Genome Biology

METHODOLOGY

Open Access

Evaluating the potential and limitations of nanopore adaptive sampling for targeted transcriptome sequencing

Nicole DeBruyne[1], Feng Wang[2], Yang Xu[2,3] and Lan Lin[2,4,5,6*]

*Correspondence:
linlan@chop.edu

[1] Graduate Group in Cell and Molecular Biology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
[2] Center for Computational and Genomic Medicine, Children's Hospital of Philadelphia, Philadelphia, PA, USA
[3] Division of Biostatistics and Bioinformatics, Department of Pharmacology, Physiology, and Cancer Biology, Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, PA, USA
[4] Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
[5] Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, PA, USA
[6] Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA, USA

Abstract

Long-read RNA sequencing is a powerful technology for transcriptomics, but low throughput and high cost pose challenges. Adaptive sampling, a feature of Oxford Nanopore Technologies, offers real-time enrichment by selectively ejecting non-target molecules. We evaluate adaptive sampling for human transcriptome analysis. Adaptive sampling modestly enriches target transcripts (1.3 × for cDNA sequencing, 1.9 × for direct RNA sequencing) while preserving gene expression and splicing profiles, but is significantly less effective than cDNA hybridization capture. Short read lengths and low sequencing quality limit performance. Adaptive sampling on direct RNA sequencing can boost target yield (~ 20%) within fixed run times, potentially aiding time-constrained applications.

Keywords: Transcriptomics, Long-read sequencing, RNA sequencing, Targeted sequencing, Selective sequencing, Adaptive sampling, Oxford Nanopore Technologies

Background

RNA sequencing enables comprehensive evaluation of the transcriptome. Long-read RNA sequencing is especially useful for full-length isoform analysis and proper identification of repetitive or highly homologous transcripts [1, 2]. However, in many biomedical and clinical applications, users may only be interested in a subset of transcripts. For instance, clinicians may wish to screen the expression of a small panel of disease-relevant genes [3]. Similarly, researchers may be interested in evaluating gene expression associated with a particular biological process or surveying low-abundance transcripts from a specific set of genes [4–7]. Achieving sufficient coverage for transcripts of interest may require greater sequencing depth and increased cost. Such cost considerations are particularly relevant for long-read RNA sequencing experiments, which are 3 to 10 times more expensive per unit of sequencing output compared to standard short-read platforms such as Illumina [8]. To address this issue, experimental enrichment approaches can be used to increase the proportion of target

transcripts in an RNA sequencing library before sequencing. Various approaches for targeted RNA sequencing exist, each optimized for different use cases. PCR amplicon sequencing is effective for small gene panels but becomes impractical for hundreds of targets due to primer design complexity and potential amplification bias, particularly for gene targets with numerous alternatively spliced transcript isoforms [9]. Cas9-mediated methods such as DASH are primarily used for the depletion of a few highly abundant transcripts and are also not well-suited for targeted enrichment of large panels [10, 11]. Hybridization capture is scalable to large gene panels representing a small portion of the transcriptome and yields high enrichment, but potential drawbacks to this approach include the extended library preparation time and requirement for additional resources such as large biotinylated oligo pools [12–14].

As an alternative to experimental enrichment, Oxford Nanopore Technologies (ONT) offers real-time selective sequencing from un-enriched libraries. ONT applies a voltage gradient across membrane-embedded protein pores and measures changes in the electrical current as a single molecule of DNA or RNA passes through each pore. Each series of nucleotides alters the current in a characteristic manner, which can be converted into a sequence of individual bases [15]. The nanopore sequencer can reverse the direction of the electric field across individual pores, a function used to clear out strands that become lodged in the pore (Additional file 1: Fig. S1, "unblocking") [16]. By rapidly determining whether a read originates from a region of interest, non-target strands can be ejected from the pore before they are fully sequenced (Fig. 1A). Early selective sequencing methods involved converting the reference sequence into a simulated nanopore electrical signal for comparison with live raw data from each pore, but this process was computationally demanding. Recent advancements in high-speed GPU basecalling using ONT's guppy basecaller have enabled real-time basecalling and alignment to a reference FASTA file [17]. Selective sequencing is now available directly through ONT's user interface, MinKNOW, via a function called adaptive sampling (AS). Users can elect to perform AS in either "enrichment mode" where reads aligning to the input FASTA file are accepted, or "depletion mode" where reads aligning to the reference file are rejected.

Previous studies have shown the utility of ONT AS for selective enrichment of genomic DNA, with one study reporting up to 13.5 × enrichment [18]. These studies often utilize AS to enrich specific microbial DNA from metagenomic samples or to achieve higher coverage of disease-relevant genes for identifying structural variants in time-sensitive human clinical samples [18–20]. However, few studies have utilized AS for transcriptome analysis [21, 22]. The relatively short length of mRNA molecules poses a significant challenge, as it limits the effectiveness of AS for transcript enrichment. Specifically, while genomic DNA is typically fragmented into 10–20 kb segments, human mRNA molecules average only around 1 kb in length [23]. Although ONT reports that an AS decision can be made after sequencing 200 bases [16], the process of basecalling and alignment lags behind raw data acquisition. The median number of bases sequenced before a rejection decision is made during AS has been shown to be over 500 bp for DNA molecules [18, 24]. Rejecting a non-target genomic DNA fragment can conserve over 90% of the resources that would otherwise be used sequencing it. In contrast, the benefits of rejecting non-target RNA or cDNA molecules are limited, as significant time

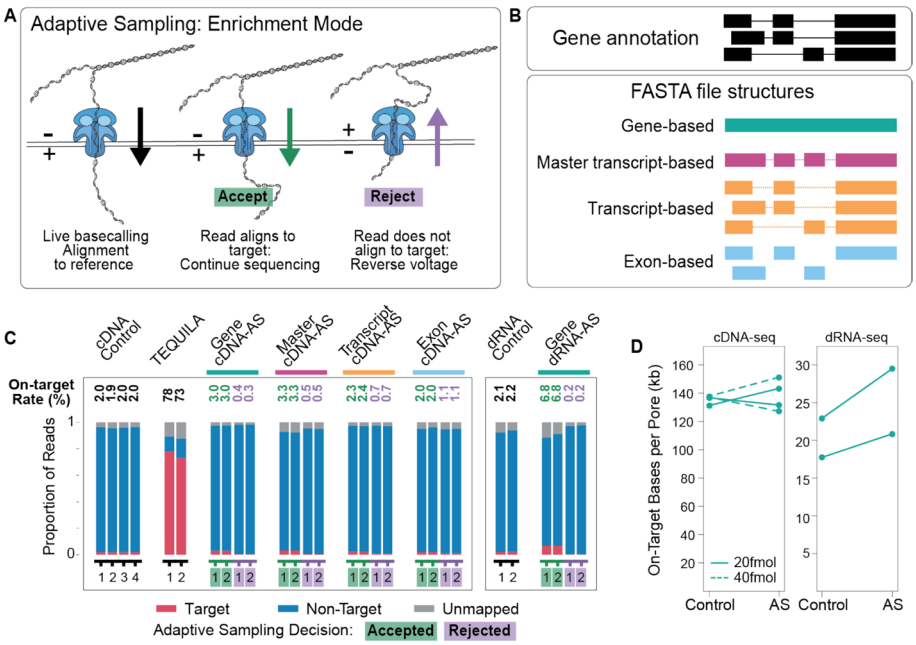

Fig. 1 **A** Overview of ONT's adaptive sampling feature. Reads undergo live basecalling and alignment to a user-supplied reference FASTA sequence. In "enrichment mode," reads that are determined not to align to target loci are ejected by reversing the voltage across the pore. **B** Four reference sequence structures. Boxes: Exons. Lines: Introns. **C** Proportion of reads mapped to 221 target splicing factors in each sequencing experiment using RNAs from SY5Y cell line with spike-in RNAs (SIRV-Set 4). "cDNA Control": standard whole transcriptome cDNA sequencing, "TEQUILA": TEQUILA-seq probe-based transcript enrichment, "Gene cDNA-AS": adaptive sampling on a cDNA library using the gene-based reference file, "Master cDNA-AS": adaptive sampling on a cDNA library using the master-transcript-based reference file, "Transcript cDNA-AS": adaptive sampling on a cDNA library using the transcript-based reference file, "Exon cDNA-AS": adaptive sampling on a cDNA library using the exon-based reference file, "dRNA Control": standard whole transcriptome direct RNA sequencing, and "Gene dRNA-AS": adaptive sampling on an RNA library using the gene-based reference file. AS: adaptive sampling. Replicates for each condition are presented along the *x*-axis as "1," "2," "3," and "4." Adaptive sampling experiments are split into accepted (green) and rejected (purple) pools. **D** The number of bases in reads mapped to 221 target splicing factors generated from sequencing SY5Y + SIRV-Set 4 samples, adjusted for the number of pores available at the first pore scan, which occurs at the very start of the sequencing run. Gene-based (teal) adaptive sampling ("AS") was performed on half (1500 channels, 6000 pores) of a PromethION flow cell, while the other half performed standard whole transcriptome cDNA or direct RNA ("Control") sequencing

and flow cell capacity are used to sequence nearly half of all non-target transcripts before their rejection.

Nearly all studies to date that have used AS for transcript enrichment have done so from native (direct) RNA libraries and have reported only modest enrichment of target transcripts [21, 22]. RNA molecules translocate through the nanopore more slowly than cDNA molecules, with RNA moving at a rate of 260 bases per second, compared to 400 bases per second for cDNA. Therefore, decisions can be made after sequencing ~ 300 bases [22]. However, due to the higher RNA input requirement (1 μg total RNA) and significantly lower sequencing output for direct RNA sequencing, cDNA-based AS is still an attractive option for targeted analysis of transcriptome profiles.

In this study, we systematically evaluate the performance of ONT's built-in AS feature for selective sequencing of cDNA and direct RNA libraries. We performed ONT AS under various sequencing configurations and compared its performance to

TEQUILA-seq, a hybridization-capture approach [14]. This study offers valuable insights into the capabilities and limitations of ONT AS for targeted analysis of complex (e.g., human) transcriptomes.

Results

Adaptive sampling of cDNA or native RNA molecules results in modest enrichment of target transcripts relative to overall sequencing output

To evaluate the performance of ONT AS on cDNA molecules (cDNA-AS), we performed AS in "enrichment mode" on cDNA libraries generated from SH-SY5Y cells. These libraries included spike-in RNA variants (SIRV-Set 4) and were used to selectively target a panel of 221 genes encoding splicing factors, 46 transcripts from the External RNA Controls Consortium (ERCC), and 5 long SIRVs (Additional file 2: Table S1 and S2) [14]. Standard ONT whole-transcriptome cDNA sequencing performed in parallel showed that SIRV-Set 4 RNAs constitute < 1% of the cDNA library. In the unenriched samples, transcripts from the 221 splicing factors account for 2% of all reads mapped to the human genome, while target ERCC/SIRV transcripts account for 30% of reads mapped to the SIRV-Set 4 genome (Additional file 2: Table S3).

RNA molecules contain large gaps corresponding to intronic regions when compared to the genomic sequence, which may influence alignment during AS depending on the structure of the reference file. ONT recommends that AS be performed by aligning sequencing reads to bacterial-sized genomes [16], which are a few million base pairs in size [25]. Therefore, the structure and file size of the reference sequence may influence the performance of AS when targeting numerous genes in complex mammalian transcriptomes, such as the human transcriptome. However, this topic has not been explicitly explored before. To investigate this, we tested four different FASTA file structures with varying data organizations and file sizes: (1) "Gene-based": a full-length sequence for each target gene locus, including all annotated exonic and intronic regions, 272 sequences, 16.6 Mb in file size; (2) "Master-transcript-based": a single "master" transcript for each target gene, generated by concatenating all annotated exons of the gene, 272 sequences, 1.5 Mb in file size; (3) "Transcript-based": full-length cDNA sequences for every annotated transcript of the target genes, 2675 sequences, 122.3 Mb in file size; and (4) "Exon-based": individual non-redundant exons from all annotated transcripts of the target genes, 7343 sequences, 2.9 Mb in file size (Fig. 1B).

Gene-based and master-transcript-based cDNA-AS demonstrated the best performance. Transcripts from the 221 splicing factors yielded an average of 3.01% gene-based (rep1: 0.54 M on-target accepted reads/18.08 M total accepted reads, rep2: 0.36 M on-target accepted reads/11.87 M total accepted reads) and 3.28% master-transcript-based (rep1: 0.44 M on-target accepted reads/13.47 M total accepted reads, rep2: 0.49 M on-target accepted reads/14.77 M total accepted reads) "accepted read proportion," which refers to the proportion of reads from on-target transcripts in the accepted pool. This corresponds to 1.55 × and 1.70 × enrichment, respectively, compared to standard ONT whole transcriptome cDNA sequencing. In contrast, transcript-based (1.23 × enrichment; 2.38%, rep1: 0.41 M on-target accepted reads/17.50 M total accepted reads, rep2: 0.43 M on-target accepted reads/17.68 M total accepted reads) and exon-based (1.04 × enrichment; 2.01%, rep1: 0.39 M on-target accepted reads/19.21 M total accepted

Table 1 Enrichment fold of cDNA- and direct RNA-based AS

Sample	Target panel[1]	Adaptive sampling file/ sequencing format	Enrichment fold	
			Accepted read proportion[2]	Base proportion[3]
SY5Y/SIRV	A	Gene/cDNA	1.55	1.31
BT549	B		1.70	1.30
T47D	B		1.62	1.30
SY5Y	A	Master-transcript/cDNA	1.70	1.34
BT549	B		1.75	1.30
T47D	B		1.74	1.34
SY5Y/SIRV	A	Transcript/cDNA	1.23	1.14
SY5Y/SIRV	A	Exon/cDNA	1.04	1.03
SY5Y/SIRV	A	Gene/dRNA	3.16	1.90

[1] Target panel A: 221 splicing factors, 46 ERCC transcripts, and 5 long-SIRVs. Target panel B: 468 actionable cancer genes. SIRV: SIRV-Set 4. For the full list of genes, see Additional file 2: Table S1 and Table S4

[2] Read proportion is calculated as the number of on-target reads in the accepted pool divided by the total number of reads in the accepted pool

[3] Base proportion is calculated as the number of bases in on-target reads divided by the total number of bases sequenced

reads, rep2: 0.38 M on-target accepted reads/19.05 M total accepted reads) cDNA-AS resulted in minimal to no enrichment in "accepted read proportion" (Fig. 1C, Additional file 1: Fig. S2, Table 1).

However, these values do not fully reflect the enrichment relative to total sequencing output, as only 62% (gene-based), 49% (master-transcript-based), 72% (transcript-based), and 87% (exon-based) of reads were accepted during the AS process (Additional file 2: Table S3). To provide a more accurate assessment of overall enrichment, we evaluated the "base proportion," defined as the proportion of bases from on-target transcripts relative to the total bases sequenced. Gene-based and master-transcript-based cDNA-AS enriched the "base proportion" by 1.31 × (2.83%, rep1: 0.62B on-target bases/22.00B total bases, rep2: 0.40B on-target bases/14.31B total bases) and 1.34 × (2.88%, rep1: 0.49B on-target bases/17.13B total bases, rep2: 0.55B on-target bases/19.17B total bases), respectively, while transcript-based and exon-based cDNA-AS yielded 1.14 × (2.47%, rep1: 0.47B on-target bases/19.36B total bases, rep2: 0.50B on-target bases/19.84B total bases) and 1.03 × (2.22%, rep1: 0.44B on-target bases/19.91B total bases, rep2: 0.45B on-target bases/20.08B total bases) enrichment in "base proportion" (Additional file 1: Fig. S3, Table 1).

It is worth noting that a small percentage of rejected reads also map to target genes. On-target reads account for 0.35% of all reads in the rejected pool for gene-based cDNA-AS and 0.55% for master-transcript-based cDNA-AS (Fig. 1C). Since these reads were rejected during sequencing, in theory they do not represent full-length transcripts and are therefore not suitable for full-length transcript isoform analyses. Overall, 6% and 14% of all on-target reads were incorrectly rejected in the gene-based and master-transcript-based cDNA-AS experiments, respectively (Fig. 2A).

We compared these results to the enrichment achieved through direct RNA AS (dRNA-AS) on the same SY5Y + SIRV-Set 4 RNA samples. Gene-based dRNA-AS yielded an "accepted read proportion" of 6.79% (rep1: 0.057 M on-target accepted

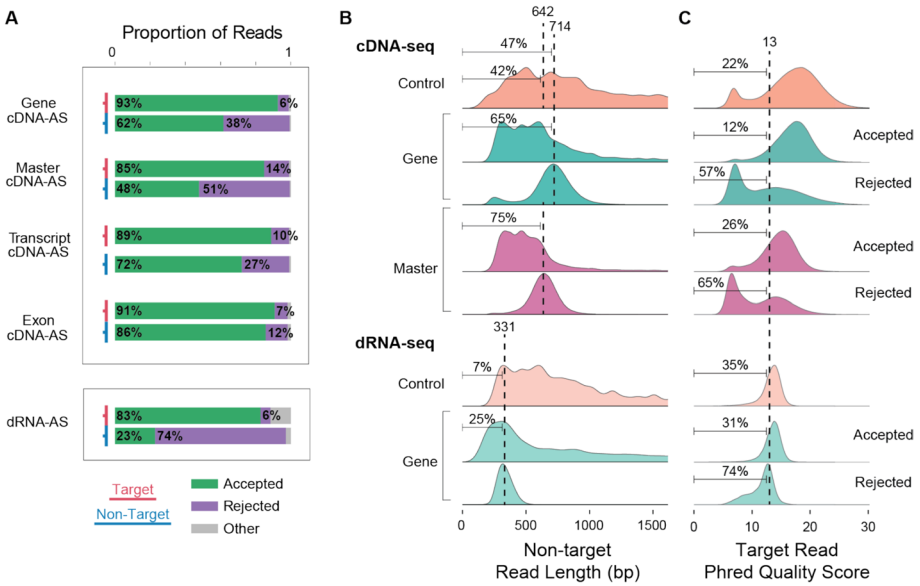

Fig. 2 **A** Proportion of target (red) and non-target (blue) reads that were accepted (green), rejected (purple), or other (gray) in adaptive sampling experiments. **B** Density plots of read lengths of non-target reads and **C** Phred quality scores (*q*-scores) of target reads in standard whole transcriptome cDNA or direct RNA ("Control") and gene-based ("Gene") and master-transcript-based ("Master") adaptive sampling experiments. Adaptive sampling experiments are split into accepted and rejected pools. Median read lengths from the rejected pools are marked by a dotted line (**B**). *Q*-score of 13 (5% error rate) is marked by a dotted line (**C**)

reads/0.83 M total accepted reads, rep2: 0.083 M on-target accepted reads/1.23 M total accepted reads), equivalent to 3.16 × enrichment compared to control direct RNA sequencing, and a "base proportion" of 4.52% (rep1: 74.98 M on-target bases/1.69B total bases, rep2: 4.99 M on-target bases/2.43B total bases) (Fig. 1C, Additional file 1: Fig. S3, Table 1), equivalent to 1.90 × enrichment. This level of enrichment is comparable to previous reports using dRNA-AS [21, 22]. For instance, Naarman de Vries et al. reported a 1.35 × enrichment in on-target bases when targeting a single transcript in an equimolar mixture of two in vitro transcripts (IVT) [22]. The slightly higher enrichment found in our experiment may in part be explained by the significant difference in the initial proportion of targeted transcripts. Lastly, as a reference, we compared the performance of cDNA-AS and dRNA-AS to a hybridization capture-based experimental enrichment approach, TEQUILA-seq [14], which achieved 36.7 × enrichment in on-target bases (Fig. 1C).

To evaluate the consistency of ONT AS, we performed gene-based and master-transcript-based cDNA-AS on two additional cell lines, BT549 and T47D [26, 27], targeting 468 actionable cancer genes (Additional file 2: Table S4). Enrichment of 1.62–1.75 × in "accepted read proportion" and 1.30–1.34 × in "base proportion" was consistent with our previous findings (Additional file 1: Fig. S4, Table 1). These results demonstrate that the performance of cDNA-AS is not affected by different samples or target gene panels.

Direct RNA-AS, but not cDNA-AS, results in modest enrichment within a fixed run time

The enrichment values discussed so far are independent of total yield, as they represent the proportional increase in on-target output relative to overall sequencing output.

However, certain long-read sequencing applications may prioritize speed over final output. For instance, rapid detection of fusion transcripts or RNA modifications may be relevant for time-sensitive clinical samples. To address this, we investigated whether AS could increase the number of on-target reads within a fixed run time. ONT recommends increasing the volume of library loaded onto the flow cell for AS experiments to enhance sequencing speed, as the pores are expected to spend more time in the "open" state in AS experiments [16]. Following this recommendation, we performed three experiments using the following: (1) the protocol-standard cDNA input concentration of 20 fmol, (2) an increased cDNA input concentration of 40 fmol, and (3) direct RNA sequencing with an initial input of 3 µg (~6 pmol). To account for variability in the number of available pores across flow cells, we utilized a "split-format" approach, in which half of the channels performed gene-based AS while the other half performed whole-transcriptome cDNA sequencing or direct RNA sequencing on the SY5Y transcriptome, targeting 221 splicing factors (Fig. 1D). cDNA-AS, regardless of input library concentration, did not result in an increase in on-target base output compared to whole transcriptome cDNA sequencing (Fig. 1D, left). In contrast, direct RNA AS resulted in a modest increase (1.23 ×) in on-target base output (Fig. 1D, right). This increase aligns with a previous report of dRNA-AS, which found a 1.27 × increase in the number of bases mapped to a target transcript in a similar split-format approach on an in vitro transcript mixture [21].

Notably, both gene-based cDNA and dRNA AS resulted in reduced "total output within a fixed run time" (Additional file 1: Fig. S5), likely due to the time each pore spends rejecting reads and waiting in the "open" state until another strand is captured. This explains the lower enrichment performance in "on-target base output within a fixed run time," despite achieving moderate enrichment in "accepted read proportion" and "base proportion" (Fig. 1C, Additional file 1: Fig. S3, Table 1).

Adaptive sampling on the human transcriptome does not impact flow cell health

We aimed to determine whether AS leads to premature flow cell failure, as the frequent ejection of reads during AS may negatively impact pore health and reduce total yield. For applications where AS is employed to maximize the number of reads generated within a set budget, the lifespan of the flow cell should be considered. Using the pore_scan.txt output file located in the "other_reports" directory, we assessed the number of available pores on each half of the flow cell at the start and end of an 18-h (cDNA) or 72-h (direct RNA) sequencing experiment. We observed no difference in the rate of pore degradation (Fig. 3, "single pore" count).

Transcript length and read quality affect the end-decision accuracy during adaptive sampling

We next investigated the factors influencing the decisions made during AS, and their contribution to the final enrichment levels. First, we calculated the fraction of correctly rejected non-target reads during AS. AS correctly rejected 74% of non-target reads in gene-based dRNA-AS, 51% in master-transcript-based cDNA-AS, 38% in gene-based cDNA-AS, 27% in transcript-based cDNA-AS, and 12% in exon-based cDNA-AS (Fig. 2A). The percentage of the correctly rejected non-target reads tracked well with the overall performance of the AS experiments.

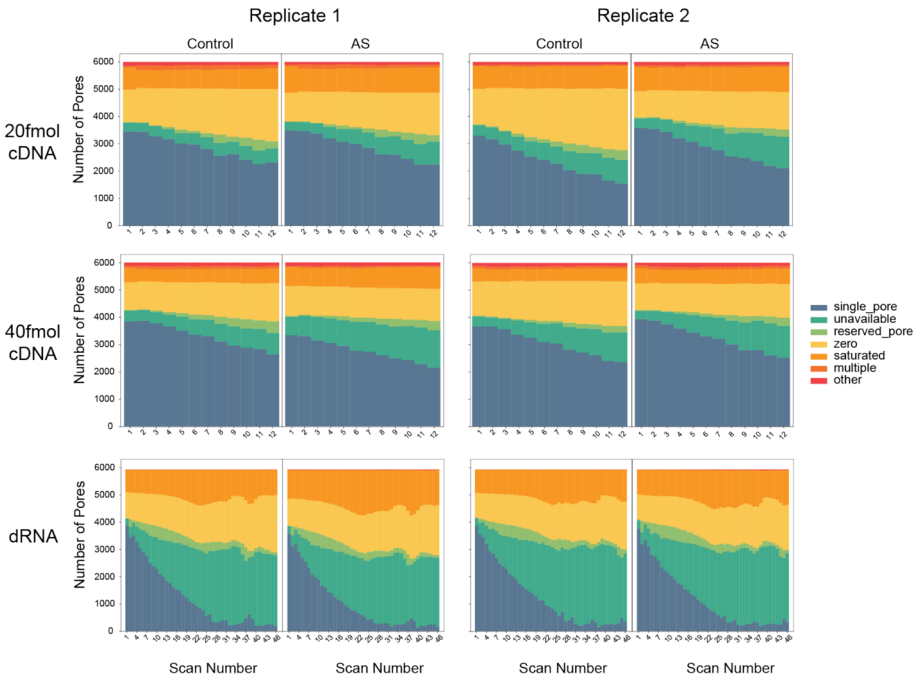

Fig. 3 State of all pores on a PromethION flow cell at each pore scan, which occurs every 90 min over the course of a sequencing experiment. On each flow cell, gene-based adaptive sampling ("AS") was performed on 1500 channels (6000 pores). The remaining 1500 channels performed standard whole transcriptome cDNA or direct RNA (dRNA) sequencing ("Control"). Healthy pores that are available for sequencing fall into the "single_pore" category. Pores that are irreversibly blocked, missing, or damaged fall into "unavailable," "zero," and "saturated" categories

We compared the read lengths from AS experiments to those from control experiments (Fig. 2B, Additional file 1: Fig. S6). The median read length was 736 nt from control cDNA experiments and 842 nt from control direct RNA experiments. In contrast, the median length of rejected reads, representing the amount of sequencing completed before a "rejection" decision could be executed, was 714 nt for gene-based cDNA-AS, 642 nt for master-transcript-based cDNA-AS, and 331 nt for gene-based dRNA-AS. The shorter length of rejected reads in dRNA-AS reflects the slower translocation rate of RNA through the nanopore. Additionally, 65% of wrongly accepted reads (i.e., non-target reads that were fully sequenced) were shorter than the median length required for rejection in gene-based cDNA-AS experiments, and 75% in master-transcript-based cDNA-AS experiments (Fig. 2B). This suggests that short read length accounted for the majority of wrongly accepted non-target reads. However, a notable fraction of non-target reads, particularly in gene-based dRNA-AS, spanning several kb, were fully sequenced and stored in the "accepted" bin (Fig. 2B, Additional file 1: Fig. S6A and S6C). This phenomenon cannot be explained by the "minimum decision time" requirement and remains unclear.

We then examined whether sequencing quality could affect end-decision accuracy during AS. Low quality on-target reads may be less likely to map correctly to the reference sequence in the provided FASTA file during AS, leading to their incorrect rejection. The Phred quality score (q-score) is a standard metric for sequencing quality [28], where higher q-scores indicate more confident base calls. A q-score of > 13 (equivalent to

a 5% error rate) is considered standard for a successful ONT long-read sequencing run [15, 29]. We plotted the *q*-scores for all experiments (Fig. 2C, Additional file 1: Fig. S6B and S6C). As expected, a significant proportion of the on-target but wrongly rejected reads had *q*-scores below 13 in gene- (57%) or master-transcript- (65%) based cDNA-AS, as well as in gene-based dRNA-AS (74%) (Fig. 2C). In general, on-target rejected reads exhibited lower *q*-scores compared to their accepted counterparts. In gene-based cDNA-AS experiments, the median q-score for on-target accepted reads was 17.2, compared to 11.6 for rejected reads. Similarly, in master-transcript-based cDNA-AS experiments, the median q-score was 14.8 for on-target accepted reads, compared to 10.0 for rejected reads. This trend was observed to a lesser extent in gene-based dRNA-AS, with the median q-score for on-target accepted reads at 13.6, compared to 12.1 for rejected reads.

Adaptive sampling does not bias on-target transcript abundance or affect the detection of novel transcripts

To assess if AS preserves the relative abundance of target transcripts, we added spike-in RNA variants (SIRV-Set 4) to SH-SY5Y RNA samples before cDNA synthesis and library construction. SIRV-Set 4 contains 96 ERCC transcripts with initial concentrations spanning six orders of magnitude and 15 equimolar long-SIRV transcripts up to 12 kb in length. We compared the detected abundance of each SIRV transcript to the respective ground-truth spike-in concentrations. The observed abundance correlated well with the initial concentration across all conditions (Additional file 1: Fig. S7A) and showed no significant length bias (Additional file 1: Fig. S7B). Additionally, the relative transcript abundances for the 211 splicing factors in the SH-SY5Y cell line during AS experiments all correlated well with whole transcriptome sequencing results (Additional file 1: Fig. S8). However, it is notable that most AS experiments showed minimal enrichment for on-target transcripts, as indicated by the overlapping regression lines for on-target and non-target transcript groups. Only master-transcript-based cDNA-AS and gene-based dRNA-AS experiments showed modest enrichment, with slight increases in on-target transcript abundance compared to non-target transcripts. These results contrast with the significantly enriched on-target transcripts using the hybridization-capture-based TEQUILA-seq method (Additional file 1: Fig. S7 and Fig. S8).

We then assessed the ability of AS to recover novel transcripts, particularly those with start or termination sites outside the annotated gene boundaries defined in the various FASTA file structures. ONT recommends including a buffer around regions of interest when creating reference FASTA files for AS of genomic DNA. Because gDNA is randomly fragmented, some fragments covering target loci may also contain adjacent non-target regions. Extending reference sequences into these surrounding regions helps prevent incorrect rejection of such reads. Although RNA sequencing does not involve random fragmentation, we considered the possibility that some reads might originate from novel transcripts starting well before the annotated start position or extending far beyond the annotated end position of a gene. Similarly, reads containing unannotated exonic regions may not align as well to the master-transcript-based FASTA file, which lacks these unannotated exons, potentially resulting in lower enrichment of these transcripts. To minimize the variation introduced by sequencing depth, data from SH-SY5Y,

BT549, and T47D cell lines were downsampled to an equal number of on-target reads in each condition. This step was necessary because samples with higher read depth have an increased likelihood of detecting novel transcripts. We then used ESPRESSO to identify and quantify annotated and novel transcript isoforms across control cDNA/dRNA sequencing, gene-based cDNA/dRNA-AS, master-transcript-based cDNA-AS, and TEQUILA-seq. No significant difference was observed in the number of identified novel transcripts across the sequencing conditions (Additional file 1: Fig. S9). Therefore, AS does not introduce bias in isoform composition within enriched sequencing data.

Discussion

ONT AS serves as a convenient method for targeted gDNA sequencing to enhance coverage of genes of interest. It has been shown to have robust enrichment on genomic DNA when the targeted fraction constitutes less than 10% of the total genome [16]. Applications of AS in RNA sequencing have primarily been evaluated using sequencing of native RNAs (direct RNA sequencing) [21, 22]. However, there is growing interest in applying AS for targeted cDNA sequencing due to the significantly higher yield achievable with ONT cDNA sequencing compared to direct RNA sequencing. Furthermore, the optimal choice of reference file format for AS on mammalian transcriptomes remains unclear, as mapping RNA sequencing reads to genome FASTA files presents unique challenges not encountered with gDNA reads. In this study, we systematically evaluated the performance of both cDNA- and dRNA-AS on the human transcriptome, exploring different parameters and readouts of targeted enrichment.

We note that the performance of AS is likely to be strongly influenced by computational parameters, including the structure of the reference FASTA file, the configuration of the aligner, and the speed of live basecalling. For a typical user, only the reference FASTA file can be directly controlled through the MinKNOW interface. Due to the lack of prior systematic evaluation, it was unclear what sequence structure is optimal for ONT AS applications in complex transcriptomes. To address this, we evaluated the performance of AS using four different reference FASTA file structures (Fig. 1B). The best performing reference sequence structures were as follows: (1) sequences covering the entire uninterrupted gene region (gene-based) or (2) sequences comprising all annotated exons merged into a single "master" transcript (master-transcript-based) (Fig. 1B). Conversely, the FASTA file containing separate sequences for all annotated exons of target genes (exon-based) showed poor performance, rejecting very few reads (Fig. 2A), likely due to challenges in aligning short individual exon sequences. Similarly, the file containing cDNA sequences for each annotated target transcript (transcript-based) also exhibited poor performance, potentially due to the large file size (Fig. 1B).

We performed cDNA-AS on three cell lines: SY5Y (a neuroblastoma line), BT549 (a mesenchymal-like breast cancer line), and T47D (an epithelial-like breast cancer line). These models were selected for their transcriptome complexity and relevance to distinct biological states. By spanning different cell lineages and epithelial-to-mesenchymal phenotypes, they serve as suitable representative systems for studying human transcriptomes.

We chose target gene panels of 221 human splicing factors or 468 actionable cancer genes, each of which has a moderate baseline gene expression level of 2–2.5% in the

human transcriptome (Additional file 2: Table S3). AS resulted in 1.30–1.34 × enrichment for on-target yield relative to total sequencing base output when using gene-based or master-transcript-based cDNA-AS (Sup. Figure 3, Table 1). However, it did not improve on-target yield within a fixed time frame compared to standard whole transcriptome cDNA sequencing (Fig. 1D), limiting its potential for resource savings. Consistent with previous reports [21, 22], direct RNA AS yielded higher enrichment (1.90 × enrichment relative to total sequencing output) (Sup. Figure 3, Table 1), likely owing to the slower translocation rate of RNA molecules through the nanopore. Despite producing a lower overall yield (Sup. Figure 5), dRNA-AS generated more data from target transcripts than control dRNA sequencing within a fixed run time (Fig. 1D).

AS correctly rejected up to 74% of non-target reads (gene-based dRNA-AS) when targeting 221 splicing factors, which represent 2% of the SY5Y transcriptome (Fig. 2A). This less-than-perfect non-target-read rejection rate contributes to the generally low enrichment observed in AS. For example, achieving a 5 × enrichment in accepted read proportion for a panel representing 2% of a library would require an over 82% rejection rate in the best case scenario (in which no on-target reads are wrongly rejected), while a 10 × enrichment would require an even higher rejection rate of over 92%. The enrichment ratios are further reduced when using the "base proportion," as the large number of reads rejected by the AS process still take up sequencing resources, i.e., the available pores in the flowcell.

To further understand the factors influencing AS performance, we assessed the read length and quality scores of reads for which incorrect decisions were made. We found that the majority of incorrectly accepted non-target reads were shorter than the median length required for rejection (Fig. 2B). As a result, these reads were fully sequenced before a rejection decision could be made. This rejection read length requirement explains the main reason for the lower performance of cDNA-AS compared to gDNA-AS and the relatively better performance of dRNA-AS. On the other hand, target reads with lower sequencing quality *q*-scores were more likely to be incorrectly rejected (Fig. 2C).

Rapid transcript enrichment via AS holds potential for applications where cost and time savings are critical. However, in its current state, the application of AS on large and complex transcriptomes, such as human, shows suboptimal performance. Capture-based enrichment approaches such as TEQUILA-seq achieve substantially higher fold enrichment and should remain the method of choice when time and hands-on experimental effort are not limiting factors. Nonetheless, for time-sensitive experiments or those requiring native RNA sequencing, such as RNA modification analysis, dRNA-AS may still offer a valuable alternative. While dRNA-AS offers only modest enrichment (1.23 ×) of target transcripts over a set time period, unlike hybridization capture, it can be easily automated and requires no additional sample handling beyond uploading a reference FASTA file.

Recently developed third-party technologies such as RISER and PROFIT-seq [31, 32], which can be implemented through MinKNOW's ReadUntil interface, achieve slightly better enrichment compared to the built-in AS feature implemented through MinKNOW. For example, RISER, a third-party selective sequencing application designed specifically for direct RNA sequencing, makes rejection decisions faster than ONT AS.

However, it has only been rigorously assessed for depletion of highly abundant transcripts, rather than targeted enrichment of specific transcripts of interest. Similarly, PROFIT-seq specifically addresses the obstacles introduced by the short read length of cDNA compared to gDNA by incorporating rolling circle amplification to generate long concatemeric reads, allowing for more time for each off-target strand to be rejected. Additionally, without requiring oligo(dT)-based reverse transcription, PROFIT-seq can target non-polyadenylated transcripts, unlike many standard enrichment methods. Although they each perform moderately better than AS, both methods result in $< 2\times$ enrichment. Furthermore, for PROFIT-seq, the modest saving of sequencing resources through selective sequencing may be offset by the substantially increased time and experimental complexity introduced during library preparation.

Our analysis suggests that ONT AS is not currently an optimal enrichment strategy for target enrichment of complex transcriptomes. Through a systematic evaluation of reference FASTA file structure, we find that smaller, less redundant reference files with individual sequences that encompass the entire length of target transcripts (e.g., gene-based or master-transcript-based) enable more efficient alignment and contribute to better enrichment. We anticipate that continued advances in live basecalling and alignment speed by ONT or third-party developers could gradually improve the performance and utility of ONT AS for targeted enrichment of complex transcriptomes in specific applications.

Conclusions

This study systematically evaluates ONT AS for human transcriptome analysis and compares it to TEQUILA-seq, a hybridization-capture approach. While ONT AS enables modest enrichment of target genes in human transcriptome analysis, it is much less effective than hybridization-capture (Fig. 1C). Direct RNA-based AS may be suitable for rapid transcript analysis in time-sensitive clinical applications. As the modest enrichment of AS is mainly due to short off-target read lengths and low sequencing quality of on-target reads, enhancing live basecalling speed and computational analysis may be necessary to improve AS for targeted long-read RNA sequencing.

Methods

Cell culture

SH-SY5Y human neuroblastoma cells (ATCC, CRL-2266) were cultured in DMEM/F-12 (Gibco, 11330032) with 10% fetal bovine serum (FBS, Corning, MT35010CV) and 100 U/ml penicillin–streptomycin (Gibco, 15140122) at 37 °C in 5% CO_2. The SH-SY5Y cell line was verified to be mycoplasma-free by the Lonza MycoAlert assay and authenticated by short tandem repeat analysis. BT459 and T47D breast cancer cell lines were purchased from the American Type Culture Collection (ATCC, Manassas, VA, USA 30–4500 K™) and cultured according to ATCC recommendations in RPMI-1640 Medium (Gibco, 11875093) with 10% fetal bovine serum (FBS, Corning, MT35010CV), 0.023 units/mL human insulin (BT549) (Sigma-Aldrich, I2643), or 0.2 units/mL bovine insulin (T47D) (Cell Applications, 128100) at 37 °C in 5% CO_2. The breast cancer cell lines were verified to be mycoplasma-free by the Lonza MycoAlert assay and authenticated using short tandem repeat analysis by the supplier.

RNA extraction, cDNA synthesis, and TEQUILA-seq

Total RNA was extracted from SH-SY5Y, BT549, and T47D cell lines using TRIzol reagent (Invitrogen, 15596018). Spike-in RNA variants (SIRV) were added to total SH-SY5Y RNA at a ratio of 1:1200. RNA concentrations were measured using a NanoDrop 2000 Spectrophotometer, and RNA integrity was examined on an Agilent 4200 TapeStation.

For whole transcriptome cDNA sequencing, reverse transcription was performed on 300 ng of total SH-SY5Y RNA with 250 pg SIRV-Set 4 spike-in RNA (Lexogen, 141.01) using Maxima H minus reverse transcriptase (Thermo Fisher, EP0752), dNTP (Invitrogen, 18427088), template-switching oligo (TSO) (IDT), and RNase inhibitor (Promega, N2615) at 42 °C for 90 min, followed by enzymatic inactivation at 85 °C for 5 min. PCR amplification was performed with KAPA HiFi HotStart ReadyMix (Fisher Scientific, 50–196–5217) using the following program: 95 °C initial denaturing for 3 min, followed by 12 cycles of 98 °C for 20 s, 67 °C for 20 s, 72 °C for 5 min, and a final extension at 72 °C for 8 min. Amplified cDNA was purified using 0.7X SPRIselect beads (Beckman Coulter, B23318).

TEQUILA probes were synthesized and cDNA transcripts were captured as previously described [14].

All oligos are listed in Additional file 2: Table S5.

Nanopore library construction and sequencing

Nanopore direct RNA libraries were prepared from 3600 ng of total RNA from SY5Y cells with 3 ng of spike-in RNAs following the standard SQK-RNA004 protocol. First-strand cDNA was synthesized by incubating RNA with T4 DNA Ligase (2 M U/ml—NEB, M0202), Quick Ligation Reaction Buffer (NEB, B6058), RNAseOUT (Invitrogen, 10777019), and the supplied RT Adapter at room temperature for 10 min, then at 50 °C for 50 min with 10 mM dNTP mix (Thermo Scientific, 18427013), SuperScript III Reverse Transcriptase and the associated 5 × First-strand buffer (Invitrogen, 18080093), and the supplied DTT, followed by a 10-min enzymatic inactivation at 70 °C. The resulting RNA–DNA hybrid was purified with 1.8X Agencourt RNAClean XP beads (Beckman Coulter, A63987). The supplied RNA Ligation Adapter was ligated at room temperature with NEBNext Quick Ligation Reaction Buffer (NEB, E6056) and T4 DNA Ligase (2 M U/ml—NEB, M0202) for 10 min, then purified with 0.4 × Agencourt RNAClean XP beads (Beckman Coulter, A63987) and supplied Wash Buffer, then eluted with the supplied RNA Elution Buffer. Libraries were immediately loaded onto RNA PromethION flow cells (FLO-PRO004RA) and sequenced on a P2 Solo.

Nanopore whole transcriptome libraries were prepared from 200 ng of amplified cDNA following the standard ONT SQK-LSK114 protocol. The cDNA products underwent end-repair and dA-tailing with the NEBNext Ultra II End Repair/dA-Tailing Module (NEB, E7546) by incubating at 20 °C for 5 min and 65 °C for 5 min. The cDNA was then purified with 1X AMPure XP beads (Beckman Coulter, A63881). Adapter ligation was carried out at room temperature for 10 min using NEBNext Quick T4 DNA ligase (NEB, E6056). The ligated libraries were purified with 0.4 × volumes of AMPure XP beads (Beckman Coulter, A63881) and the supplied Ligation Adapter and Short Fragment Buffer, then eluted with the supplied Elution Buffer. Replicates for each SY5Y + SIRV-Set 4 experiment (cDNA control, gene-based cDNA-AS,

master-transcript-based cDNA-AS, transcript-based cDNA-AS, exon-based cDNA-AS, and TEQUILA-seq), as well as BT549/T47D experiments (cDNA control, gene-based cDNA-AS, and master-transcript-based cDNA-AS), were performed on a single R10.4.1 PromethION flow cell (FLO-PRO114M) on a P2 Solo or P24. Flow cells were washed in between replicates using the flow cell wash kit (ONT, EXP-WSH004). BT549 and T47D TEQUILA-seq experiments were sequenced on R9.4.1 flow cells (FLO-MIN106) on a GridION. Unless otherwise specified, 20 fmol of each cDNA library was loaded onto the flow cell.

For adaptive sampling experiments, reference FASTA files were uploaded and selected under the "Enrich or deplete sequences" in the "Adaptive Sampling" section of the experiment setup within MinKNOW. When indicated, adaptive sampling was performed only on channels 1–1500, while channels 1501–3000 performed standard 1D cDNA or direct RNA sequencing.

Detailed sequencing summary statistics, including library information, chemistry, flow cell types, and sequencing output, can be found in Additional file 2: Table S3.

Construction of reference FASTA files for adaptive sampling

FASTA files were generated from a reference GTF using scripts provided at https://github.com/Lin-rnalab/ONT-adaptive-sampling [33, 34]. Briefly, the GTF file was filtered for the relevant annotation type (gene, transcript, or exon) in target genes. For exon-based adaptive sampling, duplicate exons (i.e., those with identical start and stop coordinates as an exon in another transcript of the same gene) were removed. For master-transcript-based adaptive sampling, all exons for a given gene were concatenated into a single sequence. These modified GTF files were then converted into FASTA format.

Basecalling and alignment of nanopore sequencing data

Basecalling of raw nanopore data was performed using super-high accuracy dorado (v.0.6.0) with the following settings: --emit-fastq --device cuda:all. The "dna_r10.4.1_e8.2_400bps_sup@v4.3.0" model was used for basecalling of cDNA sequencing data, and "rna004_130bps_sup@v3.0.1" was used for basecalling of direct RNA data. Where adaptive sampling was performed on only a portion of channels on a flow cell, pod5 files were split according to channel number before basecalling was performed. Basecalled reads were aligned to either the GRCh38/hg38 primary reference genome or to the SIRV-Set 4 reference genome using minimap2 (v2.25) with parameters: "-ax splice -ub---secondary=no -t 8 -k 14 -w 4." A bed file containing annotated splice junctions was created using minimap2's "paftools.js gff2bed" tool and supplied to the aligner with "--junc-bed."

Processing of adaptive sampling data

End decisions, target status, lengths, and Phred qualities were extracted for each read in adaptive sampling and non-adaptive sampling experiments. For adaptive sampling experiments, reads that were denoted "data_service_unblock_mux_change" under "end_reason" in the sequencing_summary.txt file were categorized as rejected, while those denoted "signal_positive" were classified as accepted. All other reads were categorized as "other" and not used for downstream analysis. On-target reads

were identified with `samtools view -F 3844 {BAM} -L {BED}`, where the BED file contained on-target gene coordinates. Non-target reads were classified as those that map to the genome (i.e., part of the subset identified from `samtools view -F 3844 {BAM}`) but were not part of the on-target subset. Unmapped reads were those which were not part of the mapped subset. Read lengths were extracted from the sequence line (line 2) of the corresponding entry in the FASTQ file generated after basecalling. Phred qualities for each base were extracted from the quality line (line 4) and converted to an error probability with $P = 10^{-Q/10}$, where P is the error probability and Q is the quality score of that base. The average error probability across all bases in a read was then converted back to a Phred quality with $Q = -10 * \log10(P)$ to calculate the overall Phred quality of the read.

Analysis script is provided at https://github.com/Lin-rnalab/ONT-adaptive-sampling [33, 34].

Transcript identification and quantification

Novel and annotated transcripts were quantified using ESPRESSO (v1.4.0) guided by a gene annotation GTF file using default parameters [30].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03813-1.

Additional file 1: Supplementary Figures S1-S9. Fig S1: Pore activity during whole transcriptome and adaptive sampling cDNA sequencing experiments on SH-SY5Y cells. Fig S2: Top 1000 most highly expressed genes in SH-SY5Y cells detected using different methods. Fig S3: Proportion of bases mapped to target genes in SH-SY5Y cells using different methods. Fig S4: Proportion of reads and bases mapped to target genes in BT549 and T47D breast cancer cells using different methods. Fig S5: Total base output in side-by-side "split-format" experiments of whole transcriptome and adaptive sampling on SH-SY5Y cells. Fig S6: Read lengths and Phred quality scores for experiments on SH-SY5Y cells using different methods. Fig S7: Transcript detection and quantification using different methods. Fig S8: Pairwise comparisons of estimated abundances for transcript isoforms of target and non-target genes in SH-SY5Y cells using different methods. Fig S9: Number of target gene isoforms detected in SH-SY5Y, BT549, and T47D cells using different methods.

Additional file 2: Supplementary Tables S1-S5. Table S1: Panel of 221 genes encoding human splicing factors. Table S2: Contents of SIRV-Set 4 modules. Table S3: Summary statistics for all sequencing experiments. Table S4: Panel of 468 actionable cancer genes. Table S5: List of oligonucleotides, primers, and probe sequences.

Acknowledgements
We thank Dr. Yi Xing for assistance with the manuscript.

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
L.L. and N.D. conceived and designed the study; N.D. generated and analyzed the results; F.W. contributed sample and generated data; Y.X. contributed methods; N.D. and L.L. wrote the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by National Institutes of Health grants R01GM121827 and R35GM158057.

Data availability
The data generated during the current study are available in FASTQ file format in the Sequence Read Archive at https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA1169789 [35]. Processed data are available in the Gene Expression Omnibus at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse307605 [36]. All analysis scripts are provided at https://github.com/Lin-rnalab/ONT-adaptive-sampling. A snapshot of this repository has also been archived with a DOI via Zenodo to ensure long-term accessibility: https://doi.org/10.5281/zenodo.15732007. The published scripts are licensed under the GPLv3 open source license.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
L.L. and F.W. are named inventors on a patent application (application number: PCT/US22/79537, in process) filed by The Children's Hospital of Philadelphia covering the TEQUILA method. This patent will not affect the use of the TEQUILA method for reproducing the results in this manuscript. The other authors declare no competing interests.

Received: 25 March 2025 Accepted: 24 September 2025
Published online: 09 October 2025

References

1. Park E, Pan Z, Zhang Z, Lin L, Xing Y. The expanding landscape of alternative splicing variation in human populations. Am J Hum Genet. 2018;102(1):11–26.
2. Chu C, Borges-Monroy R, Viswanadham VV, Lee S, Li H, Lee EA, et al. Comprehensive identification of transposable element insertions using multiple sequencing technologies. Nat Commun. 2021;12(1):3836.
3. Heyer EE, Deveson IW, Wooi D, Selinger CI, Lyons RJ, Hayes VM, et al. Diagnosis of fusion genes using targeted RNA sequencing. Nat Commun. 2019;10(1):1388.
4. Norkin M, Ordóñez-Morán P, Huelsken J. High-content, targeted RNA-seq screening in organoids for drug discovery in colorectal cancer. Cell Rep. 2021;35(3):109026.
5. Tourancheau A, Margaillan G, Rouleau M, Gilbert I, Villeneuve L, Lévesque E, et al. Unravelling the transcriptomic landscape of the major phase II UDP-glucuronosyltransferase drug metabolizing pathway using targeted RNA sequencing. Pharmacogenomics J. 2016;16(1):60–70.
6. Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, et al. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat Biotechnol. 2011;30(1):99–104.
7. Bussotti G, Leonardi T, Clark MB, Mercer TR, Crawford J, Malquori L, et al. Improved definition of the mouse transcriptome via targeted RNA sequencing. Genome Res. 2016;26(5):705–16.
8. Hook PW, Timp W. Beyond assembly: the increasing flexibility of single-molecule sequencing technology. Nat Rev Genet. 2023;24(9):627–41.
9. Dillon LW, Hayati S, Roloff GW, Tunc I, Pirooznia M, Mitrofanova A, et al. Targeted RNA-sequencing for the quantification of measurable residual disease in acute myeloid leukemia. Haematologica. 2019;104(2):297–304.
10. Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, et al. Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. Genome Biol. 2016;4(17):41.
11. Hardigan AA, Roberts BS, Moore DE, Ramaker RC, Jones AL, Myers RM. CRISPR/Cas9-targeted removal of unwanted sequences from small-RNA sequencing libraries. Nucleic Acids Res. 2019;47(14):e84.
12. Mercer TR, Clark MB, Crawford J, Brunck ME, Gerhardt DJ, Taft RJ, et al. Targeted sequencing for gene discovery and quantification using RNA Captureseq. Nat Protoc. 2014;9(5):989–1009.
13. Davis CP, West JA. Purification of specific chromatin regions using oligonucleotides: capture hybridization analysis of RNA targets (CHART). Methods Mol Biol. 2015;1262:167–82.
14. Wang F, Xu Y, Wang R, Zhang B, Smith N, Notaro A, et al. TEQUILA-seq: a versatile and low-cost method for targeted long-read RNA sequencing. Nat Commun. 2023;14(1):4760.
15. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021;39(11):1348–65.
16. Adaptive sampling | Oxford Nanopore Technologies. [cited 2025 Feb 6]. Available from: https://nanoporetech.com/document/adaptive-sampling.
17. Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. Nat Biotechnol. 2021;39(4):442–50.
18. Martin S, Heavens D, Lan Y, Horsfield S, Clark MD, Leggett RM. Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. Genome Biol. 2022;23(1):11.
19. De Meulenaere K, Cuypers WL, Gauglitz JM, Guetens P, Rosanas-Urgell A, Laukens K, et al. Selective whole-genome sequencing of Plasmodium parasites directly from blood samples by nanopore adaptive sampling. mBio. 2024;15(1):e0196723.
20. Miller DE, Sulovari A, Wang T, Loucks H, Hoekzema K, Munson KM, et al. Targeted long-read sequencing identifies missing disease-causing variation. Am J Hum Genet. 2021;108(8):1436–49.
21. Wang J, Yang L, Cheng A, Tham CY, Tan W, Darmawan J, et al. Direct RNA sequencing coupled with adaptive sampling enriches RNAs of interest in the transcriptome. Nat Commun. 2024;15(1):481.
22. Naarmann-de Vries IS, Gjerga E, Gandor CLA, Dieterich C. Adaptive sampling for nanopore direct RNA-sequencing. RNA. 2023;29(12):1939–49.
23. GENCODE - Human Release 44. [cited 2025 Feb 6]. Available from: https://www.gencodegenes.org/human/release_44.html.
24. Ulrich JU, Epping L, Pilz T, Walther B, Stingl K, Semmler T, et al. Nanopore adaptive sampling effectively enriches bacterial plasmids. mSystems. 2024;9(3):e0094523.

25. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, et al. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 2015;15(2):141–61.
26. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. J Mol Diagn. 2015;17(3):251–64.
27. Fiala EM, Jayakumaran G, Mauguen A, Kennedy JA, Bouvier N, Kemel Y, et al. Prospective pan-cancer germline testing using MSK-IMPACT informs clinical translation in 751 patients with pediatric solid tumors. Nat Cancer. 2021;2:357–65.
28. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The sanger FASTQ file format for sequences with quality scores, and the solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010;38(6):1767–71.
29. Ni Y, Liu X, Simeneh ZM, Yang M, Li R. Benchmarking of Nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. Comput Struct Biotechnol J. 2023;21:2352–64.
30. Gao Y, Wang F, Wang R, Kutschera E, Xu Y, Xie S, et al. ESPRESSO: Robust discovery and quantification of transcript isoforms from error-prone long-read RNA-seq data. Sci Adv. 2023;9(3):eabq5072.
31. Sneddon A, Ravindran A, Shanmuganandam S, Kanchi M, Hein N, Jiang S, et al. Biochemical-free enrichment or depletion of RNA classes in real-time during direct RNA sequencing with RISER. Nat Commun. 2024;15(1):4422.
32. Zhang J, Hou L, Ma L, Cai Z, Ye S, Liu Y, et al. Real-time and programmable transcriptome sequencing with PROFIT-seq. Nat Cell Biol. 2024;26(12):2183–94.
33. DeBruyne N. ONT-adaptive-sampling. Github. 2025. https://github.com/Lin-rnalab/ONT-adaptive-sampling.
34. DeBruyne N. 2025. ONT AS data processing Zenodo. 2025. https://doi.org/10.5281/zenodo.15675045.
35. DeBruyne N. Nanopore cDNA and direct RNA sequencing with adaptive sampling on SY5Y, BT549, and T47D cell lines. Sequence Read Archive. 2025. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1169789.
36. DeBruyne N. Evaluating the potential and limitations of nanopore adaptive sampling for targeted transcriptome sequencing. Datasets. Gene Expression Omnibus. 2025. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse307605.

Publisher's Note