated gene
DTU	Differential transcript usage
FDR	False discovery rate
G-C	Guanine-cytosine
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
JSD	Jensen–Shannon Divergence
KEGG	Kyoto Encyclopedia of Genes and Genomes
Log_2FC	Log2 fold change
MF	Molecular functions
nt	Nucleotide
ONT	Oxford Nanopore Technologies
PCA	Principal component analysis
Poly(A)	Polyadenine
PCR	Polymerase chain reaction
RNA-seq	RNA-sequencing
TP	True positive
TPM	Transcripts per million
TN	True negative

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12879-025-11078-z.

Additional File 1: Figures. Supporting supplementary figures 1-9.

Additional File 2: Tables. Supporting supplementary tables 1-10.

Acknowledgements

We are grateful to all the parents and children participating this study, and the clinical and research teams who contributed to study setup, recruitment of patients, data collection and entry and monitoring. We thank the Australian Centre for Ecogenomics (ACE) Sequencing Facility of the School of Chemistry and Molecular Biosciences, University of Queensland for performing the RNA preparation and sequencing. We acknowledge all investigators of the Rapid Pediatric Infection Diagnosis in Sepsis (RAPIDS) Study Group, listed below.

Rapid Paediatric Infection Diagnosis in Sepsis (RAPIDS) Study Group:
Luregn J. Schlapbach[3,6+], Sainath Raman[3,8], Natalie Sharp[3], Natalie Phillips[3,9], Adam Irwin[10,11], Ross Balch[11], Amanda Harley[11], Kerry Johnson[11], Zoe Server[11], Shane George[3,12,13], Keith Grimwood[13,14], Peter J. Snelling[12,13], Arjun Chavan[15], Eleanor Kitkatt[15], Luke Lawton[15], Allison Hempenstall[16], Pelista Pilot[16], Kristen S. Gibbons[3], Renate Le Marsney[3], Antje Blumenthal[5], Carolyn Pardo[3], Jessica Kling[5], Stephen McPherson[5], Anna D. McDonald[3+], Seweryn Bialasiewicz[3], Trang Pham[3], Devika Ganesamoorthy[2,3], Lachlan Coin[1,2,7]
8 Paediatric Intensive Care Unit, Queensland Children's Hospital, Brisbane, Australia.
9 Emergency Department, Queensland Children's Hospital, Children's Health Queensland, Brisbane, QLD 4101, Australia.
10 Faculty of Medicine, UQ Centre for Clinical Research, University of Queensland, Brisbane, QLD 4029, Australia.
11 Infection Management and Prevention Services, Queensland Children's Hospital, Children's Health Queensland, Brisbane, QLD 4101, Australia.
12 Department of Emergency Medicine, Gold Coast University Hospital, Southport, QLD 4215, Australia.
13 School of Medicine and Dentistry, Griffith University, Southport, QLD 4222, Australia.

14 Department of Infectious Disease and Paediatrics, Gold Coast Health, Southport, QLD 4215, Australia.
15 Paediatric Intensive Care Unit, Townsville University Hospital, Townsville, QLD 4814, Australia.
16 Thursday Island Base Hospital, Thursday Island, QLD 4875, Australia.
We are grateful to all the parents and children participating this study, and the clinical and research teams who contributed to study setup, recruitment of patients, data collection and entry and monitoring. We thank the scientific teams for performing the RNA preparation and sequencing. We acknowledge all investigators of the Rapid Pediatric Infection Diagnosis in Sepsis (RAPIDS) Study Group.
+Representative: chirp@uq.edu.au

Authors' contributions
JH, DG, and LJMC developed the methodology and designed parts of the study. DG, ST, SN, HL, AB, SJM, JCK, and LJS conducted wet-lab experiments and generated the data for analysis. JH and JZ carried out the data analysis. JH, DG and LJMC contributed to the first draft of the manuscript. JH, DG, JJYC, SLT, SN, HL, AB, KSG, LJS and LJMC were involved in reviewing and editing the manuscript. All authors contributed to the article and approved the submitted version.

Funding
Research reported in this publication was supported by the Medical Research Future Fund under grant numbers GHFM76734 and MRFF9100000 (LJS) and the National Health and Medical Research Council GNT1195743 (LJMC). The study has been further funded by grants from the Children's Hospital Foundation, Brisbane, Australia (LJS, AB); Children's Hospital Foundation, Brisbane, Australia and The University of Queensland Faculty of Medicine EMCR Seed Funding (DG); Brisbane Diamantina Health Partners, Brisbane, Australia (LJS); Australian Infectious Diseases Research Centre, Brisbane, Australia (LJS, AB). LJS received a Practitioner Fellowship from the National Health and Medical Research Council (NHMRC) Australia, and support from the NOMIS foundation. KSG is supported by an NHMRC Investigator Grant. AB acknowledges support by an Australian Research Council Future Fellowship (FT220100487). None of the funding bodies had any involvement in study design, conduct, data collection, analysis, interpretation, writing of the manuscript, and the decision to submit.

Data availability
All processed sequence data such (including count matrices) are available from https://github.com/abcdtree/dRNA-ONT-blood. The raw sequence data has been deposited in UQ e-space, with the accession UQ:ac61d77. It can be accessed via the following DOI https://doi.org/10.48610/ac61d77.

Declarations

Ethics approval and consent to participate
The Children's Health Queensland Hospital and Health Service Human Research Ethics Committee; Queensland, Australia. approved the study on June 9, 2017 (HREC/17/QRCH/85). Written informed consent or delayed consent was obtained for all participants from their parents/carers. The study adhered to the WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Participants.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Clinical Pathology, The University of Melbourne, Parkville, Australia. [2]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Australia. [3]Children's Intensive Care Research Program, Child Health Research Centre, The University of Queensland, Brisbane, Australia. [4]Department of Microbiology and Immunology, The University of Melbourne, Parkville, Australia. [5]Frazer Institute, The University of Queensland, Brisbane, Australia. [6]Department of Intensive Care and Neonatology, and Children's Research Center, University Children's Hospital Zurich, University of Zurich, Zurich, Switzerland. [7]Department of Infectious Disease, Imperial College London, London, UK.

Received: 20 January 2025 Accepted: 2 May 2025
Published online: 13 May 2025

References
1. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3): R25.
2. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.
3. Marguerat S, Bahler J. RNA-seq: from technology to biology. Cell Mol Life Sci. 2010;67(4):569–79.
4. Kukurba KR, Montgomery SB. RNA sequencing and analysis. Cold Spring Harb Protoc. 2015;2015(11):951–69.
5. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Comput Biol. 2017;13(5): e1005457.
6. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53–9.
7. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. The impact of amplification on differential expression analyses by RNA-seq. Sci Rep. 2016;6(1): 25533.
8. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.
9. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-Seq data. BMC Bioinformatics. 2011;12(1): 480.
10. Feng J, Li W, Jiang T. Inference of isoforms from short sequence reads. J Comput Biol. 2011;18(3):305–21.
11. Chang JJY, Yang X, Teng H, Reames B, Corbin V, Coin L. Using synthetic RNA to benchmark poly(A) length inference from direct RNA sequencing. bioRxiv 2024.10.25.620206
12. Eckmann CR, Rammelt C, Wahle E. Control of poly(A) tail length. Wiley Interdiscip Rev RNA. 2011;2(3):348–61.
13. Biziaev N, Shuvalov A, Salman A, Egorova T, Shuvalova E, Alkalaeva E. The impact of mRNA poly(A) tail length on eukaryotic translation stages. Nucleic Acids Res. 2024;52(13):7792–808.
14. Lima SA, Chipman LB, Nicholson AL, Chen Y-H, Yee BA, Yeo GW, Coller J, Pasquinelli AE. Short poly(A) tails are a conserved feature of highly expressed genes. Nat Struct Mol Biol. 2017;24(12):1057–63.
15. Shien J-H, Su Y-D, Wu H-Y. Regulation of coronaviral poly(A) tail length during infection is not coronavirus species- or host cell-specific. Virus Genes. 2014;49(3):383–92.
16. Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP. Poly(A)-tail profiling reveals an embryonic switch in translational control. Nature. 2014;508(7494):66–71.
17. Chang H, Lim J, Ha M, Kim NV. TAIL-seq: genome-wide determination of Poly(A) tail length and 3′ end modifications. Mol Cell. 2014;53(6):1044–52.
18. Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. Nat Biotechnol. 2016;34(5):518–24.
19. Jain M, Olsen HE, Paten B, Akeson M. The Oxford nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 2016;17(1):239.
20. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021;39(11):1348–65.
21. Viehweger A, Krautwurst S, Lamkiewicz K, Madhugiri R, Ziebuhr J, Holzer M, Marz M. Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. Genome Res. 2019;29(9):1545–54.
22. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. Biol Direct. 2009;4(1): 14.
23. Schlapbach LJ, Ganesamoorthy D, Wilson C, Raman S, George S, Snelling PJ, Phillips N, Irwin A, Sharp N, Le Marsney R, et al. Host gene expression signatures to identify infection type and organ dysfunction in children evaluated for sepsis: a multicentre cohort study. Lancet Child Adolesc Health. 2024;8(5):325–38.

24. Gleeson J, Leger A, Prawer YDJ, Lane TA, Harrison PJ, Haerty W, Clark MB. Accurate expression quantification from nanopore direct RNA sequencing with NanoCount. Nucleic Acids Res. 2022;50(4): e19.
25. Prjibelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL, Tilgner HU. Accurate isoform discovery with IsoQuant using long reads. Nat Biotechnol. 2023;41(7):915–8.
26. Anders S, Pyl PT, Huber W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9.
27. Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, Love MI, Goke J. Context-aware transcript quantification from long-read RNA-seq data with Bambu. Nat Methods. 2023;20(8):1187–95.
28. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525–7.
29. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. GENCODE: the reference human genome annotation for the ENCODE project. Genome Res. 2012;22(9):1760–74.
30. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47(D1):D766–73.
31. Oxford Nanopore Technologies D. https://github.com/nanoporet,ech/dorado.
32. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation (Camb). 2021;2(3):100141.
33. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. J Stat Softw. 2017;82(13):1–26.
34. Allen M, Poggiali D, Whitaker K, Marshall TR, van Langen J, Kievit RA. Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Res. 2019;4:63.
35. Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomas J, Amorin R, Estevan-Morio E, Liu T, Nanni A, McIntyre L, et al. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. Nat Methods. 2024;21(5):793–7.
36. Nowicka M, Robinson MD. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. F1000Res. 2016;5:1356.
37. Van den Berge K, Soneson C, Robinson MD, Clement L. stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. Genome Biol. 2017;18(1):151.
38. Temperley RJ, Wydro M, Lightowlers RN, Chrzanowska-Lightowlers ZM. Human mitochondrial mRNAs–like members of all families, similar but different. Biochim Biophys Acta. 2010;1797(6–7):1081–5.
39. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat Methods. 2019;16(12):1297–305.
40. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
41. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, Wadi L, Meyer M, Wong J, Xu C, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc. 2019;14(2):482–517.
42. Joly JH, Lowry WE, Graham NA. Differential gene set enrichment analysis: a statistical approach to quantify the relative enrichment of two gene sets. Bioinformatics. 2021;36(21):5247–54.
43. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284–7.
44. Passmore LA, Coller J. Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression. Nat Rev Mol Cell Biol. 2022;23(2):93–106.
45. Ji Q, Li H, Cai Z, Yuan X, Pu X, Huang Y, Fu S, Chu L, Jiang C, Xue J, et al. PYGL-mediated glucose metabolism reprogramming promotes EMT phenotype and metastasis of pancreatic cancer. Int J Biol Sci. 2023;19(6):1894–909.
46. Ferreira M, Beullens M, Bollen M, Van Eynde A. Functions and therapeutic potential of protein phosphatase 1: Insights from mouse genetics. Biochim Biophys Acta Mol Cell Res. 2019;1866(1):16–30.
47. Luo X, Zhang Y, Ruan X, Jiang X, Zhu L, Wang X, Ding Q, Liu W, Pan Y, Wang Z, et al. Fasting-induced protein phosphatase 1 regulatory subunit contributes to postprandial blood glucose homeostasis via regulation of hepatic glycogenesis. Diabetes. 2011;60(5):1435–45.
48. Chen B, Chen X, Hu R, Li H, Wang M, Zhou L, Chen H, Wang J, Zhang H, Zhou X, et al. Alternative polyadenylation regulates the translation of metabolic and inflammation-related proteins in adipose tissue of gestational diabetes mellitus. Comput Struct Biotechnol J. 2024;23:1298–310.
49. Silverstein RL, Febbraio M. CD36, a scavenger receptor involved in immunity, metabolism, angiogenesis, and behavior. Sci Signal. 2009;2(72):re3.
50. Carow B, Rottenberg ME. SOCS3, a major regulator of infection and inflammation. Front Immunol. 2014;5:58.
51. Jang DI, Lee AH, Shin HY, Song HR, Park JH, Kang TB, Lee SR, Yang SH. The Role of Tumor Necrosis Factor Alpha (TNF-α) in Autoimmune Disease and Current TNF-α Inhibitors in Therapeutics. Int J Mol Sci. 2021;22(5):2719.
52. Sulczewski FB, Martino LA, Salles D, Yamamoto MM, Rosa DS, Boscardin SB. STAT3 signaling modulates the immune response induced after antigen targeting to conventional type 1 dendritic cells through the DEC205 receptor. Front Immunol. 2022;13: 1006996.
53. Taylor H, Laurence ADJ, Uhlig HH. The role of PTEN in innate and adaptive immunity. Cold Spring Harb Perspect Med. 2019;9(12):a036996.
54. Altman A, Kong KF. Protein kinase C inhibitors for immune disorders. Drug Discov Today. 2014;19(8):1217–21.
55. Clark MB, Wrzesinski T, Garcia AB, Hall NAL, Kleinman JE, Hyde T, Weinberger DR, Harrison PJ, Haerty W, Tunbridge EM. Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain. Mol Psychiatry. 2020;25(1):37–47.
56. Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain S. A comprehensive examination of nanopore native RNA sequencing for characterization of complex transcriptomes. Nat Commun. 2019;10(1):3359.
57. Li S. Regulation of ribosomal proteins on viral infection. Cells. 2019;8(5):508.
58. Wang W, Jin Y, Zeng N, Ruan Q, Qian F. SOD2 facilitates the antiviral innate immune response by scavenging reactive oxygen species. Viral Immunol. 2017;30(8):582–9.
59. Philips JA, Rubin EJ, Perrimon N. Drosophila RNAi screen reveals CD36 family member required for mycobacterial infection. Science. 2005;309(5738):1251–3.
60. Hoebe K, Georgel P, Rutschmann S, Du X, Mudd S, Crozat K, Sovath S, Shamel L, Hartung T, Zahringer U, et al. CD36 is a sensor of diacylglycerides. Nature. 2005;433(7025):523–7.
61. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;21(1):30.
62. Lagarrigue S, Martin L, Hormozdiari F, Roux PF, Pan C, van Nas A, Demeure O, Cantor R, Ghazalpour A, Eskin E, et al. Analysis of allele-specific expression in mouse liver by RNA-Seq: a comparison with Cis-eQTL identified using genetic linkage. Genetics. 2013;195(3):1157–66.
63. Turro E, Su SY, Goncalves A, Coin LJ, Richardson S, Lewin A. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. Genome Biol. 2011;12(2): R13.
64. Beucher G, Blondot ML, Celle A, Pied N, Recordon-Pinson P, Esteves P, Faure M, Métifiot M, Lacomme S, Dacheux D, et al. Bronchial epithelia from adults and children: SARS-CoV-2 spread via syncytia formation and type III interferon infectivity restriction. Proceedings of the National Academy of Sciences. 2022;119(28):e2202370119.
65. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, Mcpherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, et al. A survey of best practices for RNA-seq data analysis. Genome Biology. 2016;17(1):13.
66. Eisen TJ, Eichhorn SW, Subtelny AO, Lin KS, McGeary SE, Gupta S, Bartel DP. The dynamics of cytoplasmic mRNA metabolism. Mol Cell. 2020;77(4):786-799 e710.
67. Chang JJY, Gleeson J, Rawlinson D, De Paoli-Iseppi R, Zhou C, Mordant FL, Londrigan SL, Clark MB, Subbarao K, Stinear TP, et al. Long-read RNA sequencing identifies polyadenylation elongation and differential transcript usage of host transcripts during SARS-CoV-2 in vitro infection. Front Immunol. 2022;13:832223.
68. Chen J-C, Xie T-A, Lin Z-Z, Li Y-Q, Xie Y-F, Li Z-W, Guo X-G. Identification of key pathways and genes in SARS-CoV-2 infecting human intestines by bioinformatics analysis. Biochem Genet. 2022;60(3):1076–94.
69. Fang KY, Liang GN, Zhuang ZQ, Fang YX, Dong YQ, Liang CJ, Chen XY, Guo XG. Screening the hub genes and analyzing the mechanisms in

discharged COVID-19 patients retesting positive through bioinformatics analysis. J Clin Lab Anal. 2022;36(7):e24495.
70. Khalid Z, Huan M, Sohail Raza M, Abbas M, Naz Z, Kombe Kombe AJ, Zeng W, He H, Jin T. Identification of novel therapeutic candidates against SARS-CoV-2 infections: an application of RNA sequencing toward mRNA based nanotherapeutics. Frontiers in Microbiology. 2022;13:901848.
71. Chen J, Xu F. Application of nanopore sequencing in the diagnosis and treatment of pulmonary infections. Mol Diagn Ther. 2023;27(6):685–701.
72. Delahaye C, Nicolas J. Sequencing DNA with nanopores: troubles and biases. PLoS One. 2021;6(10): e0257521.

Publisher's Note