Song *et al. Genome Biology* (2025) 26:76
https://doi.org/10.1186/s13059-025-03536-3

Genome Biology

**METHODOLOGY** **Open Access**

# DEMINERS enables clinical metagenomics and comparative transcriptomic analysis by increasing throughput and accuracy of nanopore direct RNA sequencing

Junwei Song[1†], Li-an Lin[1†], Chao Tang[1,3†], Chuan Chen[2,4†], Qingxin Yang[1], Dan Zhang[1], Yuancun Zhao[1], Han-cheng Wei[2,3], Kepan Linghu[1], Zijie Xu[1], Tingfeng Chen[1], Zhifeng He[1], Defu Liu[1], Yu Zhong[3], Weizhen Zhu[6], Wanqin Zeng[1], Li Chen[1], Guiqin Song[4], Mutian Chen[2], Juan Jiang[5], Juan Zhou[2], Jing Wang[5], Bojiang Chen[5], Binwu Ying[2], Yuan Wang[2], Jia Geng[2*], Jing-wen Lin[2,3*] and Lu Chen[1*]

[†]Junwei Song, Li-an Lin, Chao Tang and Chuan Chen contributed equally to this work.

*Correspondence: geng.jia@scu.edu.cn; lin.jingwen@scu.edu.cn; luchen@scu.edu.cn

[1] Department of Laboratory Medicine, Key Laboratory of Birth Defects and Related Diseases of Women and Children of MOE, State Key Laboratory of Biotherapy, West China Second University Hospital, Sichuan University, Chengdu 610041, China
[2] Department of Laboratory Medicine, State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, Sichuan University and Collaborative Innovation Center, Chengdu 610041, China
Full list of author information is available at the end of the article

## Abstract

Nanopore direct RNA sequencing (DRS) is a powerful tool for RNA biology but suffers from low basecalling accuracy, low throughput, and high input requirements. We present DEMINERS, a novel DRS toolkit combining an RNA multiplexing workflow, a Random Forest-based barcode classifier, and an optimized convolutional neural network basecaller with species-specific training. DEMINERS enables accurate demultiplexing of up to 24 samples, reducing RNA input and runtime. Applications include clinical metagenomics, cancer transcriptomics, and parallel transcriptomic comparisons, uncovering microbial diversity in COVID-19 and $m^6A$'s role in malaria and glioma. DEMINERS offers a robust, high-throughput solution for precise transcript and RNA modification analysis.

**Keywords:** Nanopore direct RNA sequencing, Demultiplex, Basecalling, RNA modification, Machine learning

## Background

The advance of third-generation sequencing technologies has revolutionized our ability to perform long-read sequencing for genomes and transcriptomes [1]. Oxford Nanopore Technologies' (ONT) direct RNA sequencing (DRS) exemplifies this revolution, enabling the direct sequencing of DNA, poly(A) RNA, and their modifications without fragmentation or amplification steps. This technology employs an array of protein nanopores in a synthetic membrane, allowing the sequencing of RNA molecules directly [2]. As motor enzymes ratchet a single RNA molecule through the pore, it causes electric current fluctuations when different nucleotides block the pore, a process that can be recorded to determine the RNA sequence using basecalling algorithms [3].

DRS technology, capable of producing long sequencing reads up to 21 kb [4], is powerful in obtaining full-length transcripts [5, 6] and nearly complete RNA genomes [7, 8]. It is particularly effective in measuring RNA poly(A) tail length and analyzing the regulation function of these tails over gene expression [9–11]. Furthermore, DRS can detect RNA modification directly [12–15], such as N6-methyladenosine (m$^{[[6]]}$A), 7-methylguanosine (m7G) [16], N1-methylpseudouridine [17], pseudouridine (Ψ) and 2′-O-methylation (Nm)[15, 18], and others, due to the unique current fluctuations these modifications produce.

The amplification-free library preparation method for DRS can sidestep amplification or reverse transcription (RT) bias [19], making it particularly useful in identifying exitrons (exonic introns) that likely arise from reverse transcription artifacts [20]. DRS is particularly useful in identifying isoform-specific RNA structure [21], tRNA [22, 23], and rRNA [16] modification, as well as in genome sequencing of RNA viruses [7, 8, 24, 25]. Another application of DRS is to analyze the dynamics of RNA metabolism by labeling nascent RNAs with base analogs (for example, 5-ethynyluridine [26] and 4-thiouridine [27]). DRS has been applied in studies on (epi)transcriptome analyses across a wide range of species, including humans [4, 28], animals (nematodes [6, 9], insects [29, 30], etc.), plants [31–33], and microorganisms including bacteria [34, 35], archaea [36], yeast [37], zooplankton [38], viruses [11, 39–44], and parasites [45–47]. Additionally, DRS has been applied in pathogen identification in clinical samples [24, 48, 49].

Despite the advantages, to date DRS still faces several challenges. Firstly, DRS requires high poly(A) RNA input which is especially challenging when dealing with rare or limited biological samples. Secondly, one RNA sample per flow cell leads to high cost and technical complexity, posing a substantial barrier in large-scale studies. Multiplexing samples in one flow cell will reduce the required poly(A) RNA input, the effort in library construction, and the sequencing cost without compromising data quality [50]. Moreover, the multiplexing strategy minimizes batch effects that arise from processing samples individually, leading to more consistent and reliable data across different sequencing runs. Current methods like Poreplex [51] and DeePlexiCon [52], limited to demultiplexing only four samples, highlight the need for improved methodologies to enhance throughput, reduce costs, and minimize technical variability in DRS applications.

However, basecalling during the demultiplexing process is quite challenging due to the lower sampling rate of RNA at 70 bases per second through nanopores, in contrast to 450 bases per second of DNA. This significant difference necessitates the development of specialized basecalling methods tailored to the distinct signal characteristics inherent to RNA. Currently, the median basecalling accuracy for DRS stands at about 86% [53], necessitating a leap over the 90% threshold for high basecalling accuracy to reliably document genetic elements such as short exons and exon junctions [53]. This is especially crucial for RNA viruses and pathogens with high mutation rates. Although the utilization of convolutional neural networks (CNN) has improved basecalling accuracy [54], no studies have developed species-specific basecalling to mitigate the challenge of low accuracy in species-specific studies.

Here, we introduce DEMINERS (*D*emultiplexing and *E*valuation using *M*achine-learning *I*ntegrated in *N*anopor*e* direct *R*NA *S*equencing), an innovative machine-learning framework that significantly improved the efficiency and scalability for DRS. DEMINERS employed a robust demultiplexing workflow that utilizes a Random Forest classifier named DecodeR to accurately demultiplex up to 24 samples per run. DEMINERS therefore reduces the required poly(A) RNA input and sequencing costs, minimizing batch effects while maintaining high accuracy, thereby improving the consistency and reliability of comparative studies in RNA biology. Furthermore, DEMINERS incorporated a novel basecaller named Densecall built on an optimized CNN architecture, achieving state-of-the-art basecalling performance across different DRS datasets from animals, plants, and microorganisms. Additionally, DEMINERS enables comprehensive downstream applications, including gene/isoform expression profiling, RNA variant and modification identification, assembly the genome of RNA virus, and meta-genomic/transcriptomic analyses, providing reliable performance for diverse research purposes. Overall, DEMINERS improves the scalability and accuracy, while reduces the cost of DRS, facilitating the exploration of complex transcriptomic features in different organisms.

## Results

### Overview of DEMINERS

We first developed an experimental workflow for RNA multiplexing that integrates adapter-ligation and sample barcoding to ligate each RNA sample to a unique RNA transcription adapter (RTA) (Fig. 1a). The pooled RNA library is processed through a nanopore sequencer, generating raw electrical signals. For demultiplexing, DEMINERS utilized a Random Forest algorithm named DecodeR to classify the segmented current signals, allowing up to 24 samples to be demultiplexing simultaneously (Fig. 1b).

Next, to improve the accuracy of DRS basecalling, we reconstructed a convolutional basecaller named Densecall inspired by DenseNet [55], originally designed for image feature extraction. We adopted the reconstructed architecture to process one-dimensional electrical signals for basecalling, ensuring direct connection of each layer to the subsequent ones (Fig. 1c). Moreover, DEMINERS provides species-specific basecalling models to further increase the accuracy for diverse species (Methods).

For downstream analysis, we developed a comprehensive gene/isoform analysis workflow that facilitates the identification of novel transcripts by employing an expectation–maximization (EM) algorithm to estimate isoform abundance (Methods). DEMINERS also enables detection of RNA variations and modifications, genome assembly for RNA viruses, and meta-genomic/transcriptomic analyses (Fig. 1d).

### Maximizing throughput and accuracy in demultiplexing

We designed and synthesized 48 RNA transcription adapters (RTAs), containing barcodes with lengths varying between 22 and 28 nucleotides (nt) (Additional file 3: Data S1). The design led to an increased Hamming distance amongst barcodes as their length extended, facilitating better differentiation (Additional file 1: Fig. S1a). These RTAs were then ligated to 51 in vitro transcribed (IVT) RNAs and 5 DRS runs were performed, yielding a total of 6,276,168 valid reads uniquely aligned to the reference sequences.

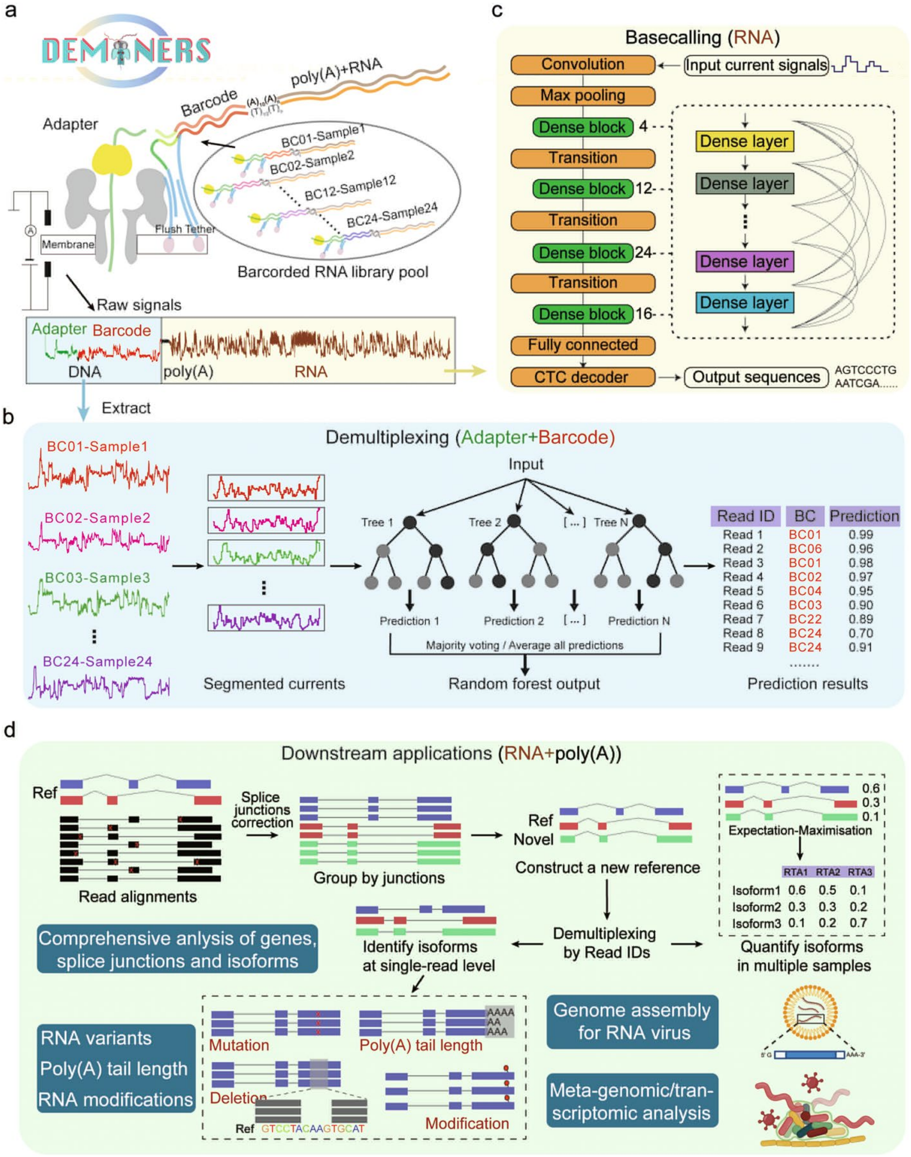

**Fig. 1** The overview of DEMINERS. **a** Experimental workflow of barcoding and direct RNA sequencing pipeline. Each sample is ligated to an RNA transcription adapter (RTA) containing an RNA adaptor, a barcode (BC), and poly(T). These barcoded RNA samples are sequenced in a flowcell, producing raw signals of multiplexed barcodes and RNAs. **b** Schematic illustration of DEMINERS machine-learning classifier based on Random Forest. The current signals of adapters and barcodes were extracted based on the distinct current changes introduced by poly(A) tails. The currents were then segmented into 100 segments/units according to the current changepoints. The normalized and segmented current signals were used as input for barcode classification based on random forest algorithm. **c** Representation of DEMINERS basecaller built on an optimized convolutional neural network. The basecalling architecture employs an inter-layer connection strategy to foster feature reuse and mitigate the vanishing gradient. The basecaller incorporates a convolutional layer for denoising, followed by max pooling and 4 densely connected convolutional networks (Dense blocks) to decode raw current signals. The 4 dense blocks containing 6, 12, 24, and 16 dense layers, respectively. Then a fully connected layer with a log softmax activation is used for classification and a connectionist temporal classification (CTC) decoder outputs nucleotide sequence. **d** Overview of downstream applications of DEMINERS. In this study, we show that the DRS reads retrieved by DEMINERS can be used for comprehensive analysis of genes, splice junctions, and isoforms. The splice junctions are corrected and grouped by junctions to construct a new reference. Isoforms are quantified using an expectation–maximization algorithm. By matching the demultiplexed read IDs, the mutations, deletions, poly(A) tail lengths, and RNA modifications can be identified at the single-read level. Additionally, DEMINERS can perform genome assembly for RNA viruses and support meta-genomic/transcriptomic analysis.

Considering the impact of sequencing depth on subsequent model training, we selected 24 barcodes with over 80,000 sequencing reads for further study (Additional file 3: Data S1). Notably, we observed that barcodes of mixed lengths significantly improved the performance of DEMINERS compared to using barcodes with uniformed 20-nt length (Additional file 1: Fig. S1b).

To advance the demultiplexing capability, we extracted signals of adapters and barcodes from the raw electrical signals based on the distinct current changes introduced by poly(A) tails. The optimal signal changepoints were identified according to the mean and variance of the normalized currents. The normalized currents were then segmented into small units, with the average current value of each segment/unit assigned as the feature values, which is used to train the classifier using different machine-learning algorithms (Fig. 1c, Additional file 1: Fig. S1c, Methods). This step retained the patterns of current signals while reduces the noises. We evaluated 6 different classification algorithms, including *k*-nearest neighbor (KNN), neural network (NNET), naïve Bayes (NB), classification and regression tree (CART), AdaBOOST, and Random Forest (RF) to classify 4 barcodes in the test datasets (Additional file 3: Data S1). After plotting Receiver Operator Characteristic (ROC) and precision-recall (PR) curves for the 6 classifiers, we found that the area under the ROC curve (AUROC) were all above 0.93 (Additional file 1: Fig. S1d) and the area under the precision-recall curve (AUPRC) all greater than 0.83 (Additional file 1: Fig. S1e). Notably, RF outperformed the remaining methods in all parameters we measured, achieving the highest AUROC of 0.9961, the highest AUPRC of 0.9906 and the highest accuracy of 95.67%, and the highest recall of 95.66% across all reads (Additional file 1: Fig. S1d-f, Additional file 2: Table S1). The barcode classification of DEMINERS was thus built on RF algorithm.

Next, we evaluated the performance of DEMINERS in classifying different numbers of barcodes (from 2 to 24) and found the AUROCs were all higher than 0.99 (Fig. 2a). Moreover, AUPRCs were higher than 0.98 when classifying 10 barcodes, but still reached 0.95 when classifying 24 barcodes (Fig. 2b, Additional file 2: Table S2). We then assessed the accuracy and recovery rates of different numbers of barcodes (Fig. 2c). For instance, with a predicted probability cutoff set-off of 0.5, DEMINERS achieved an accuracy of

(See figure on next page.)

**Fig. 2** Performance of different demultiplexing and basecalling methods. **a**,**b** Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves showing the performance of DEMINERS in demultiplexing direct RNA-seq (DRS) data generated with 2 to 24 barcodes. AUROC is the area under the ROC curve, and AUPRC is the area under the PR curve. **c** Accuracy and recovery rates of DEMINERS in demultiplexing DRS data with different numbers of barcodes. The cross symbol represents the cutoff of predicted probability is at 0.5. **d** Bar charts representing the precision, recall, and F1-score of DEMINERS and Poreplex [51] classifying 4 Poreplex barcodes. **e** Bar charts representing precision, recall, and F1-score of DEMINERS and DeePlexiCon classifying 4 DeePlexiCon barcodes. **f** Comparison of DEMINERS and DeePlexiCon [52] classifying 24 custom-designed barcodes. **g** Box plots of the accuracy, mismatch, deletion, and insertion rates of DEMINERS, RODAN [56], and Guppy in basecalling of the 10-species test set. The boxes show the median and lower/upper quantile, the dots indicate the outliers, and the *P* values were determined by Wilcoxon test. *P* values < 0.05 are highlighted in red. **h** Box plots showing the accuracy of DEMINERS, RODAN, and Guppy in basecalling DRS data of 5 different species. Each dataset was run for 8 times to ensure robustness, and *P* values were determined by Wilcoxon test. *P* values < 0.05 are highlighted in red. **i** Bar chats representing the accuracy, mismatch, deletion, and insertion rates of DEMINERS (mouse-specific and general modes), RODAN, and Guppy.

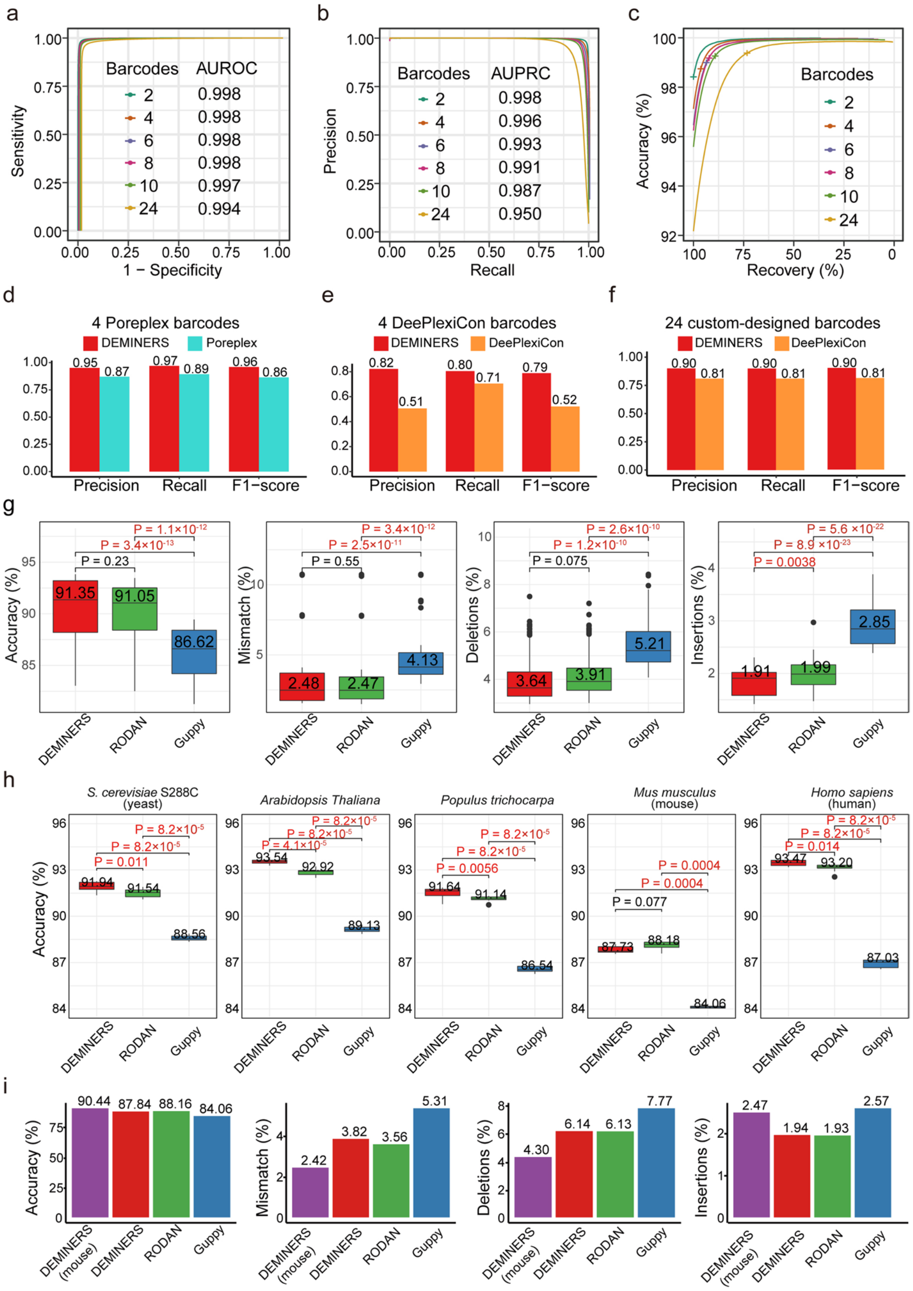

**Fig. 2** (See legend on previous page.)

99.3% and an 89.2% recovery rate when classifying 10 barcodes, and 99.4% accuracy and 73.3% recovery for 24 barcodes (Fig. 2c, Additional file 1: Fig. S2a, Additional file 3: Data S1).

Furthermore, we conducted comparison analysis of DEMINERS with two existing methods, Poreplex [51] and DeePlexiCon [52]. The evaluation was carried out using DRS datasets generated with the 4 barcodes originally designed for Poreplex or DeePlexiCon, and 24 barcodes designed in this study (Additional file 4: Data S2). When comparing DEMINERS to Poreplex using Poreplex 4-barcode dataset, the classification accuracies of DEMINERS ranged from 95.82 to 97.83% (mean ± standard deviation: 96.98 ± 0.84), whereas the accuracies of Poreplex ranged from 83.4 to 93.18% (89.4 ± 4.21) (Additional file 1: Fig. S2b). When demultiplexing DeePlexiCon 4-barcode dataset, the classification accuracies of DEMINERS were 90.57 to 94.43% (92.29 ± 1.60), much higher than that of DeePlexiCon, ranging from 51.33 to 85.22% (68.39 ± 15.16) (Additional file 1: Fig. S2c). Moreover, DEMINERS excelled DeePlexiCon and Poreplex in parallel comparisons in all measured characteristics, including precision, recall, and F1-score (Fig. 2d-f), as well as accuracy, sensitivity, and specificity (Additional file 1: Fig. S2d-f, Additional file 2:Table S3). When demultiplexing 24 barcodes, DEMINERS achieved accuracies between 83.24 and 93.78% (89.86 ± 0.03), higher than the accuracies of DeePlexiCon, ranging from 69.67 to 87.97% (80.61 ± 0.05) (Additional file 1: Fig. S3a). Although the sensitivity and specificity of DEMINERS and DeePlexiCon were comparable (Additional file 1: Fig. S3b), DEMINERS significantly outperformed DeePlexiCon in terms of precision, accuracy, and recovery (Additional file 1: Fig. S3c-d). We further assessed the performance of DEMINERS and DeePlexiCon when processing the same amount of DRS reads. The system CPU time of DEMINERS was significantly faster than DeePlexiCon (approximately 12 times and 9 times faster than that of DeePlexiCon in CPU and GPU model, respectively, Additional file 1: Fig. S3e) with a much lower CPU usage (DEMINERS, 185.95 ± 14.93% compared to DeePlexiCon 926.67 ± 32.34%, Additional file 1: Fig. S3f). Moreover, the maximum memory consumption of DEMINERS was around seven times less than that of DeePlexiCon (Additional file 1: Fig. S3g). In summary, DEMINERS effectively classified up to 24 barcodes with high accuracy and precision, outperforming existing methods like Poreplex and DeePlexiCon in accuracy, precision, recall, and running time.

### Improving the basecalling accuracy

To improve the accuracy of DRS basecalling, we reconstructed a convolutional architecture inspired by DenseNet [55] and provided a training module for species-specific basecalling (Methods). The basecaller of DEMINERS incorporated convolutional layers for denoising, max pooling for down-sampling, 4 dense blocks with transition layers for feature processing, and connectionist temporal classification (CTC) [57] loss for gradient descent (Fig. 1c). In addition, we employed a memory optimization technique [58] that significantly reduces memory consumption during model training, exemplified by reducing GPU memory usage from 6628 to 4700 MB with a batch size of 32 and a chunk size of 4096.

To evaluate the accuracy, we first trained the basecaller on a dataset containing 4 species: including human, *C. elegans*, *E. coli*, and *Arabidopsis thaliana*, previously used by RODAN [54]. For the test set, we used an integrated 10 species dataset that includes 2 previous published datasets [11, 54] and the in-house datasets, including four viruses [SARS-CoV-2, Porcine reproductive and respiratory syndrome virus (PRRSV), Seneca Valley virus (SVV), Porcine epidemic diarrhea virus (PEDV)], two plants (*A. Thaliana* and *Populus trichocarpa*), one fungus (*S. cerevisiae*), one eukaryotic parasite (*Plasmodiun berghei*), and two mammals (human and mouse) (Additional file 2: Table S4). The accuracy of basecalling the test sets of each species was evaluated by DEMINERS, RODAN [54], and ONT-Guppy. We found that DEMINERS and RODAN significantly outperformed Guppy with higher accuracy, and lower mismatch, deletion, and insertion rates (Fig. 2g). Compared to RODAN, DEMINERS had comparable accuracy, and mismatch and deletion rates, but fewer insertions (Fig. 2g).

Notably, DEMINERS showed higher basecalling accuracy than RODAN for 4 species, including *S. cerevisiae*, *A. Thaliana*, *P. trichocarpa*, and human, except for mouse (Fig. 2h, Additional file 2: Table S4). To further improve the accuracy, we integrated a species-specific model training mode in DEMINERS. Using the RODAN mouse dataset to train the mouse-specific model, the basecalling accuracy for DEMINERS was improved from 87.84 to 90.44%, surpassing 88.16% accuracy of RODAN and 84.06% of Guppy (Fig. 2i). The mismatch and deletion rates were lower in DEMINERS mouse-specific model, but the insertion rate was higher (Fig. 2i, Additional file 2: Table S4). Overall, DEMINERS demonstrated higher basecalling accuracy and lower mismatch rates compared to Guppy and RODAN, especially after integrating a species-specific model, leading to improved performance in basecalling species-specific datasets, and facilitating transcript assembly of the non-model organisms.

### Accurate pathogen identification, variant calling, and genome assembly of RNA virus from multiplexed RNA samples

To assess whether DEMINERS can correctly distinguish different species and call variants from multiplexed RNA samples, we tested the performance of DEMINERS on three experimentally pooled RNA samples (Fig. 3a), containing 3 or 5 different species (Additional file 5: Data S3). The following 8 pathogens were used in the study: RNA viruses [Seneca Valley virus (SVV), Porcine epidemic diarrhea virus (PEDV), Porcine reproductive and respiratory syndrome virus (PRRSV), and Getah virus (GETV)], bacteria (*E. coli* and *S. enteritidis*), fungus (*S. cerevisiae*), and a parasite (*P. berghei*). For all three experiments, DEMINERS correctly identified the species with AUROC above 0.98 (Fig. 3b) and AUPRC above 95% (Fig. 3c), while achieving accuracy above 96% (Fig. 3d). Next, we merged the 3 datasets and successfully identified all species by metagenomic analysis (Fig. 3e) and found that the genomes of RNA viruses were significantly longer than transcripts of nonviral microorganisms (RNA viruses, mean: 2093; nonviral microorganisms, mean: 880; $P < 2.2\mathrm{e}-16$, Wilcoxon test). These results demonstrated the reliability of our method and its potential in meta-genomic/transcriptomic applications.

Next, we assessed whether DEMINERS retrieved-DRS sequences can be used for genome assembly of RNA virus. For SVV (genome 7,310-nt in length) and PRRSV (15,428-nt) where the sequence depths were adequate, the constructed genomes had 96.5 and 95.6% coverage with 96.1 and 95.6% identity compared to their reference genomes, respectively. Notably, there were 835 reads longer than 7000-nt aligned to the SVV genome and 42 reads longer than 15,000-nt aligned to the PRRSV genome (Fig. 3f, Additional file 1: Fig. S4a), showing the capability of DEMINERS in assembly of RNA virus genomes.

Since RNA viruses have high mutation rates, we further analyzed the single-nucleotide variants (SNVs) in their genomes. For SVV, 414 and 423 SNVs were identified in two sequencing experiments (Exp1 and Exp2), of which 96.48% (411) were common with a high correlation rate (Fig. 3g, Person correlation $R = 0.94$). For example, a C-to-V mutation and a 3-bp deletion of SVV were identified in both experiments (Fig. 3h, Additional

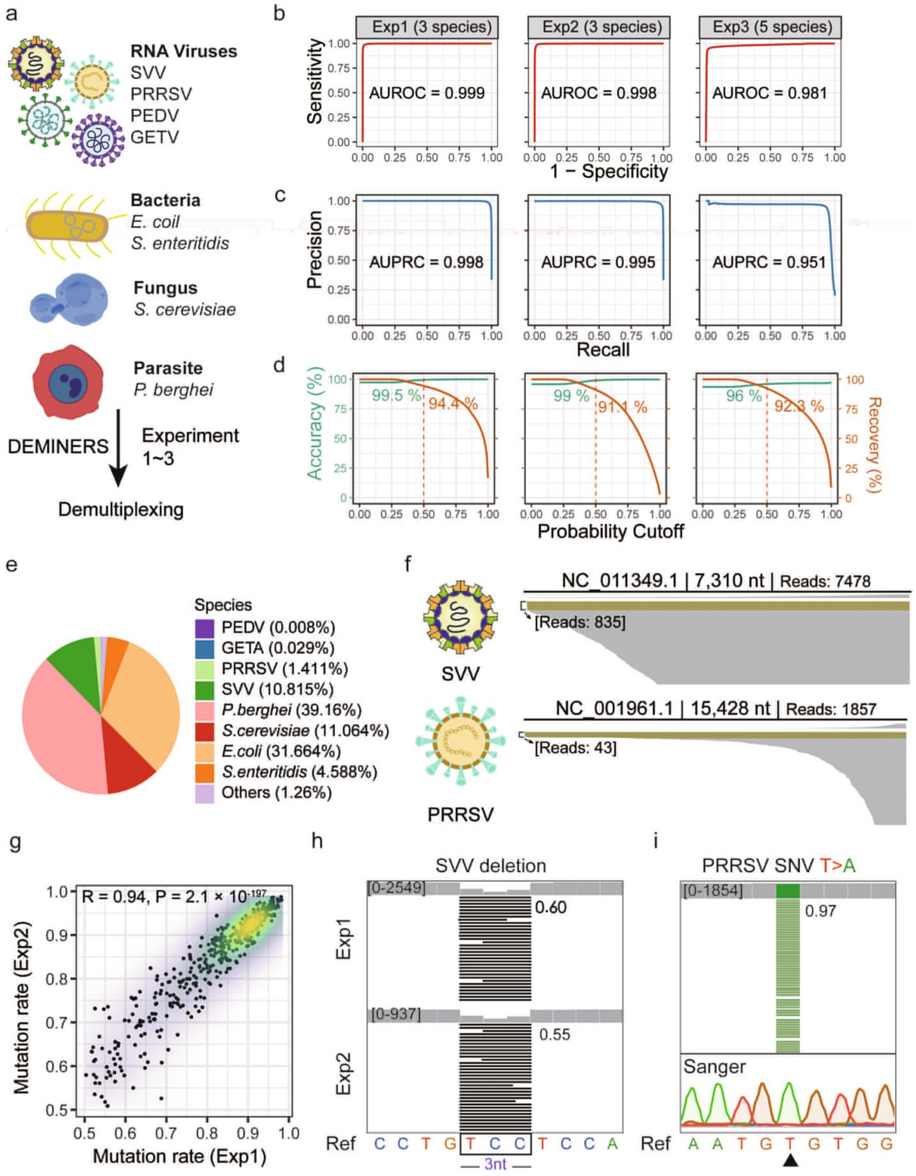

**Fig. 3** Performance of DEMINERS in pathogen identification, variant calling, and genome assembly of RNA virus from multiplexed RNA samples. **a** Experimental design to evaluate demultiplexing performance of DEMINERS. The multiplexed samples containing RNA isolated from various pathogens, including RNA viruses, bacteria, fungus, and parasite. **b**,**c** Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves of DEMINERS in demultiplexing the samples of three DRS experiments (Exp). AUROC, the area under the ROC curve; AUPRC, the area under the PR curve. **d** The accuracy and recovery rates at various predictive probability cutoffs. The red dashed lines indicate a predictive probability cutoff of 0.5. **e** Pie chart depicting the distribution of species identified in pseudo-metagenomic analysis of the combined 3 DRS datasets. **f** Integrative genomics viewer (IGV) visualizes genome coverage of Seneca Valley virus (SVV) (7,310 nt) and Porcine Reproductive and Respiratory Syndrome virus (PRRSV) (15,428 nt). Reads longer than 7000 nt for SVV and longer than 15,000 nt for PRRSV are colored in yellow and the number of reads were shown in brackets. **g** Reproducibility assessment of single-nucleotide variants (SNVs) identified from demultiplexed SVV reads in two DRS experiments (Exp1 and Exp2). Pearson correlation coefficient (R) and relative P value were shown. **h** IGV visualization of a 3-nt deletion (4022 to 4024) in SVV genome identified by DEMINERS in two experiments. The deleted sequences were boxed in the reference sequence (Ref). The numbers represent the average deletion frequencies. **i** IGV visualization of a SNV (T-to-A at position 15,307) in PRRSV genome identified by DEMINERS. The reference sequence (Ref) and Sanger sequencing chromatograms depicting the T-to-A variant (arrowhead) were shown. The number represents the frequency of the 'A' variant.

file 1: Fig. S4b), indicating the high consistency between our multiplexed DRS experiments. Next, we aim to assess the accuracy of the identified SNVs. A total of 145 SNVs in PRRSV genome with high sequencing depth were all validated by Sanger sequencing (Fig. 3i, Additional file 5: Data S3). Combined, DEMINERS can accurately demultiplex experimentally pooled RNA samples, identify pathogens, call variants with high specificity and precision, and assemble genomes of RNA viruses.

### Unveiling altered respiratory tract microbial diversity in COVID-19

To test whether DEMINERS can be applied in clinical metagenomics, we collected 11 nasopharyngeal and 13 oropharyngeal swabs from 24 individuals infected with SARS-CoV-2. The total 24 samples were multiplexed and subjected to DRS following DEMINERS workflow; meanwhile, NGS of SARS-CoV-2 genome enriched by multiplex PCR was performed (Fig. 4a, Additional file 6: Data S4). In total, 374,204 reads were retrieved after quality control, with an average length of 432-nt (ranging from 101 to 503,147-nt) (Fig. 4a, Additional file 1: Fig. S4c-d).

Taxonomic classification analysis showed that around 60% reads belonged to bacteria, and the rest belonged to human (15.72%), viruses (0.1%), or archaea (0.01%) (Fig. 4b). A total of 377 genera were classified, 322 of which present in the nasopharynx and 329 in the oropharynx. While we observed a similar level of microbial diversity (measured by three methods, Chao1 [59], Shannon [60], and Simpson [61]) (Additional file 1: Fig. S4e-f), the dominant bacterial genera were different (Fig. 4c). For instance, *Veillonella* (23,647 reads, 8.14%) was most prevalent in the nasopharynx, while *Streptococcus* was predominant in the oropharynx (8326 reads, 11.09%) ($P = 2.04 \times 10^{-143}$, Two proportion $Z$ test) (Fig. 4c,Additional file 6: Data S4), consistent with a previous research on COVID-19 microbiome [62]. Principal component analysis (PCA) showed distinct clustering of the nasopharynx and oropharynx samples based on their microbial

(See figure on next page.)

**Fig. 4** Metagenomic analysis of swab samples by DEMINERS. **a** Study design. Eleven nasopharyngeal and thirteen oropharyngeal swabs were collected from 24 individuals infected with SARS-CoV-2. The isolated RNA were multiplexed and subjected to DEMINERS followed by metagenomics analysis. Meanwhile, the same individual samples were subjected to next-generation sequencing (NGS) of SARS-CoV-2 genome enriched by multiplex PCR. **b** Taxonomic classification analysis of the DEMINERS retrieved reads. **c** Pie charts depicting the distribution of genus bacteria identified in the swab samples. **d** Bar plot showing the differential distribution of microorganisms in nasopharyngeal (Naso) and oropharyngeal (Oro) swabs. The values represent the mean percentage of each microbial species in Naso samples minus the mean percentage in Oro samples. **e** UpSet plot showing m$^6$A site overlaps among Nanom6A [13], TandemMod [64], and m6Anet [63]. The bar chart in lower left shows the total number of m$^6$A sites identified by each software, and the lower right chart illustrates the counts of intersecting or unique sites identified by each software. **f** Distinct ionic current signals indicating RNA modifications at position of 29,385 in SARS-CoV-2 genome. Red lines, SARS-CoV-2 from swab specimens (clinical); black lines, SARS-CoV-2 maintained in vitro culture (IV). **g** IGV visualization of a SARS-CoV-2 SNV (A-to-C at position 29,510, arrowhead) identified by DEMINERS and NGS in 3 swab samples. The numbers represent the read frequencies of the SNV. **h** IGV visualization of a 26-nt deletion (29,734 to 29,759) in SARS-CoV-2 genome, identified by DEMINERS and NGS in 3 samples. The average deletion frequencies were shown. **i** Density plot showing SNV densities flanking the identified m$^6$A sites (red) and random sites (grey). *P* value, Wilcoxon test. **j** Scatter plot showing Pearson correlation between Ct value of SARS-CoV-2 *N* and *Orf1ab* genes and microbial diversity (Chao1). R, Pearson correlation coefficient; N, relative sample size. **k** Box plots showing the microbial diversity (Chao1) in nasopharyngeal (Naso-) or oropharyngeal (Oro-) swabs with high-Ct value (Ct > 21) or low-Ct value (Ct ≤ 21). P values were determined by Wilcoxon test.

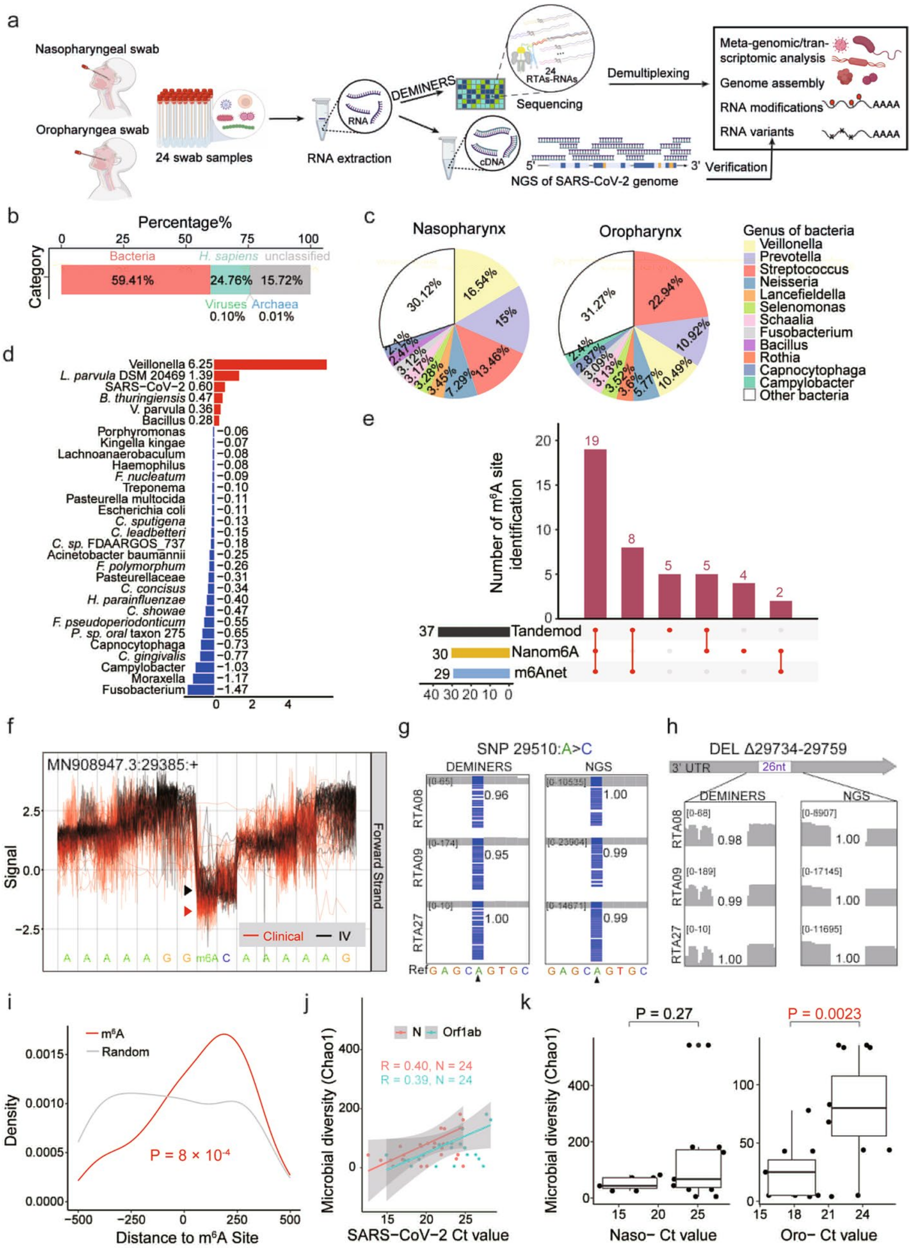

**Fig. 4** (See legend on previous page.)

compositions (Additional file 1: Fig. S4g). Among the identified species, 30 of which showed differential distribution between the two sites (Fig. 4d). We also analyzed RNA modifications of the identified species. A total of 110 m$^6$A sites were identified in *Streptococcus* using both m6Anet [63] and TandemMod [64] (Additional file 1: Fig. S4h), and these m$^6$A sites were consistent with the DRACH motifs (Additional file 1: Fig. S4i).

Despite the low amount of viral RNA recovered from these clinical samples, we successfully mapped high-quality reads to the SARS-CoV-2 genome (mean quality 46.6) and assembled high-identity SARS-CoV-2 contigs (98.42%) (Additional file 6: Data S4). The reads mainly originated from the *N* and *ORF10* genes known to generate more subgenomic RNAs, consistent with a previous DRS study using SARS-CoV-2 maintained in Vero cell culture [11] (Additional file 1: Fig. S5a-b). Using Nanom6A [13], TandemMod [64], and m6Anet [63], we identified 43 m$^6$A modification sites consistent with the DRACH motifs in the SARS-CoV-2 genome, 5 of which have been reported previously [11, 65] (Fig. 4e,f, Additional file 1: Fig. S5c-d, Additional file 6: Data S4). Furthermore, we identified 26 point-mutations and two deletions (26- and 9-nt deletion), all validated by the parallel NGS (Fig. 4g,h, Additional file 1: Fig. S5e, Additional file 6: Data S4). An enrichment of SNVs around m$^6$A peaks was found in our swab samples, consistent with the previous study on cultured SARS-CoV-2 [11] (Fig. 4i), indicating a potential relationship between mutation and m$^6$A.

Notably, we found a positive correlation between the Ct values of SARS-CoV-2 PCR test and microbial diversity ($R = 0.4$ and 0.39 for *N* and *Orf1ab* genes, respectively) (Fig. 4j, Additional file 1: Fig. S5f). Further analysis revealed that microbial diversity was significantly higher in oropharynx samples with high Ct-values ($Ct > 21$) than those with low Ct-values ($Ct \leq 21$), but the diversity is not significantly different in the nasopharynx samples (Fig. 4k, Additional file 1: Fig. S5g).

Taken together, DEMINERS enabled DRS of clinical samples with low RNA amount and the reads retrieved can be used for metagenomics, variant calling, and RNA modification analyses. Our analysis uncovered that SARS-CoV-2 abundance negatively impact on the oropharyngeal microbiota.

### Uncovering stage-specific transcriptomic features of the malaria parasite

The malaria parasite invades erythrocytes in the blood stage, classified into ring, trophozoite, and schizont stages. The mature blood-stage parasites, schizonts withdraw from the circulating blood and cytoadhere to the microvasculature of internal organs which causes severe complications (Fig. 5a). To analyze the transcriptional and post-transcriptional features in different stages of malaria parasites, we performed DRS on 3

(See figure on next page.)

**Fig. 5** Parallel comparative analysis of transcriptomic features in different stages of malaria parasites. **a** Schematics of the blood-stage malaria parasites. Trophozoite (troph) and Schizont (schz) stages of parasite were analyzed in this study using DEMINERS. **b,c** Scatter plot showing Pearson correlation between m$^6$A modification ratios and log2 gene expression or poly(A) tail length across all samples. R, Pearson correlation coefficient; *P*, *P* value; N, sample size. **d** Violin plot showing the distribution of poly(A) tail lengths of all transcripts transcribed in trophozoite or schizont stages. The number of transcripts and the mean poly(A) tail length are shown. *P* value, Wilcoxon test. **e** Venn diagram of m$^6$A sites identified in trophozoites and schizonts. **f** Heatmap illustrating the expression level of ribosome-related genes and their mean m$^6$A modification levels in trophozoites and schizonts. **g** Point plots showing the normalized gene expression of *PBANKA_1245821* and the ratios of m$^6$A (chr12_v3:1,735,101) in trophozoites and schizonts. *P* values, Wald test for gene expression, *T*-test for m$^6$A. Reads for gene, the ratio of m$^6$A (marked in red), and PlasmoDB annotation are shown in the right. **h** Bar charts showing the percentages and numbers of five major types alternative splicing events in trophozoites and schizonts. RI: retained intron, A5: alternative 5′ splice site, A3: alternative 3′ splice site, SE: skipped exon, AF: alternative first exon. **i** Sashimi plot and read alignments for *PBANKA_1316400* in trophozoites and schizonts. The red boxes indicate the novel exon, the purple boxes indicate the retained introns, and the arrows in the left indicate the reads of novel isoforms. Annotated transcript and novel isoforms with more than 2 DRS reads are shown below. The inversed triangles indicate the predicted start sites the stars indicate the predicted stop site and the numbers in the brackets showing the length of the predicted translated proteins. **j** Pymol visualization of predicted protein structures of annotated or novel isoforms. The dashed blue box indicates the missing region of novel isoforms, and the dashed purple box highlights missing region of novel isoform 2

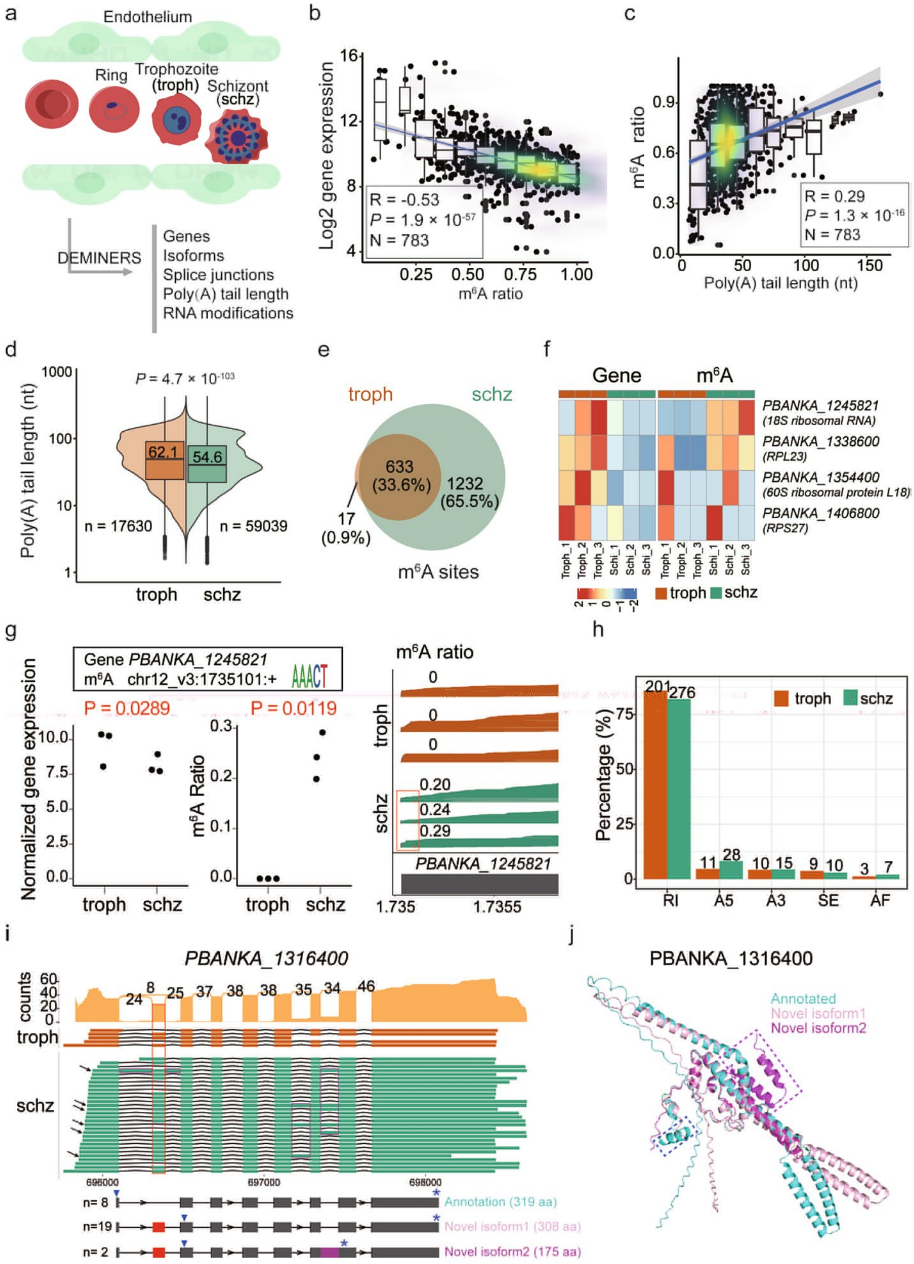

**Fig. 5** (See legend on previous page.)

trophozoite RNA and 3 schizont RNA multiplexed and sequenced in one flow cell (Additional file 7: Data S5) and analyzed the expression level of genes and isoforms, poly(A) tail length of transcripts, and m$^6$A modifications (Fig. 5a).

Interestingly, we found that the gene expression level was significantly negatively correlated with m$^6$A ratio ($R = -0.53$) (Fig. 5b), yet marginally correlated with poly(A) tail length ($R = -0.15$) (Additional file 1: Fig. S6a). In contrast, a positive correlation between m$^6$A ratio and poly(A) tail length ($R = 0.29$) was observed (Fig. 5c).

Notably, the poly(A) tail length and m$^6$A sites differed between the two stages, with an average of 8-nt longer poly(A) tail in trophozoites (Fig. 5d), and 65.5% (1232) of the identified $m^6A$ sites present only in schizont, in sharp contrast to only 17 (0.9%) trophozoite-specific sites and 663 common sites (33.6%) (Fig. 5e, Additional file 7: Data S5). We further found that the genes upregulated in the developing stages, such as ribosome-related genes, were marked with $m^6A$ at the schizont stage, possibly leading to downregulation of these genes (Fig. 5f). For example, 18S ribosomal RNA *PBANKA_1245821* harbored a schizont-specific $m^6A$ site (chr12_v3:1,735,101) and its expression level was significantly downregulated at the schizont stage (Fig. 5g), indicating a negative regulation of $m^6A$ on gene expression in the mature stage.

Next, we investigated the isoforms expressed in trophozoites and schizonts. Intron retention (IR) was the most frequent alternative splicing type, accounting for 85.9 and 82.1% of all splicing events in trophozoites and schizonts, respectively (Fig. 5h). We identified 2065 unannotated/novel isoforms in trophozoites and 2510 in schizonts (Additional file 1: Fig. S6b), 22.5% (580) of the novel isoforms were stage-specific, mostly (502) schizont-specific (Additional file 1: Fig. S6c). Among the novel splice junctions (SJs), 18.2 and 51.2% were specific in trophozoites and schizonts, respectively (Additional file 1: Fig. S6d-e). Schizont stages showed both higher numbers of predicted novel isoforms and novel SJs, indicating that the isoform diversity is higher in this stage (Additional file 1: Fig. S6b-e). For example, for *PBANKA_1316400*, predicted to encode for a U1 snRNP-associated protein [66], we identified an unannotated 80-bp exon, creating an alternative start site (Fig. 5i). The resulting new isoform is predicted to encode for a 11-amino-acid shorter protein isoform, lacking an alpha helix at the N-terminus (Fig. 5j, Additional file 7: Data S5). Moreover, 4 introns of this gene showed IR events, generating 6 different types of isoforms, mainly in schizonts. One of schizont-specific IR isoform encodes for a small protein isoform lacking both N- and C-terminus (Fig. 5i,j). For *PBANKA_0701800* (conserved protein with unknown function), we also identified a novel schizont-specific IR isoform (Additional file 1: Fig. S6f), introducing a premature stop codon possibly resulting in a non-coding RNA (Additional file 1: Fig. S6g, Additional file 7: Data S5). In addition, a novel exon-skipping isoform was identified for this gene, potentially encoding for a shorter protein isoform (Additional file 1: Fig. S6g).

Taken together, using DEMINERS, we analyzed stage-specific RNA biology of malaria parasites and identified schizont-specific features such as shorter poly(A) length, higher $m^6A$ modification levels, and increased isoform diversity.

### Identifying $m^6A$-associated isoforms in glioma

RNA modification and alternative splicing have been shown to play important roles in cancer progression and development [67–70]. To demonstrate the application of DEMINERS in analyzing RNA modifications in clinical cancer samples, we multiplexed 3 diffuse midline glioma-H3K27M mutant samples (DMG) and 1 glioblastoma sample (GBM) in one DRS run and classified the reads using DEMINERS (Fig. 6a). In addition, single-sample DRS (ss-DRS) was performed for 2 additional DMG and 3 GBM samples and NGS was also performed for verification (Additional file 8: Data S6).

First, we examined whether DEMINERS can achieve high precision in identifying mutations and deletions at human genome (GRCh38). A total of 17,587 mutations were identified, and 97% of mutations (1167) in GBM, and 98% of mutations (11,791) of DMG were verified by ss-DRS (Additional file 1: Fig. S7a). For example, two

DGM-specific mutations (chr6:29,945,653:C > T and chr6:29,945,609:C > G in *HLA-A*), one GBM-specific (chr16:2,519,933:C > A in *ATP6V0C*), and one common mutation (chr17:44,906,273:G > A in *GFAP*) identified by DEMINERS were all validated by NGS and Sanger sequencing (Fig. 6b, Additional file 1: Fig. S7b-d). Furthermore, the mutation rates were highly correlated between DEMINERS and NGS ($R = 0.93$; Additional file 1: Fig. S7e, Additional file 8: Data S6). We also identified a 5-bp deletion mutation (chr9:12,972,603–12,972,607 in *SNORD137*) in all four glioma samples (Fig. 6c) using DEMINERS.

We compared the levels of gene expression and m$^6$A modification identified by DEMINERS and ss-DRS. Although ss-DRS and DEMINERS were performed on different samples and lower read depth were obtained in the multiplexed samples, DEMINERS showed high correlation with ss-DRS in both gene expression ($R = 0.91$) and m$^6$A modification ($R = 0.96$) (Fig. 6d, Additional file 8: Data S6). In detail, 99% of the 2158 m$^6$A sites were identified both by DEMINERS and ss-DRS, 39% of which were reported in m$^6$A Altas [71]. The identified m$^6$A modifications had DRACH motifs with an enrichment in the 3'UTR region (Additional file 1: Fig. S7f), suggesting that DEMINERS can correctly identify and quantify m$^6$A level.

Next, we explored the relationship between m$^6$A modification and the expression level of genes and isoforms. Interestingly, we observed that m$^6$A ratio had a more prominent negative correlation to isoform expression level ($R = -0.61$) than the gene ($R = -0.49$) (Fig. 6e). For example, *GFAP* (glial fibrillary acidic protein), a marker of astrocyte differentiation and glioma diagnosis and prognosis [72], had two isoforms, *GFAP-214* (a protein-coding isoform) and *GFAP-230* (with an unspliced intron, likely a ncRNA) detected in the three DMG samples. *GFAP-214* had a lower expression level compared to *GFAP-230* and exhibited significantly higher m$^6$A modification at 3 sites (chr17:44,905,973,

(See figure on next page.)

**Fig. 6** RNA variants, isoforms and m$^6$A modification in human glioma. **a** Schematic representation of multiplexed human glioma samples analyzed by DEMINERS. **b** IGV visualization of a mutation (C-to-T at chr6:29,945,653 in *HLA-A* ) identified by DEMINERS and NGS in all tumor samples. The numbers represent the read frequencies of the mutations. The Sanger sequencing chromatograms and the reference sequences (Ref) were shown. **c** IGV visualization of a 5-bp deletion (12,972,603 to 12,972,607 at chr9 in *SNORD137* ) identified by DEMINERS. The reference sequence and the average deletion frequencies are shown. **d** Scatter plot showing Pearson correlation of rlog-normalized gene expression and m$^6$A modification between DEMINERS and single-sample DRS. The boxplots represent binned values of DEMINERS data, while the blue regression lines indicate the linear fit of the data. R, Pearson correlation coefficient; *P* , relative *P* value; N , sample size. **e** Scatter plot showing Pearson correlation between m$^6$A ratio and log2 gene or isoform expression. R, Pearson correlation coefficient; P , relative *P* value; N , sample size. **f** Proportions of two isoforms ( *GFAP-214* and *GFAP-230* ) and their m$^6$A ratios with motif at chr17:44,905,973, chr17:44,906,138, and chr17:44,906,690 in three DMG samples (left). P values, T -test. Total reads for *GFAP-214* and *GFAP-230* isoforms and m$^6$A sites in DMG samples are shown (right). The colored dots represent the m$^6$A sites and the numbers indicate ratios. The transcript structure based on Ensembl annotation is shown at the bottom, indicating the locations of the associated six splice junctions (SJs). The 6 SJs are located on chr17 with positions of 44,914,089 — 44,915,025 (SJ1), 44,913,824 — 44,914,027 (SJ2), 44,913,431 — 44,913,727 (SJ3), 44,911,798 — 44,913,268 (SJ4), 44,911,457 — 44,911,671 (SJ5), and 44,910,659 — 44,911,235 (SJ6). **g** Box plot showing the normalized expression of SJ6 in METTL3-KO ( $n = 2$) and control ($n = 2$) K562 cells (ENCODE CRISPR RNA-seq data [75]). P value, Wilcoxon test. h Box plots showing normalized expression of 6 SJs of *GFAP-214* in high ($n = 133$) and low ($n = 33$) expression groups in the TCGA-GBM cohort [76]. P values, Wilcoxon test. i Kaplan–Meier survival curves showing overall survival in the TCGA-GBM cohort [76]. The survival curves were grouped by high ($n = 139$) and low ($n = 25$) expression of the *GFAP* gene (upper panel), or grouped by high ( $n = 133$) and low ($n = 31$) expression of 6 SJs of *GFAP-214* (lower panel). P values, log-rank test

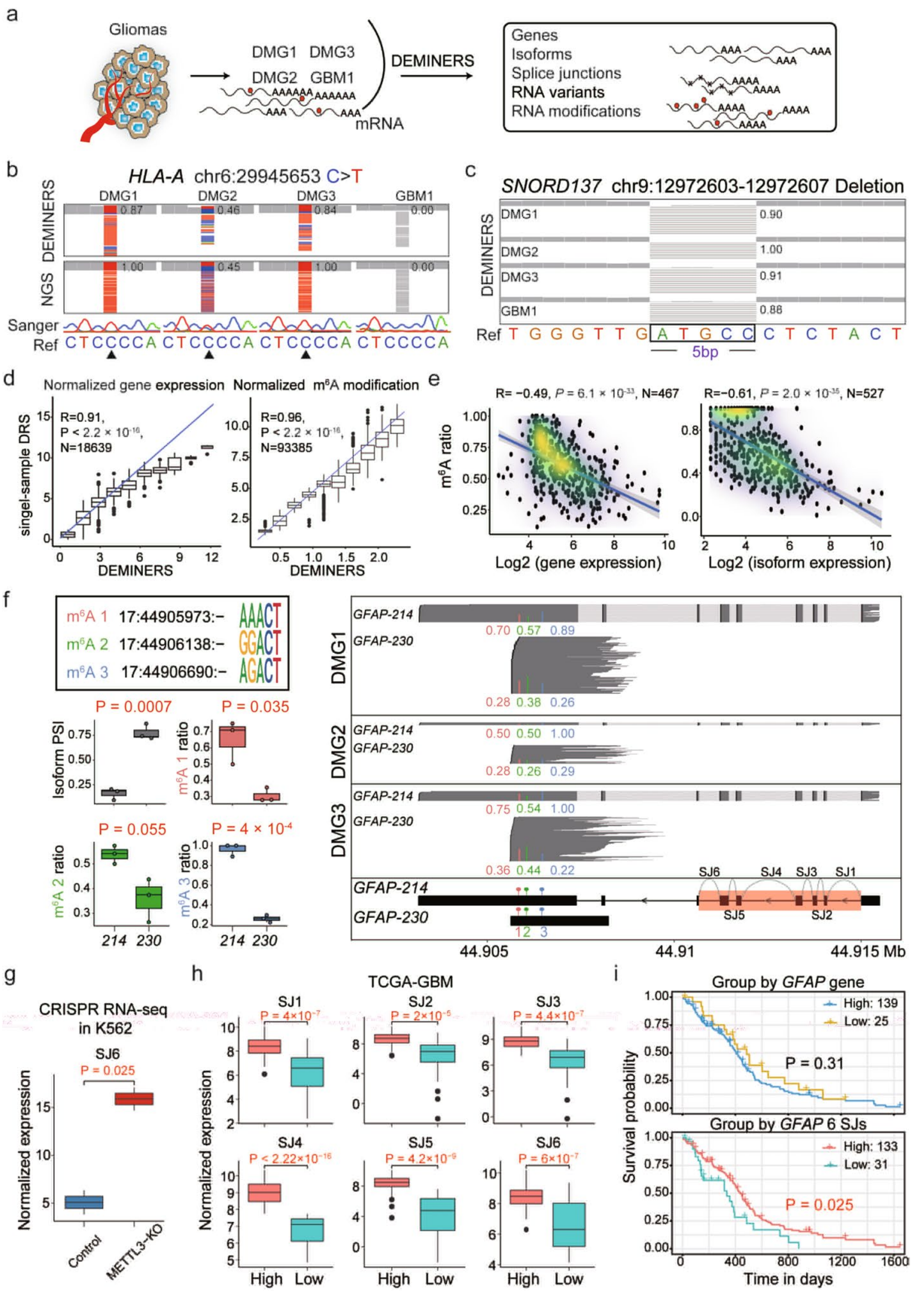

**Fig. 6** (See legend on previous page.)

chr17:44,906,138, and chr17:44,906,690) (Fig. 6f). *METTL3*-mediated m$^6$A modification have been reported to regulate cancer progression [73, 74]. We thus assessed whether the *GFAP-214* is regulated by the m$^6$A writer, *METTL3*. A K562 *METTL3* knock-out (KO) dataset from ENCODE [75] was used. We found that the splice junction 6 of *GFAP-214* was significantly upregulated in *METTL3*-KO cells (Fig. 6g), which is consistent with the hypothesis that the loss of *METTL3* leads to downregulation of m$^6$A and may contribute to the upregulated expression of *GFAP-214*. To investigate the importance of the isoforms of *GFAP*, we analyzed the RNA-seq data of GBM in TCGA [76]. We found high expression of 6 SJs of *GFAP-214* associated with better prognosis ($P=0.025$), whereas high- or low-expression levels of *GFAP* showed no significant difference in overall survival ($P = 0.31$) (Fig. 6h,i). Our findings showed that m$^6$A potentially play a role in regulating the expression of *GFAP* isoforms, potentially contributing to cancer progression through complex post-transcriptional regulatory networks.

## Discussion

DRS enables direct sequencing of RNA molecules without fragmentation, reverse transcription or amplification, making it effective in obtaining information of full-length transcripts, poly(A) tail length, and RNA modifications. Despite its advantages, DRS faces challenges such as low throughput, low basecalling accuracy yet high input requirements. DEMINERS significantly improved the accuracies of barcode classification and basecalling, thereby increasing the throughput of DRS and facilitating the complex analysis of RNA biology. Importantly, DEMINERS allows parallel comparisons of transcriptomic features in different biological conditions, increasing statistical power and reducing batch effects and sequencing cost.

Hitherto, the barcoding strategy offered by ONT is only suitable for DNA and cDNA sequencing, but not for DRS, and DeePlexiCon [52] and Poreplex [51] provide demultiplexing of only 4 barcodes. In this study, we scaled up the demultiplexing capability to 24 samples while maintaining higher accuracy. We considered the GC content, melting temperature, and secondary structure in the barcode design [77]. To ensure the accuracy of barcode classification, we opted for barcodes with different length and implemented a Hamming distance error correction algorithm [78].

Next, we developed a novel signal smoothing method for the random forest algorithm, which preserved the major changes in the current signals while it reduced the random noises by segmenting the current signals and speeding up the classification step. As a result, DEMINERS exhibits superior running speed, taking only 1/12 of the CPU time of DeePlexiCon [52]. Meanwhile, DEMINERS shows high flexibility, allowing preference for high classification accuracy or high read recovery by adjusting the classifier prediction probability.

The differences in current signals and translocation speeds between DNA and RNA lead to a high error rate in basecalling. We developed a DRS-specific basecalling method reconstructed from DenseNet [55], leveraging its advantages of facilitating information flow, feature reuse, and efficient gradient propagation. Meanwhile, we employed the memory optimization technique [58], which significantly reduces memory consumption during model training. Moreover, we provide species-specific basecalling modes, further improving the accuracy of basecalling.

With improved barcode classification and basecallling, we demonstrated the application of DEMINERS in real-world samples and showcased its capability of metagenomics, genome assembly of RNA viruses, and analyses of transcripts, mutation, and RNA modification. The real-time ONT sequencing is particularly useful in rapid detection of infection, antibiotic/antimicrobial resistance and metagenomics[1, 79, 80]. Although compared to cDNA/DNA sequencing DRS provides additional information, such as RNA modifications. The technical difficulties remain for application of DRS in infection detection because of the low amount of viral RNAs in swab samples. It is estimated that there are only an average of 58 SARS-CoV-2 RNA copies/μL in nasopharyngeal swabs and 14 copies/μL in oropharyngeal swabs [81]. Thanks to the increased accuracy in barcode classification and basecalling, DEMINERS can demultiplex up to 24 samples and the required RNA input is thus substantially reduced. We demonstrated the application of DEMINERS in metagenomic analysis of 24 multiplexed nasal/oropharyngeal swabs and revealed that the microbiota was altered in COVID-19, in line with previous studies [62, 82, 83].

We demonstrated the application of DEMINERS in parallel comparison studies of different biological conditions without library construction nor sequencing batch effect. We not only identified many unannotated exons and spicing junctions, but also revealed stage-specific characteristics, including shorter poly(A) length, and higher transcript diversity and $m^6A$ modification in the mature stage of malaria parasites.

RNA modifications, particularly $m^6A$ methylation, have emerged as pivotal regulatory elements in microorganisms [84], including viruses [85], and cancers [86, 87]. Using DEMINERS, we not only accurately identified $m^6A$ modification sites and identified a negative correlation between $m^6A$ and gene expression in malaria parasites and gliomas alike to previous studies [88], but also uncovered an even more prominent negative correlation between $m^6A$ and isoforms in gliomas. For example, $m^6A$ likely involved in splicing regulation of *GFAP* and is associated with glioma prognosis.

Together, DEMINERS brings together high accuracy barcode classifier based on machine-learning and improved baseballer based on a convolutional neural network and provides users with an economical solution for DRS, especially more suitable for SQK-RNA004 version with increased throughput to greater than 30 M reads, reducing both the RNA input and the sequencing cost and enables exploration of RNA biology in diverse biological processes.

## Conclusion

DEMINERS enhances RNA sequencing by integrating a multiplexed workflow, Random Forest-based classifier, and convolutional neural network basecaller. It outperforms existing methods in demultiplexing and basecalling, enabling accurate mutation calling, RNA virus genome assembly, and RNA modification detection. DEMINERS also offers valuable insights into transcriptomics of COVID-19, malaria, and glioma, demonstrating its potential in advancing clinical metagenomics and transcriptomics research.

## Methods

### Generation of multiplexed direct RNA sequencing data

#### *RNA transcription adapter design*

A set of 40 RNA transcription adapters (RTAs) were designed, each comprising an RNA adaptor (RMX), a barcode, and a poly(T) sequence. In addition, the standard ONT RTA, 3 RTAs for DeePlexiCon [52], and 4 RTAs for Poreplex [51] were also included, making up a total of 48 RTAs that were used in our study (Additional file 3: Data S1).

The barcode length ranged from 20 to 28 nucleotides (nt) increasing by 2-nt. We did not use ultra-long barcodes to prevent potential interference during the magnetic bead-based purification step removing free RTA. The OligoAnalyzer™ Tool was used to evaluate the GC content, melting temperature, and secondary structure of DNA oligonucleotides to assess the quality of the DNA barcodes. The barcode sequences are listed in Additional file 3: Data S1.

#### *Preparation of RTAs and in vitro transcribed RNAs*

Forward (1.54 μM) and reverse (1.4 μM) strands of the RTA sequences were mixed in the buffer (10 mM Tris–HCl pH 7.5, 50 mM KCl). The extra forward strands were used to avoid left-over reverse strands, which may reduce the yield of RNA-RTA products. The custom RTAs were annealed following the process: 95 °C for 5 min, 65 °C, 50 °C, 37 °C, and 22 °C for 30 min at each temperature, then stored at 4 °C. The formation of RT adapters was verified by 4% agarose gel electrophoresis.

Fifty-one in vitro transcribed (IVT) RNAs, with lengths ranging from 177 to 748 bp, were generated for barcode testing and model training. The genes and primers used to generate cDNA are listed in Additional file 3: Data S1. The PCR products were determined by agarose gel electrophoresis. IVT RNAs were then generated by T7-RNA polymerase (NEB, M0251) using 1 μg of purified PCR products and purified using RNA Clean & Concentrator-5 kits (Zymo, R1015) following the manufacturer's recommendations. The integrity and concentration of IVT RNAs were measured by Qsep100 (BiOptic Inc) and Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific), respectively.

#### *RNA multiplexing, library preparation and sequencing*

RNA samples were ligated to pre-annealed custom RT adaptors (RTA) using NEBNext Quick Ligation Module (NEB, E6056) in separated reactions and reverse-transcribed by SuperScript IV Reverse Transcriptase (Thermo Fisher, 18090200) to form RNA/DNA duplexes avoiding the formation of the secondary structure of RNA. The products were purified using 1.8X Agencourt RNAClean XP beads (Beckman, A63987). Next, equimolar amounts of reverse-transcribed RNA/DNA hybrids from each reaction were pooled, and the RNA adaptors mix (RMX) were ligated to the 3′ end of RNA/DNA hybrid at room temperature for 30 min. After ligation, the mixture was purified using 0.4X Agencourt RNAClean XP beads (Beckman, A63987), washed twice with the wash buffer, and eluted in the elution buffer. The RNA sequencing library was prepared using Direct RNA Sequencing Kit (SQK-RNA002, Oxford Nanopore Technologies) following the Direct RNA sequencing protocol (DRCE_9079_v2_revI_14Aug2019, ONT). The sequencing libraries were eluted in the RNA running buffer, loaded onto a primed R9.4.5 flowcell (ONT, FLO-MIN112 or FLO-PRO002), and sequenced on a MinION or PromethION sequencer (ONT) until all pores were inactivated. Five sequencing runs were performed with various combinations of RTA and IVTs and the details of read counts and quality are listed in Additional file 3: Data S1. For subsequent experiments, eleven sequencing runs were conducted in total, with detailed descriptions of the data, barcodes used, and accession IDs provided in Additional file 2: Table S5.

### Model training for demultiplexing DRS data

#### *Basecalling and alignment*

The raw current signal data from MinKNOW (Oxford Nanopore Technologies, v21.06.0) were used to obtain the RNA sequence data using Guppy (ONT, v6.5.7) with high-accuracy basecalling model (config: rna_r9.4.1_70bps_hac.cfg). Basecalled reads (FASTQ) were aligned to the reference sequences using minimap2 [89] (v2.15, -ax map-ont) with default parameters. The reference FASTA sequences were extracted from the reference genomes and listed in Additional file 3: Data S1. We filtered out the multiple alignments and the reads with mapping quality less than 60 for downstream analysis. The remaining high-quality and uniquely mapped reads were grouped based on the coupled RNA sequence, distinguished by mapped reference sequences.

### *Signal transformation*

The raw current signals were extracted from the FAST5 files using R package rhdf5 (v2.30.1, https://github.com/grimbough/rhdf5). First, we implemented the standard double-exponential moving average (DEMA) algorithm from the R package smoother (v1.1, https://CRAN.R-project.org/package=smoother) to reduce the noise of raw current signals. In this step, we set the averaging period to 20 units, meaning that the average was calculated over 20 consecutive data points. Secondly, we extracted the signals of the adapter and barcode of each read from the denoised electrical signals, according to the characteristic higher and more stable current change generated by poly(A) tails and the lower current signals of DNA molecules. Specifically, we used the *cpt.meanvar* function of the R package changepoint [90] (v2.2.2) to calculate the optimal change position from the first 20,000 denoised electrical signals which contains the adapter, barcode, and ploy(A) signals for almost all reads. A single changepoint is denoted as the first observation of the new segment with a significant difference in the mean and variance using the *cpt.meanvar* function with AMOC method [91]. Signals before the changepoint were extracted as adapters and barcodes, and at least 90% of adapter and barcode signals were smaller than the mean of poly(A) signals originating from the first stable segment after the changepoint. Finally, for the extracted adapter and barcode signals, the approximate (BinSeg) method [92] of *cpt.meanvar* function was used to identify multiple changepoints with penalty. For each read, we identified 99 changepoints to divide the current signals into 100 units/segments and calculated the average current value of each unit as feature values.

### *Model training for demultiplexing*

The feature values, defined from the mean current of 100 units, were used as predictors of each read. A matrix of 100 features generated from all reads was eventually used for model training, testing, and independent validation. To assess the accuracy of our model, we established the ground truth in which the abovementioned IVT RNAs were linked to the corresponding barcodes, that is, we have the correct correspondences between IVT RNA and barcode. This ground truth information was then used to measure the accuracy of the training and testing sets. To compare and identify optimal machine learning classification algorithms, we selected 4 barcodes from sequencing Run 4 as a test (Additional file 3: Data S1). For each barcode, we randomly selected 10,000 reads as a training set and the remaining reads were used as a test set.

Model training was performed using the R package caret [93] (v6.0–88) to evaluate six machine-learning models for barcode classification, including random forest (RF), naive

Bayes (NB), K nearest neighbours (KNNs), bagged classification trees (CART), adaptive boosting classification trees (AdaBoost), and neural networks (NNet). Hyperparameter optimization for each model was achieved using a grid search strategy. To define the granularity of the parameter grid, we used the *tuneLength* parameter within the *train* function, which allows caret [93] (v6.0–88) to automatically generate an optimal number of hyperparameter values for evaluation. By setting *tuneLength* parameter to a specific integer (e.g., 30), a comprehensive grid of configurations was created and assessed. The best set of hyperparameters was then identified based on performance metrics such as accuracy and the Kappa statistic. To ensure robust performance evaluation and reduce the risk of overfitting, we employed tenfold cross-validation (repeated 10 times), specified using the *trainControl* function.

As RF algorithm outperformed other methods, it was selected for subsequent method development. We selected the 2, 4, 6, 8, 10, and 24 barcodes with the most reads from sequencing Run4 for RF training (Additional file 4: Data S2). Because the read numbers of each barcode differed, we randomly sampled 30,000 reads of each barcode from Run4 as a training set using the downSample function of R package caret [93] (v6.0–88). The rest of the reads of the 24 barcodes were used as a test set and the reads from the Run3 were used as an independent validation set (Additional file 4: Data S2). To increase the computational efficiency, a parallel processing framework was employed in the model training using the R package doParallel (v1.0.16, https://CRAN.R-project.org/package=doParalle).

### *Method comparison*

In the comparative study of DEMINERS and DeePlexiCon [52] (v1.2.0), we randomly selected 24,000 reads (10,000 each from 24 barcodes with the most reads from Run4) to train the prediction models using DEMINERS and DeePlexiCon. Similarly, the rest of the reads of the 24 barcodes were used as a test set and the reads from Run3 were used as an independent validation set. In the comparison of DEMINERS and Poreplex [51] (v0.5), because we did not have access to the raw data of Poreplex, and Poreplex (v0.5) does not support user-supplied data to train new models, we evaluated their pre-trained models on our data (Additional file 4: Data S2). We selected all reads of the 4 barcodes used by Poreplex (v0.5) from all sequencing runs, and randomly extracted 100,000 reads of each barcode as a training set. The remaining reads were used as a test set to evaluate and compare the performance of DEMINERS and Poreplex. For Poreplex, the classification of barcodes were performed using the recommended parameters [51].

### *Performance evaluation*

The receiving operator characteristic curve (ROC) and the associated precision-recall curves were charted using the R package multiROC (v1.1.1, https://CRAN.R-project.org/package=multiROC), and the area under the curves (AUC) were calculated as AUROC and AUPRC, respectively. The relative accuracy, sensitivity, specificity, precision, recall, and F1 score of each model were determined using the *confusionMatrix* function of R package caret [93] (v6.0–88). The CPU time was calculated as the sum of the user time and system time evaluated using GNU time (https://www.gnu.org/software/time/).

## Basecalling model training

### *Architecture of the basecaller*

In our architecture, three convolutional layers were used to denoise the signals, followed by a max pooling layer with a pooling size of 10 for downsampling. DEMINERS includes four dense blocks with 6, 12, 24, and 16 layers, respectively. Each dense layer starts with DenseNet typically started with a bottleneck layer that includes Batch Normalization (BN), a Sigmoid Linear Unit (SiLU) [94], and then a $1 \times 1$ convolution to reduce dimensionality, followed by another BN, SiLU, and a $3 \times 3$ convolution. Layers are connected feed-forwardly, receiving concatenated feature maps from all previous layers to enhance feature reuse and reduce parameters.

In contrast to DenseNet having $3 \times 3$ convolutional layers, we increased the number of channels and the kernel sizes used in each dense layer to capture a broader context of the base position, up to 1024 channels and a kernel size of 99 in the last dense block. Each dense block, except the last one, was followed by a $1 \times 1$ convolutional layer to reduce features and accelerate processing. The final output was connected to a fully connected layer with a log softmax activation for classification. Our network used connectionist temporal classification (CTC) loss [57] for gradient descent. Basecalling was performed with a beam search size of 5.

### *Data process and basecaller training*

For model training and evaluation, we utilized the Taiyaki (ONT, v5.3.0) to preprocess the training and testing datasets. Each chunk, representing a segment of the raw signal data, contained 4096 signal values. These chunks were normalized using the median absolute deviation method to ensure data consistency. The training set includes data from *Arabidopsis thaliana*, *Homo Sapiens*, *Caenorhabditis elegans*, and *Escherichia coli* from the RODAN [54] study, using one million signal chunks for training and 100,000 for validation. In addition, we employed a memory optimization technique [58] to reduce memory consumption during model training.

The test dataset includes datasets from the RODAN study, including *Homo sapiens*, *Arabidopsis thaliana*, *Mus musculus*, *S. cerevisiae* S288C, and *Populus trichocarpa*, the published SARS-CoV-2 dataset [11] and the datasets generated in this study (Additional file 5: Data S3, Additional file 8: Data S6), including *Homo Sapiens*, *P. berghei*, Seneca Valley virus (SVV), Porcine Epidemic Diarrhea virus (PEDV), and Porcine Reproductive and Respiratory Syndrome virus (PRRSV), a total of 10 species (Additional file 2: Table S4). We used PyTorch [95] (v2.0.1), set the batch size to 32 and trained for 30 epochs. The label smoothing technique of the model and basecalling function were adapted from RODAN [54].

### *Species-specific model training*

The *Mus musculus* dataset from RODAN [54] was used to train a species-specific model. The training set consists of 20,000 electrical signals, while the validation and testing sets each contained 4000 electrical signals. The HDF5 data were generated using Taiyaki (ONT, v5.3.0), and species-specific models were trained similarly to the general models, except that the data sources differed. The parameters for training were set to a learning rate of 0.002, a batch size of 32, and the process was conducted over 30 epochs. The nucleotide sequences of the test set were ultimately produced by the DEMINERS's basecaller utilizing the trained model to ensure accurate representation of the mouse transcripts.

#### Performance evaluation

Basecallers were evaluated using sequence identity defined as $Accuracy = M / (M + X + I + D)$, where $M$ is the number of matching bases, $X$ is the number of mismatches, $I$ is the number of insertions, and $D$ is the number of deletions. Sequence alignment and accuracy assessment, including quantification of mismatches, insertions, and deletions against a reference genome, were conducted using minimap2 [89] (v2.15, -cs) and the *accuracy.py* function of RODAN [54] (v1.0).

### RNA extraction from biological samples

#### S. cerevisiae

*S. cerevisiae* W303 was cultured at 30 °C in YPD (Yeast Peptone Dextrose) medium containing 20 g/L of glucose, 10 g/L of yeast extraction, and 20 g/L of peptone. The collected cells were centrifuged at 5000 rpm for 5 min, followed by washing with PBS. The resuspended cells in PBS were frozen with liquid nitrogen and then thawed in 37 °C water four times. Total RNA was extracted with the TRIzol reagent (Invitrogen, 15596026) and the mRNA was enriched according to the Dynabeads mRNA purification Kit (Thermo Fisher, 61006).

#### Viruses

Seneca Valley virus (SVV), Porcine Epidemic Diarrhea virus (PEDV), Getah virus (GETV), and Porcine Reproductive and Respiratory Syndrome virus (PRRSV) were provided by the Key Laboratory of Animal Diseases and Human Health of Sichuan Province. The total RNA of these viruses was extracted using TRIzol reagent (Invitrogen, 15596026). Isolation of SARS-CoV-2 RNA was described in the section of "Clinical specimen collection and RNA preparation."

#### Bacteria

*Escherichia coli O157:H7* (ATCC 43895) and *Salmonella enteritid* (ATCC 13076) were cultured in LB media at 37 °C to an $OD_{600}$ of 0.6. Bacteria were collected by centrifugation at 3000 rpm for 5 min and washed with PBS. The resuspended cells in PBS were frozen with liquid nitrogen and then thawed in 37 °C water for four times. Total RNA was isolated using TRIzol reagent (Invitrogen, 15596026) following the instructions of the manufacturer, and treated by DNase I (Thermo Fisher, EN0521) at 37 °C for 30 min. Poly(A) tailing was performed using *E. coli* Poly(A) Polymerase (NEB, M0276S), and the resulting product was purified using RNA Clean & Concentrator-5 kit (Zymo, R1015) and enriched by the Dynabeads mRNA purification Kit (Thermo Fisher, 61006).

### *Plasmodium berghei*

C57BL/6 J mice aged between 6 and 8 weeks were purchased from GemPharmatech (Jiangsu, China) and housed under SPF conditions (Specific Pathogen Free) at the Laboratory Animal Center of West China Second University Hospital. The animal experiments were carried out following the protocols approved by the Institutional Animal Care and Use Committee of West China Second University Hospital [(2018) Animal Ethics Approval No. 024].

Mice were intraperitoneally injected with $10^4$ red blood cells (RBCs) infected with *P. berghei* ANKA parasites. Six days after infection, the blood containing trophozoite stage parasites was collected via cardiac puncture and filtered through leukocyte filters (Bengbu Zhixing Biotech, China). The collected infected RBCs were cultured in IMDM (Thermo Fisher, 12440053) supplemented with 20% fetal bovine serum (Thermo Fisher, 10100147) with 0.5% Penicillin–Streptomycin (Thermo Fisher, 15140122) at 37 °C in an atmosphere of 5% $CO_2$ and 5% $O_2$ for 14–16 h. The maturation of the parasites was examined via Giemsa-stained blood smears. Schizont or late trophozoite stage parasites were separated using nycodenz (Alere Technologies, 1002424) density centrifugation [96]. The upper layer consisted of schizont-infected RBCs and the layer below nycodenz contained a mixture of trophozoite-infected RBCs and uninfected RBCs. The RBCs were lysed with Red Blood Cell Lysis Buffer (Solarbio, R1010), and the parasites were enriched by centrifugation and washed with PBS. Total RNAs of schizont and late trophozoites were extracted using TRIzol reagent (Invitrogen, 15596026) following the manufacturers' recommendations.

### Clinical specimen collection and RNA preparation

Fresh tumor specimens were collected into OCT from glioma patients undergoing surgical resection at West China Hospital between Sep 2020 and Mar 2021. The total RNA was extracted using the TRIzol reagent (Invitrogen, 15596026). Enrichment of poly(A) + RNA was performed using Dynabeads™ mRNA Purification Kit (Invitrogen, 61006).

The nasopharyngeal and oropharyngeal swabs were collected from suspected or confirmed SARS-CoV-2 infected individuals, the RNA was extracted from swabs utilizing the RNA extraction kit (Sichuan Maccura Biotechnology, GN7101913). Real-time RT-PCR was performed by amplifying SARS-CoV-2 two target genes, open reading frame 1ab (*ORF1ab*) and nucleocapsid protein (*N*), using the 2019-nCoV Nucleic Acid Detection Kit (Sansure Biotech Inc.)

### Nanopore direct RNA sequencing and library preparation

Single-sample direct RNA sequencing (DRS) libraries were prepared using the Direct RNA Sequencing Kit (SQK-RNA002, ONT) following the manufacturer's protocol (DRCE_9079_v2_revI_14Aug2019, ONT). For multiplexed RNA samples, the DRS libraries were prepared using the DEMINERS approach, which enables the concurrent sequencing of multiple samples within the same run. Sequencing was performed on a PromethION sequencer (ONT) using the R9.4.1 flow cell. Electrical signal data were collected using the MinKNOW software (v21.06.0, ONT).

### Read basecalling, alignment, and visualization

For basecalling of the raw electrical signals contained in FAST5 files, we utilized the DEMINERS basecaller, applying the rna_r9.4.1_hac@v1.0 model, which enabled us to acquire RNA sequence data. The reference genome FASTA files were downloaded from NCBI (https://www.ncbi.nlm.nih.gov). The *P. berghei* genome FASTA and GFF files were downloaded from the PlasmoDB (https://plasmodb.org/plasmo/app). The filtered sequence reads were aligned against the reference genomes (GenBank Accession): *Homo sapiens* (GRCh38), PEDV (NC_003436), SVV (DQ641257), PRRSV (NC_001961), SARS-CoV-2 (MN908947.3), *E. coil* (NC_000913.3), *S. enterica* (NC_003197.2), *S. cerevisiae* (NC_001133-48), *Prevotella* (NZ_CP019300.1), *Veillonella* (NZ_CABKSO010000001.1), *Streptococcus* (NZ_GL732439.1), and *P. berghei* (PlasmoDB-53, Plasmodb.org), using minimap2 [89] (v2.15-ax map-ont). The direct RNA sequencing reads distribution was assessed by the coverage function of R package GenomicAlignments [97] (v1.22.1), and visualized by R package ggplot2 (v3.3.6, https://ggplot2.tidyverse.org/). The mutations, deletions, and genome coverage were visualized using IGV [98] (v2.12.2).

### Next-generation sequencing for SARS-CoV-2 samples and data process

The RNA extracted from nasopharyngeal and oropharyngeal swabs was subjected to a PCR enrichment of the SARS-CoV-2 genome using the SARS-CoV-2 Full Length Genome Panel (Genskey, 2205) following the manufacturer's instructions. The cDNA was then used to prepare paired-end sequencing libraries through a ligation method (Genskey, 2205). According to the manufacturer's recommendations, the resulting libraries were sequenced on an MGISEQ-2000 platform using the sequencing reaction kit (Genskey, GS-2000-FCS-SE100).

The low-quality regions, adaptor sequences, and sequencing primers were trimmed using fastp [99] (v0.23.2). The clean reads were mapped to the reference genome (MN908947.3) by Bowtie2 [100] (v2.4.4). After alignment, all mapped reads were aggregated to construct a consensus sequence employing methods from prior research [101].

### Identification of SNVs and deletions

For DEMINERS, we selected alignment reads from SVV, PRRSV, SARS-CoV-2, and glioma samples. For NGS, we selected alignment reads from SARS-CoV-2, and for glioma samples, we used previously published cDNA-NGS data [102] (Genome Sequence Archive, accession ID HRA001865). These reads were aligned to the reference genome (GRch38) using Bowtie2 [100] (v2.4.4). Subsequently, we generated VCF files containing mutation information using the mpileup function in Samtools [103] (v1.9) and the call function in bcftools [104] (v1.8). Variants with a quality score below 20 were excluded. Deletions were identified based on a minimum expression level of 3, deletion length greater than 3 bp, and a deletion coverage greater than 30%.

### Isoform reconstruction and quantification

#### *Read alignment and correction*

The reads were aligned to the reference genome using minimap2 [89] (v2.15, -ax splice -k14 -uf –secondary = no). Annotated splice junctions were provided to guide the alignment. The alignment output was subsequently sorted and converted into the BAM format using Samtools [103] (v1.9). Then, TranscriptClean [105] (v2.0.2) was used to correct mismatches, insertions, deletions, and non-canonical splice junctions, using known variants and splice junctions as references. The cleaned reads were next realigned to the reference genome for subsequent analysis.

#### *Isoform identification and quantification*

The process of isoform identification was performed using a custom-developed R script (https://github.com/LuChenLab/DEMINERS/tree/main/scripts). This process began with reading BAM files and filtering reads based on the mapping quality with a minimum MAPQ score using Rsamtools (v2.6.0, https://bioconductor.org/packages/Rsamtools). Subsequently, it identifies novel splice junctions supported by user-defined read count thresholds using GenomicAlignments [97] (v1.22.1). Strand-specific analysis is also incorporated to delineate transcription directionality. Several parameters were configured, such as the minimum length of insertions or deletions and the minimum reads for various novel transcription features. Moreover, multi-core processing was enabled by using doParallel (v1.0.16, https://CRAN.R-project.org/package=doParalle). This workflow generated a FASTA file with reconstructed isoform sequences, a JSON file mapping reads to corresponding transcripts, and an annotated GTF file cataloging the novel isoforms.

Quantification was achieved via an Expectation–Maximization (EM) algorithm after isoform identification. Initially, the EM algorithm assigned a compatibility score to each isoform based on the proportion of isoform per read. During the EM cycles, isoform abundances were calculated by aggregating compatibility scores for each read aligned to the isoform and normalizing by the total compatibility scores for all alignments. The compatibility index for each read-isoform pair was updated by dividing the isoform abundance by the sum of abundances for reads aligned to that isoform. The EM iterations continue until the predefined convergence criterion is met, indicating minimal changes in cumulative isoform abundance across successive EM rounds. Upon convergence, the algorithm yields an accurate isoform abundance table from which estimated counts were derived, providing a precise measurement of isoform presence across samples based on the demultiplexing method.

### Splice events identification

The splice junction (SJ) identification and quantification were performed using the junctions function from R package GenomicAlignments [97] (v1.22.1) The alternative splicing events were extracted from aligned bam files using SUPPA2 [106] (v2.3).

### Prediction and analysis of protein structures from isoform sequences

Possible open reading frames (ORFs) were predicted based on the isoform nucleotide sequences using R package ORFhunteR [107] (v1.4.0). The longest ORF was translated into an amino acid sequence and subjected to structural prediction using AlphaFold [108] (v3). Functional predictions were performed with InterProScan5 [109] (v5.68–100.0). Finally, protein structure alignment and visualization were conducted using PyMOL (v3.0, https://www.pymol.org/).

### Assessing prognostic value of genes and splice junctions in gliomas

To verify whether *GFAP* splicing events are regulated by *METTL3*, we downloaded the RNA-seq on K562 cells treated with a CRISPR gRNA against *METTL3* from the ENCODE database [75] and identified the expression of *GFAP*-associated SJs. To explore the prognostic value of genes and splice junctions in gliomas, we downloaded the gene expression and clinical data of glioblastoma multiforme (GBM) from the TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/), and linear junctions expression data from the RJunBase database [76]. Using the R packages survminer (v0.4.9, https://CRAN.R-project.org/package=survminer) and survival (v3.2.13, https://CRAN.R-project.org/package=survival), we determined the optimal cutpoints for continuous variables, plotted survival curves, and computed *p*-values to compare survival curves.

### m$^6$A identification and quantification

Raw current signal data was processed by re-squiggled algorithm with Tombo (Oxford Nanopore Technologies, v1.5.1) for accurate reference alignment. Nanom6A [13] (v2021.10.22) extracted statistical metrics from Nanopore signals for m$^6$A modification probability post-re-squiggle. At the same time, read IDs can be used to associate modification sites from Nanom6A [13] (v2021.10.22) with isoform. In parallel, TandemMod [64] (v2023.08.18) performed feature extraction and m$^6$A modified ratio calculations based on designated motifs. Simultaneously, m6Anet [63] (v2.1.0), with signal segmentation by Nanopolish [4] (v0.13.2), predicted m$^6$A modifications, employing event scaling and signal indexing for feature extraction.

### Association between SNV and m$^6$A Site in SARS-CoV-2 genomes

Firstly, we identified m$^6$A sites in two samples with over 50 SARS-CoV-2 reads using three software tools: Nanom6A [13] (v2022.12.22), TandemMod [64] (v2023.08.18) and m6Anet [63] (v2.1.0). A site was deemed highly credible if identified more than five times by all three tools in at least two samples (Additional file 7: Data S5). From these m$^6$A sites, we extended a 500-bp window both upstream and downstream. SNV sites, identified in the NGS data of the corresponding samples, were subsequently intersected with these m$^6$A sites and their surrounding windows, followed by a calculation of SNV frequencies. Additionally, 300 random genomic sites without any SNVs were selected from the SARS-CoV-2 genome to serve as control random backgrounds.

### Differential analyses of malaria parasites

We utilized DESeq2 [110] (v 1.32.0) to identify differentially expressed genes and isoforms (Wald-test, $P$ value < 0.05). Nanopolish [4] (v0.13.2) was used to estimate the lengths of poly(A) tails. The Wilcoxon test was used to determine the significant changes in poly(A) tail length between the trophozoite and schizont stages of malaria parasites. The differential m$^6$A ratios were determined using $T$-tests with a significance cutoff of $P$ value < 0.05.

### Metagene motif distribution analysis

MetaplotR [111] (v1.0) was used to analyze the motifs of $m^6A$ peaks in DRS. The motif closest to DRACH was identified based on the annotation information of the reference genome downloaded from the UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/). The relative localization of the motif coordinates in the transcript regions (5′UTR, CDS, 3′UTR) was determined. The relative lengths of the three transcript regions were defined by scaling genes that contain at least one $m^6A$ peak.

### Metagenomic analysis

The demultiplexed DRS data were filtered according to the criteria of qual > 5 and length > 100. The U bases were converted to T bases for further analysis. The taxonomic classification and calculation of taxonomic abundance were performed using BugSeq [112] (v2023-11–27) with a default metagenomic database. For domain-level analysis, reads from all samples were pooled to identify and calculate the percentage of each domain. For genus discrimination, reads from nasopharyngeal and oropharyngeal swabs were analyzed separately, with each swab type showing the percentage of the top ten genera. For species discrimination, species with read counts above 10 were retained, and DESeq2 [110] (v1.32.0) was used to identify species that exhibited differential distribution between nasopharyngeal and oropharyngeal samples (Wald-test, $P$ value < 0.05).

### Genome assembly

For the RNA virus samples, including SVV, PRRSV, and metagenomic samples from COVID-19, we performed genome assembly using the following method. First, we utilized Canu (v2.2) [113] with default parameters to assemble the genomes. To assess the consistency with reference genomes, we used FastANI (v1.1) [114] to calculate the average nucleotide identity between the assembled genomes and the reference genomes. For the COVID-19 metagenomic samples, we used BLASTn (v2.12.0) [115] to map the assembled contigs to the NCBI non-redundant nucleotide database (nr/nt) for taxonomic annotation. We set the parameters *num_alignments* and *max_hsps* to 1 to identify the best match. Finally, we used TaxonKit (v0.16.0) [116] to complete the full taxonomic classification based on the respective TaxIDs.

### Sanger sequencing validation

For PRRSV samples, primers spanning the genome region from 13,941 to 15,340 were designed to verify all point mutations within that region. The primer sequences are listed in Additional file 5: Data S3. For glioblastoma samples, primers spanning four-point mutations were designed. The primer sequences are listed in Additional file 8: Data S6. Reverse transcription (RT) was performed using SuperScript IV (ThermoFisher Scientific, 18090050) with random hexamers. Amplification was performed using the KAPA HiFi HotStart PCR Kit (Roche, KK2501), and the sequences were determined by Sanger sequencing.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03536-3.

Additional file 1: Supplementary figures

Additional file 2: Supplementary tables (Tables S1–S5)

Additional file 3: Data S1. Data used to establish DEMINERS barcode classification algorithm

Additional file 4: Data S2. Data used to evaluate DEMINERS performance

Additional file 5: Data S3. DEMINERS performance on pathogen identification

Additional file 6: Data S4. DEMINERS performance on clinical metagenomics

Additional file 7: Data S5. DEMINERS applied on parallel comparisons of RNA features

Additional file 8: Data S6. DEMINERS performance on glioma samples

### Acknowledgements

We thank Dr. Xinqiong Li for providing us with the virus samples. We thank Coseque (Chengdu, China) for providing support in basecalling algorithm.

### Peer review information

Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

### Authors' contributions

LC, J-wL and JG conceived the project and designed the experiments. JS, CC developed a multiplexing experimental protocol and conducted the experiments. CT developed the demultiplexing pipelines. LL developed the basecalling method. YW, BY and BC assisted by JZ, JJ and JW provided the clinical samples. JS and CC assisted by YZ, H-cW, GS and MC prepared the RNA and conducted the direct RNA sequencing. JS and CT assisted by QY, DZ, KL, ZX, TC, ZH, DL, WZ and LC performed the bioinformatic analysis. LC, J-wL and JG analyzed the results. JS, LC and J-wL prepared the manuscript. JS, CT, and CC assisted by QY, DZ, ZX, YZ, and WZ generated the figures and helped prepare the manuscript. All authors read and approved the final manuscript.

### Funding

This work was supported by the National Natural Science Foundation of China (82341122, 92369116, 82370233, and 82300133), Sichuan Science and Technology Program (2021YFS0027), National Key Research and Development Program of China (2017YFA0106800), and China Postdoctoral Science Foundation (2023M742455).

### Availability of data and materials

The raw and processed sequencing data generated in this study were submitted to the NCBI BioProject database [117] (Accession number PRJNA911167). We also used previously published cDNA-NGS data [102] (Genome Sequence Archive, accession ID HRA001865, https://ngdc.cncb.ac.cn/gsa-human/browse/HRA001865). The software and related scripts for DEMINERS were uploaded to GitHub [118] (https://github.com/LuChenLab/DEMINERS) and the source scripts used in the manuscript were also uploaded to Zenodo [119] (https://zenodo.org/records/14616543). Both the software and related scripts on GitHub and Zenodo are released under the GNU General Public License v3.0. The pre-trained models for basecalling models were deposited in Figshare [120] (https://figshare.com/articles/dataset/Densecall_models/25712856). The pre-trained models for barcode demultiplexing were deposited in Figshare [121] (https://figshare.com/articles/online_resource/DecodeR_Models/22678729).

The software and related scripts for DEMINERS were uploaded to GitHub [118] (https://github.com/LuChenLab/DEMINERS) and the source scripts used in the manuscript were also uploaded to Zenodo [119] (https://zenodo.org/records/14616543). Both the software and related scripts on GitHub and Zenodo are released under the GNU General Public License v3.0.
The pre-trained models for basecalling models were deposited in Figshare [120] (https://figshare.com/articles/dataset/Densecall_models/25712856).
The pre-trained models for barcode demultiplexing were deposited in Figshare [121] (https://figshare.com/articles/online_resource/DecodeR_Models/22678729).

## Declarations

### Ethics approval and consent to participate

Tumour sample collection and the study design were approved by the Biomedical Research Ethics Committee of West China Hospital (Approval number: 2020.837). The research on COVID-19 specimens was approved by the Biomedical Research Ethics Committee of West China Hospital (Approval number 2020.100, 2020.193 and 2020.267). The swabs were obtained for routine diagnostic purposes, and the remaining RNA samples were provided for research. Written consents were obtained from all patients.

### Consent for publication

Not applicable.

**Competing interests**

Sichuan University has filed patent applications for the methods described herein, with L.C., J.-wL., J.G., J.S., C.T., L.L., and C.C. listed as inventors. The remaining authors declare no competing interests.

**Author details**

[1]Department of Laboratory Medicine, Key Laboratory of Birth Defects and Related Diseases of Women and Children of MOE, State Key Laboratory of Biotherapy, West China Second University Hospital, Sichuan University, Chengdu 610041, China. [2]Department of Laboratory Medicine, State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, Sichuan University and Collaborative Innovation Center, Chengdu 610041, China. [3]Biosafety Laboratory, International Center for Biological and Translational Research, West China Hospital, Sichuan University, Chengdu 610041, China. [4]School of Pharmacy, School of Basic Medical Sciences and Forensic Medicine, North Sichuan Medical College, Nanchong 637000, China. [5]Precision Medicine Center, Precision Medicine Key Laboratory of Sichuan Province, West China Hospital, Sichuan University, Chengdu 610041, China. [6]Department of Urology, Institute of Urology, West China Hospital, Sichuan University, Chengdu 610041, China.

Received: 22 August 2024 Accepted: 10 March 2025
Published online: 28 March 2025

**References**

1. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021;39:1348–65.
2. 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;10:3359–3314.
3. Hussain S. Native RNA-Sequencing Throws its Hat into the Transcriptomics Ring. Trends Biochem Sci. 2018;43:225–7.
4. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, Sadowski N, Holmes N, de Jesus JG, Jones KL, Soulette CM, Snutch TP, Loman N, Paten B, Loose M, Simpson JT, Olsen HE, Brooks AN, Akeson M, Timp W. Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat Methods. 2019;16:1297–305.
5. Ibrahim F, Oppelt J, Maragkakis M, Mourelatos Z. TERA-Seq: true end-to-end sequencing of native RNA molecules for transcriptome characterization. Nucleic Acids Res. 2021;49: e115.
6. Li R, Ren X, Ding Q, Bi Y, Xie D, Zhao Z. Direct full-length RNA sequencing reveals unexpected transcriptome complexity during Caenorhabditis elegans development. Genome Res. 2020;30:287–98.
7. Viehweger A, Krautwurst S, Lamkiewicz K, Madhugiri R, Ziebuhr J, Hölzer 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;29:1545–54.
8. Kim JH, Kim J, Koo B-S, Oh H, Hong J-J, Hwang E-S. Rapid whole-genome sequencing of zika viruses using direct RNA sequencing. J Bacteriol Virol. 2019;49:115.
9. Roach NP, Sadowski N, Alessi AF, Timp W, Taylor J, Kim JK. The full-length transcriptome of C. elegans using direct RNA sequencing. Genome Research. 2020;30:299–312.
10. Parker MT, Knop K, Sherwood AV, Schurch NJ, Mackinnon K, Gould PD, Hall AJW, Barton GJ, Simpson GG. Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. eLife. 2020;9:e49658.
11. Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The Architecture of SARS-CoV-2 Transcriptome. Cell. 2020;181(914–921): e910.
12. Liu H, Begik O, Lucas MC, Ramirez JM, Mason CE, Wiener D, Schwartz S, Mattick JS, Smith MA, Novoa EM. Accurate detection of m6A RNA modifications in native RNA sequences. Nat Commun. 2019;10:4079.
13. Gao Y, Liu X, Wu B, Wang H, Xi F, Kohnen MV, Reddy ASN, Gu L. Quantitative profiling of N6-methyladenosine at single-base resolution in stem-differentiating xylem of Populus trichocarpa using Nanopore direct RNA sequencing. Genome Biol. 2021;22:22.
14. Pratanwanich PN, Yao F, Chen Y, Koh CWQ, Wan YK, Hendra C, Poon P, Goh YT, Yap PML, Chooi JY, Chng WJ, Ng SB, Thiery A, Goh WSS, Goke J. Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. Nat Biotechnol. 2021;39(11):1394–402.
15. Lorenz DA, Sathe S, Einstein JM, Yeo GW. Direct RNA sequencing enables m6A detection in endogenous transcript isoforms at base-specific resolution. RNA. 2020;26:19–28.
16. Smith AM, Jain M, Mulroney L, Garalde DR, Akeson M. Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. PLoS ONE. 2019;14: e0216709.
17. Fleming AM, Burrows CJ. Nanopore sequencing for N1-methylpseudouridine in RNA reveals sequence-dependent discrimination of the modified nucleotide triphosphate during transcription. Nucleic Acids Res. 2023;51:1914–26.
18. Begik O, Lucas MC, Pryszcz LP, Ramirez JM, Medina R, Milenkovic I, Cruciani S, Liu H, Vieira HGS, Sas-Chen A, Mattick JS, Schwartz S, Novoa EM. Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. Nat Biotechnol. 2021;39:1278–91.
19. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, Jordan M, Ciccone J, Serra S, Keenan J, Martin S, McNeill L, Wallace EJ, Jayasinghe L, Wright C, Blasco J, Young S, Brocklebank D, Juul S, Clarke J, Heron AJ, Turner DJ. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods. 2018;15:201–6.

20. Schulz L, Torres-Diz M, Cortés-López M, Hayer KE, Asnani M, Tasian SK, Barash Y, Sotillo E, Zarnack K, König J, Thomas-Tikhonenko A. Direct long-read RNA sequencing identifies a subset of questionable exitrons likely arising from reverse transcription artifacts. Genome Biol. 2021;22:190.
21. Aw JGA, Lim SW, Wang JX, Lambert FRP, Tan WT, Shen Y, Zhang Y, Kaewsapsak P, Li C, Ng SB, Vardy LA, Tan MH, Nagarajan N, Wan Y. Determination of isoform-specific RNA structure with nanopore long reads. Nat Biotechnol. 2021;39:336–46.
22. Thomas NK, Poodari VC, Jain M, Olsen HE, Akeson M, Abu-Shumays RL. Direct nanopore sequencing of individual full length tRNA strands. ACS Nano. 2021;15:16642–53.
23. Lucas MC, Pryszcz LP, Medina R, Milenkovic I, Camacho N, Marchand V, Motorin Y, Ribas de Pouplana L, Novoa EM. Quantitative analysis of tRNA abundance and modifications by nanopore RNA sequencing. Nat Biotechnol. 2023;42(1):72–86.
24. Grädel C, Terrazos Miani MA, Baumann C, Barbani MT, Neuenschwander S, Leib SL, Suter-Riniker F, Ramette A. Whole-genome sequencing of human enteroviruses from clinical samples by nanopore direct RNA sequencing. Viruses. 2020;12:841.
25. Tan S, Dvorak CMT, Murtaugh MP. Rapid, unbiased PRRSV strain detection Using MinION direct RNA sequencing and bioinformatics Tools. Viruses. 2019;11:1132.
26. Maier KC, Gressel S, Cramer P, Schwalb B. Native molecule sequencing by nano-ID reveals synthesis and stability of RNA isoforms. Genome Res. 2020;30:1332–44.
27. Drexler HL, Choquet K, Churchman LS. Splicing kinetics and coordination revealed by direct nascent RNA sequencing through nanopores. Mol Cell. 2020;77:985-998.e988.
28. Fang Y, Chen G, Chen F, Hu E, Dong X, Li Z, He L, Sun Y, Qiu L, Xu H, Cai Z, Liu X. Accurate transcriptome assembly by Nanopore RNA sequencing reveals novel functional transcripts in hepatocellular carcinoma. Cancer Sci. 2021;112:3555–68.
29. Bayega A, Oikonomopoulos S, Gregoriou M-E, Tsoumani KT, Giakountis A, Wang YC, Mathiopoulos KD, Ragoussis J. Nanopore long-read RNA-seq and absolute quantification delineate transcription dynamics in early embryo development of an insect pest. Sci Rep. 2021;11: 117878.
30. He XJ, Barron AB, Yang L, Chen H, He YZ, Zhang LZ, Huang Q, Wang ZL, Wu XB, Yan WY, Zeng ZJ. Extent and complexity of RNA processing in honey bee queen and worker caste development. iScience. 2022;25:104301.
31. Zhao L, Zhang H, Kohnen MV, Prasad KVSK, Gu L, Reddy ASN. Analysis of transcriptome and epitranscriptome in plants using PacBio Iso-Seq and Nanopore-based direct RNA sequencing. Front Genet. 2019;10:253.
32. Zhang S, Li R, Zhang L, Chen S, Xie M, Yang L, Xia Y, Foyer CH, Zhao Z, Lam H-M. New insights into Arabidopsis transcriptome complexity revealed by direct sequencing of native RNAs. Nucleic Acids Res. 2020;48:7700–11.
33. Wang Y, Wang H, Xi F, Wang H, Han X, Wei W, Zhang H, Zhang Q, Zheng Y, Zhu Q, Kohnen MV, Reddy ASN, Gu L. Profiling of circular RNA N6-methyladenosine in moso bamboo (Phyllostachys edulis) using nanopore-based direct RNA sequencing. J Integr Plant Biol. 2020;62:1823–38.
34. Pitt ME, Nguyen SH, Duarte TPS, Teng H, Blaskovich MAT, Cooper MA, Coin LJM. Evaluating the genome and resistome of extensively drug-resistant Klebsiella pneumoniae using native DNA and RNA Nanopore sequencing. Gigascience. 2020;9:1–14.
35. Pust MM, Davenport CF, Wiehlmann L, Tummler B. Direct RNA Nanopore Sequencing of Pseudomonas aeruginosa Clone C Transcriptomes. J Bacteriol. 2022;204: e0041821.
36. Grunberger F, Juttner M, Knuppel R, Ferreira-Cerca S, Grohmann D. Nanopore-based RNA sequencing deciphers the formation, processing, and modification steps of rRNA intermediates in archaea. RNA. 2023;29:1255–73.
37. Wongsurawat T, Jenjaroenpun P, Wanchai V, Nookaew I: Native RNA or cDNA Sequencing for Transcriptomic Analysis. A Case Study on Saccharomyces cerevisiae. Frontiers in Bioengineering and Biotechnology. 2022;10:842299.
38. Semmouri I, De Schamphelaere KAC, Mees J, Janssen CR, Asselman J. Evaluating the potential of direct RNA nanopore sequencing: Metatranscriptomics highlights possible seasonal differences in a marine pelagic crustacean zooplankton community. Mar Environ Res. 2020;153: 104836.
39. Torma G, Tombácz D, Csabai Z, Moldován N, Mészáros I, Zádori Z, Boldogkői Z. Combined Short and Long-Read Sequencing Reveals a Complex Transcriptomic Architecture of African Swine Fever Virus. Viruses. 2021;13:579.
40. Depledge DP, Srinivas KP, Sadaoka T, Bready D, Mori Y, Placantonakis DG, Mohr I, Wilson AC. Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. Nat Commun. 2019;10:754.
41. Tombácz D, Sharon D, Szűcs A, Moldován N, Snyder M, Boldogkői Z. Transcriptome-wide survey of pseudorabies virus using next- and third-generation sequencing platforms. Scientific Data. 2018;5: 180119.
42. Boldogkői Z, Moldován N, Szűcs A, Tombácz D. Transcriptome-wide analysis of a baculovirus using nanopore sequencing. Scientific Data. 2018;5: 180276.
43. Tombácz D, Prazsák I, Szűcs A, Dénes B, Snyder M, Boldogkői Z. Dynamic transcriptome profiling dataset of vaccinia virus obtained from long-read sequencing techniques. GigaScience. 2018;7:1–17.
44. Price AM, Hayer KE, McIntyre ABR, Gokhale NS, Abebe JS, Della Fera AN, Mason CE, Horner SM, Wilson AC, Depledge DP, Weitzman MD. Direct RNA sequencing reveals m(6)A modifications on adenovirus RNA are necessary for efficient splicing. Nat Commun. 2020;11:6016.
45. Lee VV, Judd LM, Jex AR, Holt KE, Tonkin CJ, Ralph SA. Flórez De Sessions P: Direct nanopore sequencing of mRNA reveals landscape of transcript isoforms in apicomplexan parasites. mSystems. 2021;6:e0108120.
46. Ono H. Yoshida M-a: Direct RNA sequencing approach to compare non-model mitochondrial transcriptomes: An application to a cephalopod host and its mesozoan parasite. Methods. 2020;176:55–61.
47. Kruse E, Göringer HU. Nanopore-Based Direct RNA Sequencing of the Trypanosoma brucei Transcriptome Identifies Novel lncRNAs. Genes (Basel). 2023;14(3):610.
48. Zhao N, Cao J, Xu J, Liu B, Liu B, Chen D, Xia B, Chen L, Zhang W, Zhang Y, Zhang X, Duan Z, Wang K, Xie F, Xiao K, Yan W, Xie L, Zhou H, Wang J. Targeting RNA with Next- and Third-Generation Sequencing Improves Pathogen Identification in Clinical Samples. Adv Sci (Weinh). 2021;8: e2102593.

49. Vacca D, Fiannaca A, Tramuto F, Cancila V, La Paglia L, Mazzucco W, Gulino A, La Rosa M, Maida CM, Morello G, Belmonte B, Casuccio A, Maugeri R, Iacopino G, Balistreri CR, Vitale F, Tripodo C, Urso A. Direct RNA Nanopore Sequencing of SARS-CoV-2 Extracted from Critical Material from Swabs. Life (Basel). 2022;12(1):69.
50. Teufel M, Sobetzko P. Reducing costs for DNA and RNA sequencing by sample pooling using a metagenomic approach. BMC Genomics. 2022;23:613.
51. Chang H: In Poreplex, A versatile sequenced read processor for nanopore direct RNA sequencing Available online: https://www.githubcom/hyeshik/poreplex; 2019.
52. Smith MA, Ersavas T, Ferguson JM, Liu H, Lucas MC, Begik O, Bojarski L, Barton K, Novoa EM. Molecular barcoding of native RNAs using nanopore sequencing and deep learning. Genome Res. 2020;30:1345–53.
53. Jain M, Abu-Shumays R, Olsen HE, Akeson M. Advances in nanopore direct RNA sequencing. Nat Methods. 2022;19:1160–4.
54. Neumann D, Reddy ASN, Ben-Hur A. RODAN: a fully convolutional architecture for basecalling nanopore RNA sequencing data. BMC Bioinformatics. 2022;23:142.
55. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ: Densely Connected Convolutional Networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269; 2017:2261–2269.
56. Neumann D, Reddy AS. Ben-Hur AJBb: RODAN: a fully convolutional architecture for basecalling nanopore RNA sequencing data. 2022;23:142.
57. Graves A, Fernández S, Gomez F, Schmidhuber J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning. 2006;6:369–76.
58. Pleiss G, Chen D, Huang G, Li T, Van Der Maaten L, Weinberger KJapa: Memory-efficient implementation of densenets. arXiv 2017.
59. Chao A. Estimating the population size for capture-recapture data with unequal catchability. Biometrics. 1987;43:783–91.
60. Shannon CE: The mathematical theory of communication. 1963. MD Comput 1997, 14:306–317.
61. Simpson EH. Measurement of Diversity. Nature. 1949;163:688–688.
62. Ma S, Zhang F, Zhou F, Li H, Ge W, Gan R, Nie H, Li B, Wang Y, Wu M, Li D, Wang D, Wang Z, You Y, Huang Z. Metagenomic analysis reveals oropharyngeal microbiota alterations in patients with COVID-19. Signal Transduct Target Ther. 2021;6:191.
63. Hendra C, Pratanwanich PN, Wan YK, Goh WSS, Thiery A, Goke J. Detection of m6A from direct RNA sequencing using a multiple instance learning framework. Nat Methods. 2022;19:1590–8.
64. Wu Y, Shao W, Yan M, Wang Y, Xu P, Huang G, Li X, Gregory BD, Yang J, Wang H, Yu X. Transfer learning enables identification of multiple types of RNA modifications using nanopore direct RNA sequencing. Nat Commun. 2024;15:4049.
65. Liu J, Xu YP, Li K, Ye Q, Zhou HY, Sun H, Li X, Yu L, Deng YQ, Li RT, Cheng ML, He B, Zhou J, Li XF, Wu A, Yi C, Qin CF. The m(6)A methylome of SARS-CoV-2 in host cells. Cell Res. 2021;31:404–14.
66. Eliana C, Javier E, Moises W. Plasmodium falciparum spliceosomal RNAs: 3' and 5' end processing. Acta Trop. 2011;117:105–8.
67. Deng X, Qing Y, Horne D, Huang H, Chen J. The roles and implications of RNA m(6)A modification in cancer. Nat Rev Clin Oncol. 2023;20:507–26.
68. Delaunay S, Frye M. RNA modifications regulating cell fate in cancer. Nat Cell Biol. 2019;21:552–9.
69. Xue C, Chu Q, Zheng Q, Jiang S, Bao Z, Su Y, Lu J, Li L. Role of main RNA modifications in cancer. N6-methyladenosine, 5-methylcytosine, and pseudouridine. Signal Transduction and Targeted Therapy. 2022;7(1):142.
70. Kahles A, Lehmann KV, Toussaint NC, Huser M, Stark SG, Sachsenberg T, Stegle O, Kohlbacher O, Sander C. Cancer Genome Atlas Research N, Ratsch G: Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell. 2018;34(211–224): e216.
71. Tang Y, Chen K, Song B, Ma J, Wu X, Xu Q, Wei Z, Su J, Liu G, Rong R, Lu Z, de Magalhães JP, Rigden DJ, Meng J. m6A-Atlas: a comprehensive knowledgebase for unraveling theN 6-methyladenosine (m6A) epitranscriptome. Nucleic Acids Res. 2021;49:D134–43.
72. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, Ohgaki H, Wiestler OD, Kleihues P, Ellison DW. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol. 2016;131:803–20.
73. Visvanathan A, Patil V, Arora A, Hegde AS, Arivazhagan A, Santosh V, Somasundaram K. Essential role of METTL3-mediated m(6)A modification in glioma stem-like cells maintenance and radioresistance. Oncogene. 2018;37:522–33.
74. Niu X, Yang Y, Ren Y, Zhou S, Mao Q, Wang Y. Crosstalk between m(6)A regulators and mRNA during cancer progression. Oncogene. 2022;41:4407–19.
75. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
76. Li Q, Lai H, Li Y, Chen B, Chen S, Li Y, Huang Z, Meng Z, Wang P, Hu Z, Huang S. RJunBase: a database of RNA splice junctions in human normal and cancerous tissues. Nucleic Acids Res. 2021;49:D201–11.
77. Somervuo P, Koskinen P, Mei P, Holm L, Auvinen P, Paulin L. BARCOSEL: a tool for selecting an optimal barcode set for high-throughput sequencing. BMC Bioinformatics. 2018;19:257.
78. Buschmann T, Bystrykh LV. Levenshtein error-correcting barcodes for multiplexed DNA sequencing. BMC Bioinformatics. 2013;14:272.
79. Trotter AJ, Aydin A, Strinden MJ, O'Grady J. Recent and emerging technologies for the rapid diagnosis of infection and antimicrobial resistance. Curr Opin Microbiol. 2019;51:39–45.
80. Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, Rae D, Grundy S, Turner DJ, Wain J, Leggett RM, Livermore DM, O'Grady J. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat Biotechnol. 2019;37:783–92.
81. Hitzenbichler F, Bauernfeind S, Salzberger B, Schmidt B, Wenzel JJ. Comparison of Throat Washings, Nasopharyngeal Swabs and Oropharyngeal Swabs for Detection of SARS-CoV-2. Viruses. 2021;13(4):653.

82. Mostafa HH, Fissel JA, Fanelli B, Bergman Y, Gniazdowski V, Dadlani M, Carroll KC, Colwell RR, Simner PJ. Metagenomic Next-Generation Sequencing of Nasopharyngeal Specimens Collected from Confirmed and Suspect COVID-19 Patients. mBio. 2020;11(6):e01969–20.
83. Merenstein C, Bushman FD, Collman RG. Alterations in the respiratory tract microbiome in COVID-19: current observations and potential significance. Microbiome. 2022;10:165.
84. Chen S, Zhang L, Li M, Zhang Y, Sun M, Wang L, Lin J, Cui Y, Chen Q, Jin C, Li X, Wang B, Chen H, Zhou T, Wang L, Hsu CH, Zhuo W. Fusobacterium nucleatum reduces METTL3-mediated m(6)A modification and contributes to colorectal cancer metastasis. Nat Commun. 2022;13:1248.
85. Tsai K, Cullen BR. Epigenetic and epitranscriptomic regulation of viral replication. Nat Rev Microbiol. 2020;18:559–70.
86. Delaunay S, Helm M, Frye M. RNA modifications in physiology and disease: towards clinical applications. Nat Rev Genet. 2023;25(2):104–22.
87. Huang H, Weng H, Chen J. m(6)A Modification in Coding and Non-coding RNAs: Roles and Therapeutic Implications in Cancer. Cancer Cell. 2020;37:270–88.
88. Fu Y, Dominissini D, Rechavi G, He C. Gene expression regulation mediated through reversible m6A RNA methylation. Nat Rev Genet. 2014;15:293–306.
89. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
90. Killick R, Eckley IA. changepoint: An R Package for Changepoint Analysis. J Stat Softw. 2014;58:1–19.
91. Killick R, Fearnhead P, Eckley IA. Optimal Detection of Changepoints With a Linear Computational Cost. J Am Stat Assoc. 2012;107:1590–8.
92. Scott AJ, Knott M. A Cluster Analysis Method for Grouping Means in the Analysis of Variance. Biometrics. 1974;30:507–12.
93. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Softw. 2008;28:1–26.
94. Ramachandran P, Zoph B, Le QV. Searching for Activation Functions. arXiv. 2017;1710.05941.
95. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S: PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32. Curran Associates Inc.; 2019: 8024--8035
96. Janse CJ, Ramesar J, Waters AP. High-efficiency transfection and drug selection of genetically transformed blood stages of the rodent malaria parasite Plasmodium berghei. Nat Protoc. 2006;1:346–56.
97. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9: e1003118.
98. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178–92.
99. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
100. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
101. Lin JW, Tang C, Wei HC, Du B, Chen C, Wang M, Zhou Y, Yu MX, Cheng L, Kuivanen S, Ogando NS, Levanov L, Zhao Y, Li CL, Zhou R, Li Z, Zhang Y, Sun K, Wang C, Chen L, Xiao X, Zheng X, Chen SS, Zhou Z, Yang R, Zhang D, Xu M, Song J, Wang D, Li Y, Lei S, Zeng W, Yang Q, He P, Zhang Y, Zhou L, Cao L, Luo F, Liu H, Wang L, Ye F, Zhang M, Li M, Fan W, Li X, Li K, Ke B, Xu J, Yang H, He S, Pan M, Yan Y, Zha Y, Jiang L, Yu C, Liu Y, Xu Z, Li Q, Jiang Y, Sun J, Hong W, Wei H, Lu G, Vapalahti O, Luo Y, Wei Y, Connor T, Tan W, Snijder EJ, Smura T, Li W, Geng J, Ying B, Chen L. Genomic monitoring of SARS-CoV-2 uncovers an Nsp1 deletion variant that modulates type I interferon response. Cell Host Microbe. 2021;29(489–502): e488.
102. Ren Y, Huang Z, Zhou L, Xiao P, Song J, He P, Xie C, Zhou R, Li M, Dong X, Mao Q, You C, Xu J, Liu Y, Lan Z, Zhang T, Gan Q, Yang Y, Chen T, Huang B, Yang X, Xiao A, Ou Y, Su Z, Chen L, Zhang Y, Ju Y, Zhang Y, Wang Y. Spatial transcriptomics reveals niche-specific enrichment and vulnerabilities of radial glial stem-like cells in malignant gliomas. Nat Commun. 2023;14(1):1028.
103. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. Genome Project Data Processing S: The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
104. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics. 2016;32:1749–51.
105. Wyman D, Mortazavi A. TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. Bioinformatics. 2019;35:340–2.
106. Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, Eyras E. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. 2018;19:40.
107. Grinev VV, Yatskou MM, Skakun VV, Chepeleva MK, Nazarov PV. ORFhunteR: An accurate approach to the automatic identification and annotation of open reading frames in human mRNA molecules. Software Impacts. 2022;12:100268.
108. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, Bodenstein SW, Evans DA, Hung CC, O'Neill M, Reiman D, Tunyasuvunakool K, Wu Z, Zemgulyte A, Arvaniti E, Beattie C, Bertolli O, Bridgland A, Cherepanov A, Congreve M, Cowen-Rivers AI, Cowie A, Figurnov M, Fuchs FB, Gladman H, Jain R, Khan YA, Low CMR, Perlin K, Potapenko A, Savy P, Singh S, Stecula A, Thillaisundaram A, Tong C, Yakneen S, Zhong ED, Zielinski M, Zidek A, Bapst V, Kohli P, Jaderberg M, Hassabis D, Jumper JM. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630(8016):493–500.
109. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40.
110. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
111. Olarerin-George AO, Jaffrey SR. MetaPlotR: a Perl/R pipeline for plotting metagenes of nucleotide modifications and other transcriptomic sites. Bioinformatics. 2017;33:1563–4.

112. Fan J, Huang S, Chorlton SD. BugSeq: a highly accurate cloud platform for long-read metagenomic analyses. BMC Bioinformatics. 2021;22:160.
113. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH. Phillippy AMJGr: Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. 2017;27:722–36.
114. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9(1):5114.
115. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
116. Shen W, Ren H. TaxonKit: A practical and efficient NCBI taxonomy toolkit. J Genet Genomics. 2021;48:844–50.
117. Junwei S, Li-an L, Chao T, Chuan C, Qingxin Y, Dan Z, Yuancun Z, Han-cheng W, Kepan L, Zijie X, Tingfeng C, Zhifeng H, Defu L, Yu Z, Weizhen Z, Wanqin Z, Li C, Guiqin S, Mutian C, Juan J, Juan Z, Jing W, Bojiang C, Binwu Y, Yuan W, Jia G, Jing-wen L, Lu C. DEMINERS enables clinical metagenomics and comparative transcriptomic analysis by increasing throughput and accuracy of nanopore direct RNA sequencing. NCBI BioProject Database. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA911167. (2025)
118. Junwei S, Li-an L, Chao T, Chuan C, Qingxin Y, Dan Z, Yuancun Z, Han-cheng W, Kepan L, Zijie X, Tingfeng C, Zhifeng H, Defu L, Yu Z, Weizhen Z, Wanqin Z, Li C, Guiqin S, Mutian C, Juan J, Juan Z, Jing W, Bojiang C, Binwu Y, Yuan W, Jia G, Jing-wen L, Lu C. DEMINERS enables clinical metagenomics and comparative transcriptomic analysis by increasing throughput and accuracy of nanopore direct RNA sequencing. Github. https://github.com/LuChenLab/DEMINERS. (2025)
119. Junwei S, Li-an L, Chao T, Chuan C, Qingxin Y, Dan Z, Yuancun Z, Han-cheng W, Kepan L, Zijie X, Tingfeng C, Zhifeng H, Defu L, Yu Z, Weizhen Z, Wanqin Z, Li C, Guiqin S, Mutian C, Juan J, Juan Z, Jing W, Bojiang C, Binwu Y, Yuan W, Jia G, Jing-wen L, Lu C. DEMINERS enables clinical metagenomics and comparative transcriptomic analysis by increasing throughput and accuracy of nanopore direct RNA sequencing. Zenodo. https://zenodo.org/records/14616543. (2025)
120. Junwei S, Li-an L, Chao T, Chuan C, Qingxin Y, Dan Z, Yuancun Z, Han-cheng W, Kepan L, Zijie X, Tingfeng C, Zhifeng H, Defu L, Yu Z, Weizhen Z, Wanqin Z, Li C, Guiqin S, Mutian C, Juan J, Juan Z, Jing W, Bojiang C, Binwu Y, Yuan W, Jia G, Jing-wen L, Lu C. DEMINERS enables clinical metagenomics and comparative transcriptomic analysis by increasing throughput and accuracy of nanopore direct RNA sequencing. Figshare. https://figshare.com/articles/dataset/Densecall_models/25712856. (2025)
121. Junwei S, Li-an L, Chao T, Chuan C, Qingxin Y, Dan Z, Yuancun Z, Han-cheng W, Kepan L, Zijie X, Tingfeng C, Zhifeng H, Defu L, Yu Z, Weizhen Z, Wanqin Z, Li C, Guiqin S, Mutian C, Juan J, Juan Z, Jing W, Bojiang C, Binwu Y, Yuan W, Jia G, Jing-wen L, Lu C. DEMINERS enables clinical metagenomics and comparative transcriptomic analysis by increasing throughput and accuracy of nanopore direct RNA sequencing. Figshare. https://figshare.com/articles/online_resource/DecodeR_Models/22678729. (2025)

## Publisher's Note