Direct RNA sequencing (RNA004) allows for improved transcriptome assessment and near real-time tracking of methylation for medical applications

Charlotte Hewel[1], Felix Hofmann[1], Vincent Dietrich[1], Anna Wierczeiko[1], Johannes Friedrich[1], Kristi Jenson[2], Stefan Mündnich[3], Stefan Diederich[1], Stanislav Sys[1], Lukas Schartel[4,5], Susann Schweiger, Mark Helm[3], Edward A Lemke[4,6], Matthias Linke[1,$], Susanne Gerber[1,$*]

[1]Institute for Human Genetics, University Medical Center of the Johannes Gutenberg University Mainz, Germany

[2]Institute of Molecular Biology (IMB) Mainz, Germany

[3] Institute of Pharmaceutical and Biomedical Sciences, Johannes Gutenberg-University Mainz, 55128 Mainz, Germany

[4]Biocenter, Johannes Gutenberg-University, Hanns-Dieter-Hüsch-Weg 17, 55128 Mainz, Germany

[5]International Max Planck Research School on Cellular Biophysics, Hanns-Dieter-Hüsch-Weg 17, 55128 Mainz, Germany

[6]Institute of Molecular Physiology, Johannes Gutenberg University Mainz, Germany

[$]Joint senior authors

[*]To whom correspondence should be addressed

Abstract

Direct RNA sequencing (dRNA-seq) is a nanopore-based technique to sequence RNA in its native form without enzymatic conversion steps. In general, dRNA-seq holds great promise to resolve uncharted areas of human disease diagnostics and biomarker development. However - apart from a small core community - broader usage and acceptance of the method have been impeded by several methodological hindrances, such as low throughput, low accuracy, and missing large-scale modified base-calling models. Recently, Oxford Nanopore Technologies (ONT) has released a new chemistry, a new nanopore (RNA004), and new base-calling options (m6A). Here, we present the first results on two standardized test samples: UHRR (Universal Human Reference RNA) and HEK293T. Furthermore, we demonstrate current approaches for the detection of RNA modification, most notably m6A, from human blood samples as well as in an artificial RNA modification system allowing for targeted pseudouridinylation of specific RNAs in a “life-like” context. The new chemistry has significantly improved throughput and accuracy and can support real-time tracking of state-specific methylation information. From this perspective, we provide an outlook on the extent to which this RNA flow cell is suitable for bringing RNA diagnostics and the study of RNA modification disorders (RNA modopathies) into a clinical application as a routine measure.

Introduction

Classical RNA-seq is firmly established in molecular and diagnostics laboratories. The method works via transcription of RNA into complementary DNA (cDNA) and finally sequencing of small fragments.

This type of sequencing lends itself well to differential expression analysis of genes or transcripts in a high throughput setting (Soneson, Love, and Robinson 2015).

However, the conversion from RNA into DNA and the subsequent fragmentation also erases a sizable quantity of information present on native RNA. The method hinders the direct observation of both full-length transcripts and RNA modifications. Especially modifications on RNA are vastly more complex than on DNA with multitudinous levels to exert regulatory function in the living organism. Classical RNA-seq is hardly suited to sequence either full length isoforms or RNA modifications (Alfonzo et al. 2021).

A hallmark discovery describing how RNA modifications interact with a biological system has been the suppression of recognition by toll-like receptors, as part of the innate immune system. Meaning if an RNA molecule carries sufficient modifications such as pseudouridine (Ψ), the immune system will not be alerted and the immunogenicity of the transcript is reduced (Karikó et al. 2005). RNA modifications play major roles in the ageing process (e.g. in the context of autophagy) and have been found to modulate the expression level and stability of mRNA in cells (Jonkhout et al. 2017; Destefanis et al. 2021; Delaunay et al. 2024). Modification dependent immune modulation was one of the mechanisms crucial for the development of modern and effective mRNA vaccines against Covid-19 with markedly reduced immunogenic shock reactions and enhanced translation (Sahin et al. 2020).

Perhaps the most well researched and best described mRNA modification is N6- methyladenosine (m6A). In fact, papers from the 1970s already describe a modification of RNA nucleosides and speculate this modification may exert a possible biological function (Desrosiers, Friderici, and Rottman 1974). More recent papers have followed up with a more fine-grained view of transcriptional control. As an example in 2018 Yue et al. performed knockout experiments and subsequent sequencing on multiple m6A writer enzymes. The experiment could confirm that m6A downregulation has an effect on not only the general m6A abundance but also on the polyA-length, the motif-related m6A deposition and the expression level of genes in a condition and enzyme specific manner.

However, it should be noted that the modification experiments in Yue et al., had to be confirmed via laborious and time intensive multiple rounds of sequencing and modification specific assays. A view that is also prominently featured in review papers of RNA modifications. More precisely, the authors note that while there have been advances in the detection of RNA modifications, the situation is far from being resolved and the fact remains that specific assays on the level of single modifications have to be employed, when using classical approaches (Motorin and Helm 2022; Helm and Motorin 2017).

Thus, the field would benefit from a method that can detect multiple modifications at once with a standard library prep. Direct RNA sequencing using nanopores is generally regarded as a potential solution to this problem. The method is able to directly observe nucleotide specific modifications as variations of the electrical current in real time (Garalde et al. 2018). Thus, nanopore sequencing in this context would allow for gene and transcript level quantification, similar to classical RNA-seq with the added bonus of longer reads and the ability to read multiple modifications in real time and all simply via a standard sequencing library prep. This approach would be technically superior to classical methods and enable research and diagnostics on complex epitranscriptomics as a routine measure and in a high-throughput fashion.

Oxford Nanopore Technologies (ONT) has previously released a chemistry (RNA002) to facilitate direct RNA sequencing. This chemistry, while yielding results, especially in the detection of modifications for tRNA did garner mixed reviews (Lucas et al. 2024). Major points of concern were low throughput, low accuracy and the fact that modification basecalling was not enabled by default (Soneson et al. 2019; Jain et al. 2022).

To improve on this ONT has released a new iteration of RNA-sequencing chemistry called RNA004. This version of the chemistry features a new nanopore, a new motorprotein and new basecalling

models for both standard nucleotides and basepair modifications (m6A). Here, we present first results of the new RNA004 chemistry. The samples selected for sequencing were the Universal Human Reference RNA (UHRR), HEK293T cells carrying a custom vector and peripheral blood as commonly used sample in clinical practice.

Results

Sequencing yield and sample characteristics

A total of nine flow cells using either RNA002 or RNA004 compatible settings were sequenced on a PromethION P24 device. Of these nine biosamples, four were generated from the HEK293T cell line carrying a custom vector, three samples were obtained from the Universal Human Reference RNA (UHRR), and two samples were derived from the peripheral blood of a healthy control subject (see Fig. 1A). The latter two samples include sequencing of RNA extracted from whole blood and a copy of the same sample obtained by in vitro transcription (IVT) (see Tab. 1). The number of reads produced ranges from ca. 3 Mio for the IVT sample to ca. 20 Mio for the polyA selected HEK293T runs. The overall yield clearly segregates after chemistry type (RNA002 or RNA004), library preparation method (polyA-selected or not), and biosample (standardized cell line or peripheral blood). The highest yield has been observed in the polyA selected HEK293T runs on RNA004 chemistry with approximately 15 GB and 21 GB yield and above 20 Mio reads each. After that, the second total RNA run for the UHRR on RNA004 follows with approximately 9 Mio reads and 6 GB. The two runs for HEK293T on RNA002 chemistry follow with approximately 5 GB and 5-6 Mio reads. Moreover, we have the first run of UHRR with RNA004 chemistry stopping at around 4.4 Mio reads and 4 GB output. This run had a flow cell failure during the testing stage and can not be seen as representative. Nevertheless, we have decided to include this run in order to show an unbiased overview of all the data from all the test runs. Finally, the human blood samples were sequenced using RNA004 chemistry once natively and once after IVT conversion. Here, 5 GB/5 Mio reads and 2 GB/3 Mio reads were achieved, respectively.

The longest N50 of the unaligned data was reported by the first RNA004 run of UHRR with 1.367 kB. Regarding the RNA002 sequencing for the HEK293T samples and UHRR we observed N50s with around 1.3 kB as well. With an N50 of around 1 kB, the first run of the HEK293T (HEK293_A), the second run of the UHRR (RNA004_UHRR_2) and the IVT sample (RNA004_blood_IVT), are all shorter than their counterpart. 5‘-3‘ genebody coverage can be found in the Supplementary Figure (Fig. S1).

Another focus point of the new chemistry is the increase in accuracy. We approached this feature via two commonly used quality metrics. The first metric employed is the phred-based quality encoding produced by the basecaller after an estimated function and stored as quality information in the fourth line of the FASTQ file. This quality estimate is primarily useful to roughly filter the data in the pre-alignment or pre-assembly stage, or in cases where no highquality reference genome of the sample is available. The second quality metric used was consequently the percent sequence accuracy. The percent sequence accuracy is a percent estimate in relation to mismatches/matches compared to a reference sequence (here: GRCh38). The runs mostly separate by chemistries in both conditions (see Fig. 1B). Counterintuitively, the older chemistries have a higher phred-based quality score than the samples sequenced with RNA004 chemistry. When examining the percent sequence accuracy the situation is reversed and we find about 92 % accuracy for the RNA002 cell line runs and about 96 % sequence accuracy for the RNA004 cell line runs (see Fig. 1C). The RNA004 peripheral blood sample scores lower, with 95.3 % sequence accuracy for the natively sequenced RNA, and 93.5 % sequence accuracy when the sample is IVT converted.

Next, we determined whether the global and chemistry-specific increase in quality can be linked to regions that are difficult to sequence, which pertains to homopolymeric sequences, as frequent error sources for nanopore sequencing on the DNA level at least. As can be seen from Fig. 1D &E, there seems to be indeed a chemistry-specific resolution of homopolymers, that worsens with length and

shows the characteristic deletion pattern already known from DNA (Fig. S2). Additionally, we evaluated the polyA-length of the transcripts DDX17, SRP14 and OLA1, who can roughly be classified into long, middle and short polyA-tailed transcripts respectively (Fig. 1F-H).

Feature counts and transcript discovery are further classical metrics evaluated in RNA-seq. Here, the new RNA004 chemistry shows improvement by virtue of throughput. As evident from Figure 2A and 2B, the counts for novel and full-length transcripts for HEK293T increase in a chemistry specific manner. The same holds true for the features (Gencode v43 GRCh38) discovered at 1x and 10x coverage respectively (see Fig. 2C, D). For clinical applications the overlap with a so called mendeliome, meaning all genes clearly associated with disease, is of special interest. Here, we find that ~ 90-50 % of genes are covered, depending on sample and condition (Supplementary Fig. S3). Moreover, we can see that deeper sequencing would have likely yielded some additional features to be present. When examining specific notoriously difficult genes like XIST (32,068 bp) in a clinical sample it is obvious that mostly the 3' end is covered as opposed to shorter single exon genes like the mitochondrial ND3 (346 bp) (Fig. 2F).

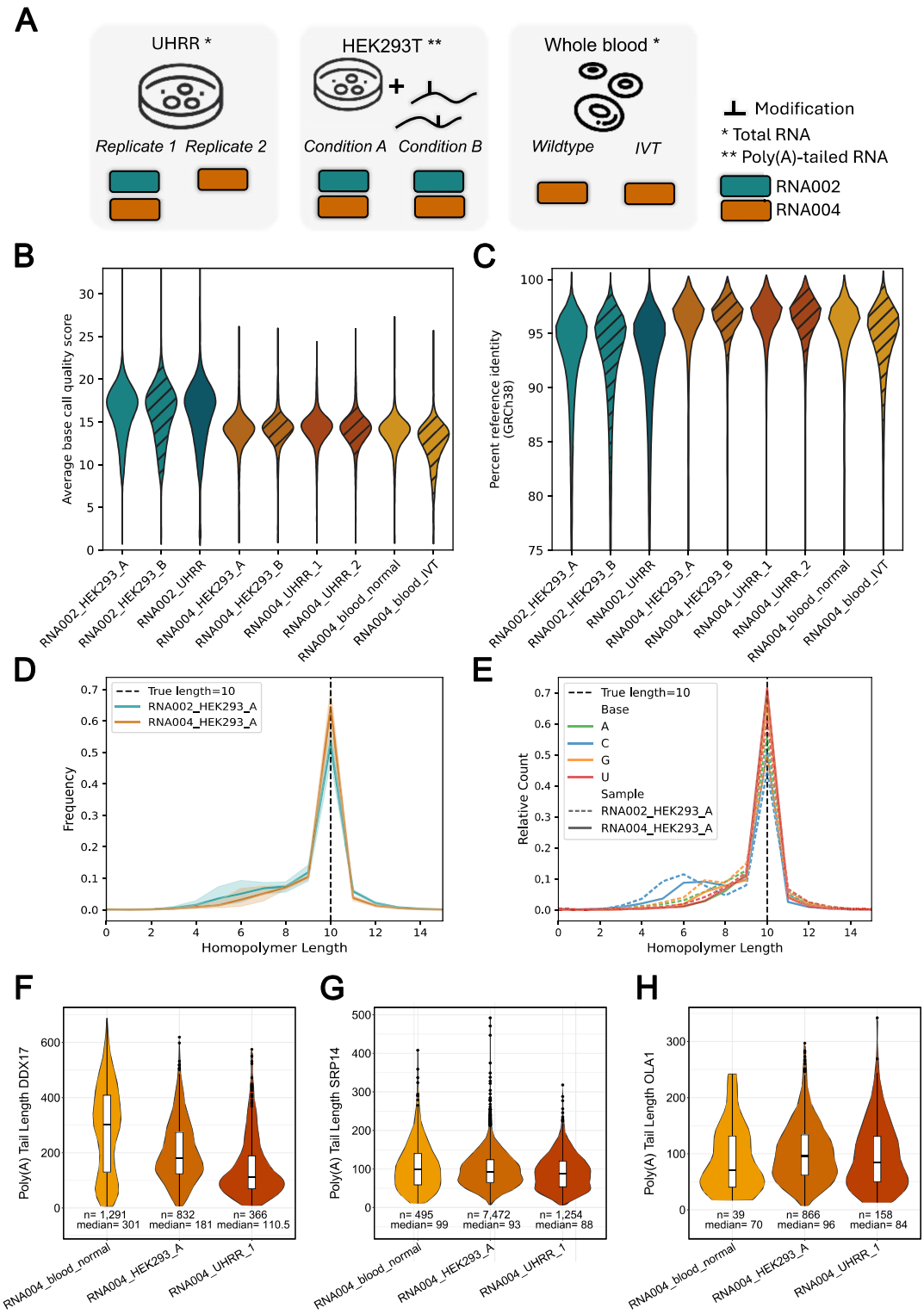
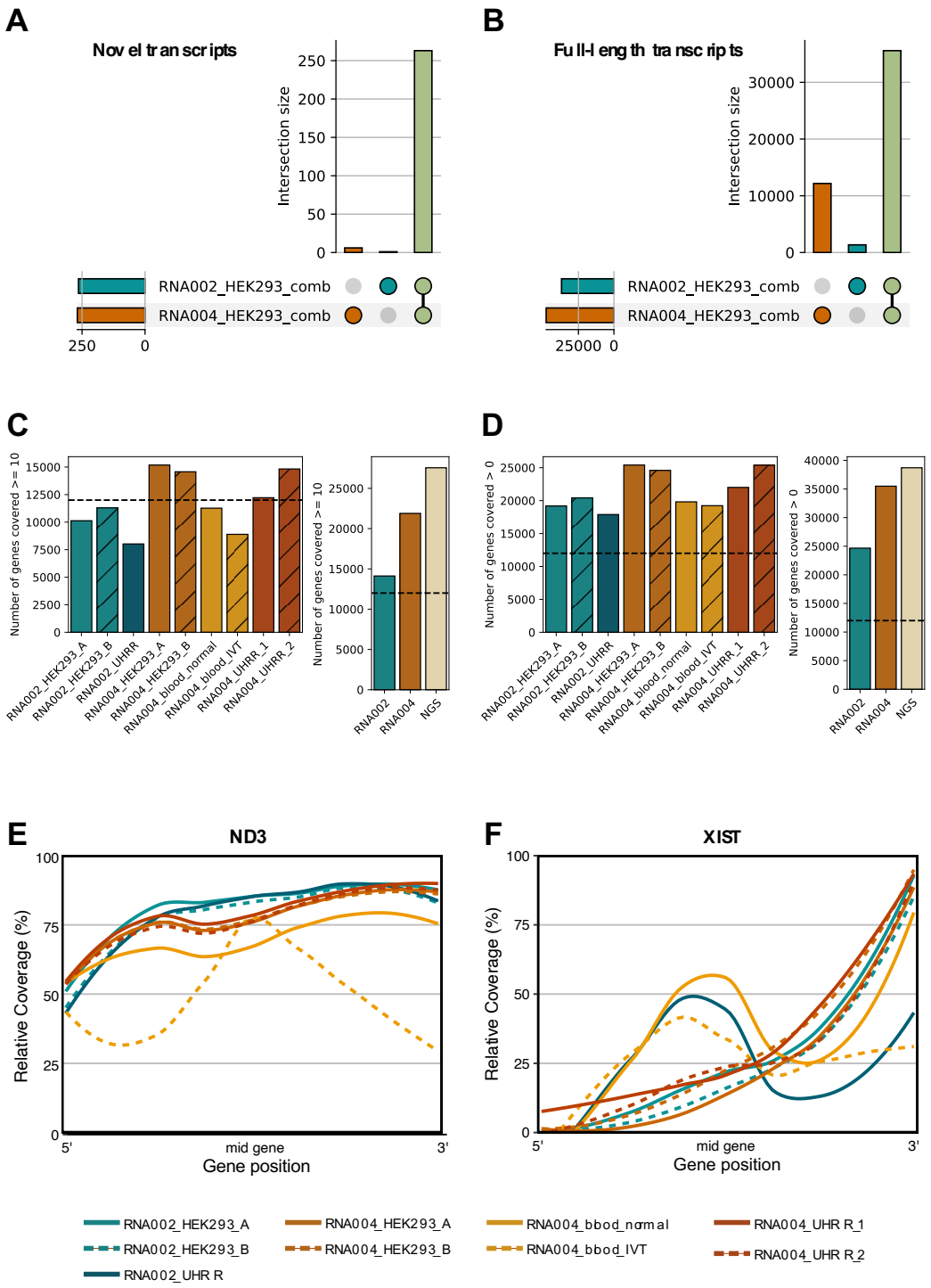


Figure 1 *General sequencing quality metrics for all samples (A) sample type overview and number of replicates sequenced, (B) violin plot of the average basecall quality score for all samples as estimated by dorado, (C) violin plot of the percent reference identity for all samples as measured against comparision with GRCh38, (D) observed vs expected homopolymer length for all homopolymers of length 10. The dashed black line represents the homopolymer length expected in the reference and the colored lines represent the homopolymer lengths observed for either chemistry RNA002 or RNA004. (E) homopolymer length as in D just split per nucleobase additionally, (F-H) violin plots of the polyA-tail length as measured by dorado for the genes DDX17, OLA1 and SRP14.*



1 *Sequencing feature characteristics (A) Upset plot displaying the number of shared and exclusive novel transcripts identified with both kits and the HEK293T samples. (B) Upset plot displaying the number of shared and exclusive full-length transcripts identified with RNA002 and RNA004 (HEK293T samples). (C) Left: Number of genes identified from RNA002 and RNA004 sequencing runs. Right: Aggregated gene counts over all samples from RNA002 and RNA004, as well as Illumina next generation sequencing. Horizontal line indicates 12000 reads. (D) Left: Number of genes identified with >10 coverage. Right: Aggregated counts of genes with a coverage >10 for RNA002 and RNA004 over all sequencing runs, and Illumina next generation sequencing. Horizontal line = 12000 reads. (E) Relative gene body coverage of ND3. (F) Relative gene body coverage of XIST.*

RNA modification detection in close to real time using chemistry RNA004

Next, we quantified a unique feature of nanopore direct RNA sequencing: the ability to detect RNA modifications natively and in a close to real time fashion. We want to highlight the general use case and the improvements due to the new chemistry on the basis of two use cases: 1) the detection of Ψ in a custom vector and 2) the global detection of m6A with a productionready basecalling model (provided by ONT basecallers) in almost real-time.

The first case relates to an artifical RNA modification system that can be used to enable efficient and mRNA selective pseudouridination. More precisely, in HEK293T cells the vectors eGFP or mCherry are targeted and modified with Ψ at a single position only. To verify the efficiency of this modification applied on mRNA level, we performed direct RNA sequencing and (due to the lack of an existing basecall model for Ψ) observed the Ψ -specific misbasecall pattern (U>C mismatch) at eGFP and mCherry for RNA002 and RNA004, respectively (Fig. 3A). The basic pattern, that a misbasecall occurs more frequently at the targeted position, could already be verified with the previous RNA002 chemistry. However, it is clearly visible that the misbasecall became considerably less noisy and more accurate using RNA004 chemistry. In Fig. 3B the SNP basecall at the targeted position is listed over time, demonstrating that the new chemistry reached a constant plateau of identification in about half the time and with greater accuracy. This is likely transferable to other applications requiring SNP-misbasecall patterns.

For the second use case, we examined m6A in a sample taken from the peripheral blood of a healthy control person. Peripheral blood is the preferred type of sample in clinical practice and routine diagnostics. For m6A ONT has already released a production ready basecall model. This means a high classification accuracy and minimal increase on total basecaller run time (see Supplementary table Tab. S1), which is crucial for „near-real time“ classifications in any type of setting. The blood sample itself was sequenced once natively and once via IVT which removes any pre-existing modifications and can be seen as a proxy for an extreme case of methylation disturbance. M6A is a frequent modification of mRNA. As has been shown by orthogonal methods of m6A labeling, m6A enriches near the STOP codon of biological transcripts, producing a distinct peak. It was possible to distinguish between the two conditions almost immediately (Fig. 3C-F). Even when the gene count seen at least once was minimal and the pattern could be tracked over time in a step-wise manner, using routine analysis tools.

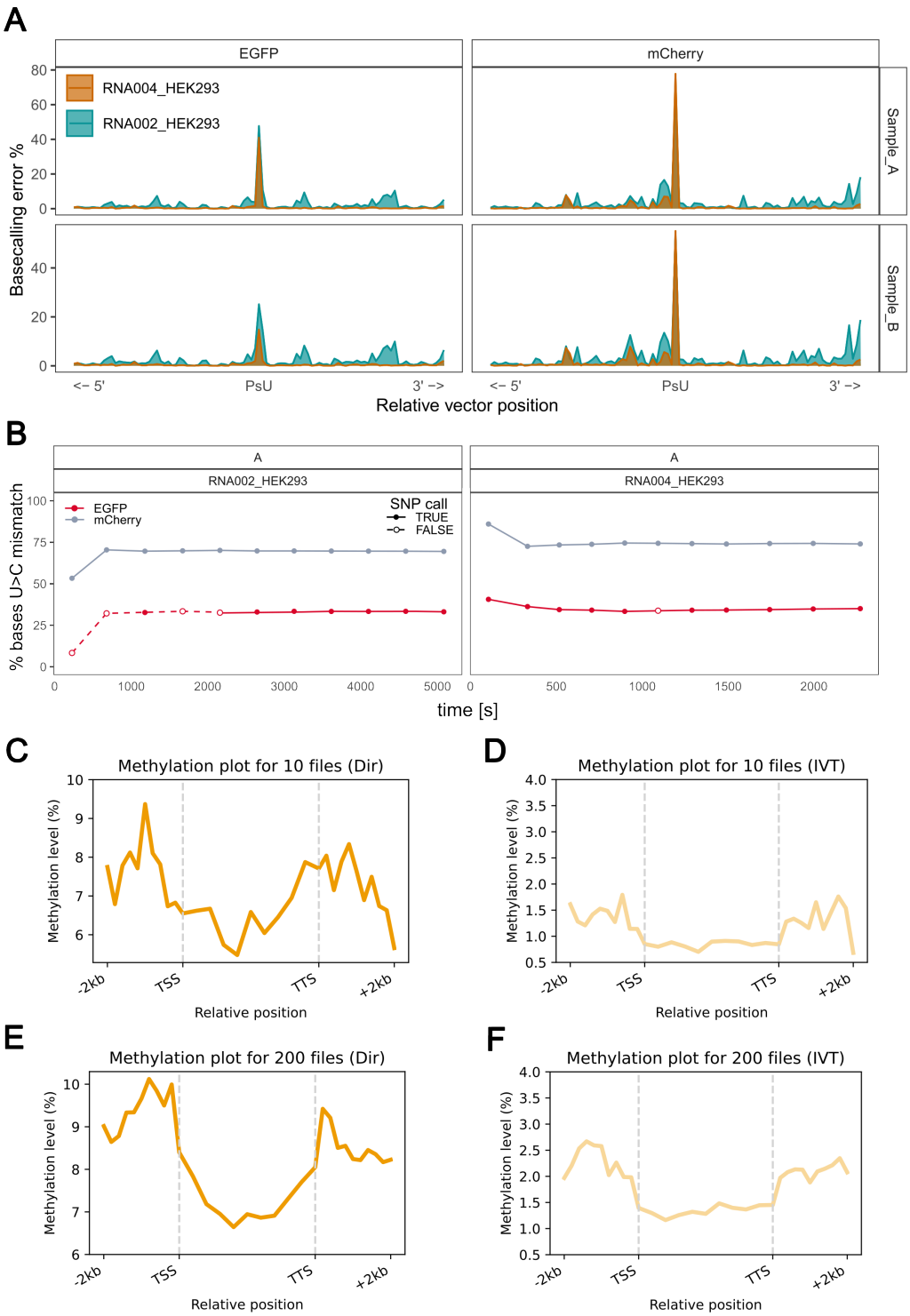


Figure 3 *Real-time methylation pattern estimation for either Ψ or m6A (A) misbasecall pattern across the entire vector sequence for either mCherry or eGFP for chemistry RNA002 or chemistry RNA004 respectively, (B) misbasecall pattern at the exact vector position where pseudourdination is expected over time and for chemistry RNA002 and chemistry RNA004 (C-F) m6A methylation percentage across genes for directly sequenced native RNA (Dir) and in-vitro transcribed RNA (IVT) from peripheral blood. Displayed metagene plots aggregate methylation levels for sequences up to 2kb upstream of the Transcription Start Site (TSS), the genebodies and sequences up to 2kb downstream of the Transcription Termination Site (TTS). Plots were created in near real-time after every 10 files output by MinKNOW, m6A methylation patterns are displayed for native RNA after 10 files (C) and 200 files (E) and IVT after 10 files (D) and 200 files (F). The methylation in the IVT sample is likely residual in nature and/or technical classification error.*

Lastly, the general methylation information can be combined with gene counts or other data on the fly to create real-time multi-omics in a single sample almost instantaneously. As an example, we observed differential m6A-site expression between HEK293T and UHRR (Fig. 4A). A global upregulation of m6A in HEK293T can be seen. This can now be cross-referenced against a list of all known m6A writer or eraser enzymes at the same time (see Supplementary Figure;Fig. S4), and the expression of ALKBH5 matched this conditional increase of methylation (Fig. 4B), without the need to sequence an additional library or to obtain more patient sample.

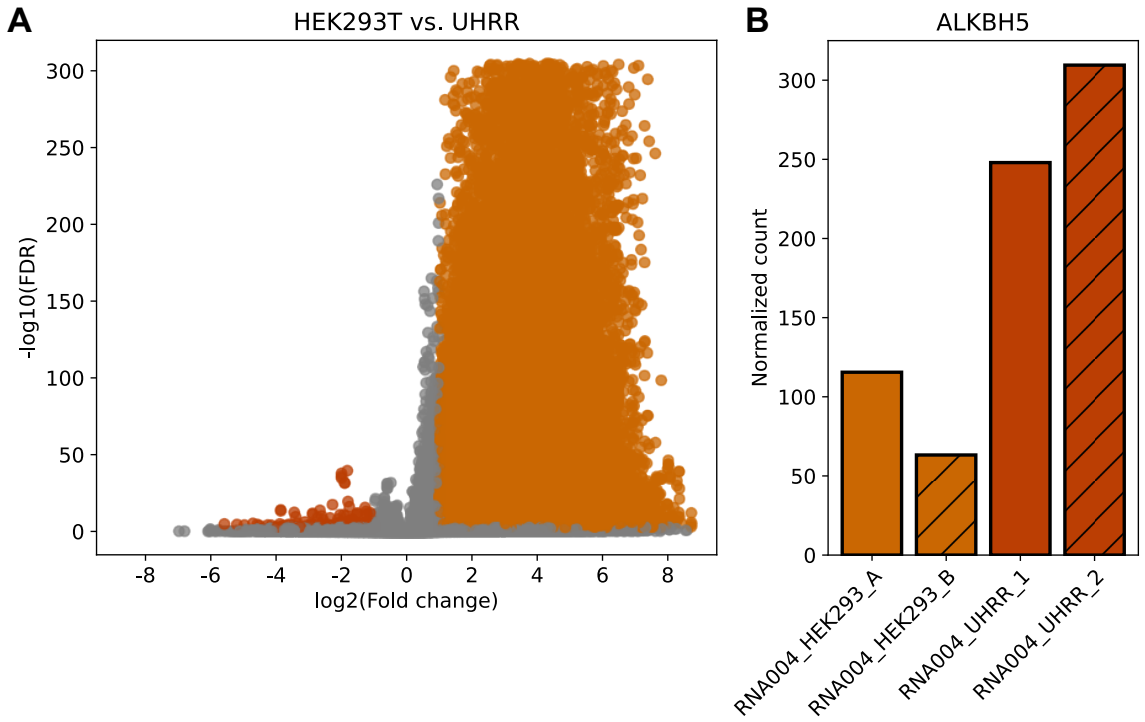


Figure 4 *Differential methylation and gene counts. (A) Volcano plot of differentially methylated loci (DML) between HEK293T and UHRR samples. Colored DML have a log2-fold change >1 or <-1 and a FDR<0.001. (B) Normalized counts of ALKBH5 RNA demethylase reads in HEK293T and UHRR samples.*

Table 1: General sample statistics

General summary	RNA002 HEK293_A	RNA002 HEK293_B	RNA002 UHRR	RNA004 HEK293_A	RNA004 HEK293_B	RNA004 UHRR_1	RNA004 UHRR_2	RNA004 blood_normal	RNA004 blood_IVT
Mean read quality	12.1	11.9	11.9	11.5	11.9	11.5	11.5	11.2	8.3
Median read quality	16.7	16.5	17.5	13.9	14.1	14.1	13.9	13.8	12.3
Number of reads	5057042	6266231	3328536	20497439	20011109	4444850	9144701	5195237	3015475
Read length N50	1343	1132	1306	964	1324	1367	940	1300	1101
Total bases unaligned	5004920908	5132882299	3011392130	15707971873	21067640688	4502746514	6714962654	5105642860	2161631790

Discussion

Nanopore direct RNA sequencing has matured considerably. For comparision, the first highquality large-scale paper of direct RNA sequencing in 2019 for a human polyA selected cell line used 30 MinION flow cells to generate 9.9 million aligned sequence reads with a median identity of 86 %, a maximum aligned length of over 21000 bases read length and the effort was spread over six different institutions (Workman et al. 2019). Both test PromethION flow cells for polyA selected HEK293T surpassed this metric individually, providing more throughput with higher quality at a fraction of the previous cost and effort.

The quality for standard basecalling models for a model organism is currently reported to be at around 96 %, which we could confirm. It is of note however that specialized condition specific basecalling models have achieved an accuracy of 97 % even for the old chemistry (Diensthuber et al. 2023).

The new chemistry has indeed shown considerable improvement on quality throughput and usability, but not all points of concern were removed. A concise list has been compiled by (Jain et al. 2022) and we would just peruse it to discuss the points in light of both the older RNA002 chemistry and the newer RNA004 chemistry.

Full length transcripts respectively the mismatch between annotated transcript length and the fraction of reads covering the annotated 5' end is still apparent. While we might argue that sheer throughput may alleviate this issue to some extent, longer complex transcripts such as XIST still remain mostly 3' end covered. Intertwined with this issue is the fact that the motor protein translocating the RNA through the pore does release the 5' end eventually, because the molecule has ended. Previous papers have shown that this leaves the last few nucleotides to be unsequenceable (Workman et al. 2019). This is also one of the reasons why state-ofthe- art transcript detection tools such as bambu struggle to correctly predict the 5' end, even on "full-length" direct RNA (Chen et al. 2023). However, special sequencing adapter library preparation concepts in addition to modifications of the read detection algorithm of MinKNOW can alleviate incomplete/missing transcripts to some point (Ugolini et al. 2022; Lucas et al. 2024).

Routine validation of RNA modification calls has also been on the wishlist. While the general concept has been proven early on, there are excellent community written tools (Stoiber et al. 2016; Liu et al. 2019; Parker et al. 2020; Jenjaroenpun et al. 2021; Pratanwanich et al. 2021; Furlan et al. 2021; Hassani et al. 2022; Abebe et al. 2022; Wan et al. 2022; Wierczeiko et al. 2023; Zhong et al. 2023; Maestri et al. 2024) .The basecallers provided by ONT such as Guppy and Dorado have not been able to directly support modification calling in real time and for high-throughput configurations.

There is no classical RNA-based method that can i) detect more than a few selected RNA modifications or ii) distinguish between different RNA modifications in a sequence where multiple types coexist. However, these distinctions are precisely the technological advances needed to gain a fundamental understanding of the regulatory role of a complex universe of RNA modifications and their effects on aging and disease. Especially for clinical applications in RNA diagnostics, reliable and easy-to-use RNA modification detection options integrated into the standard software are required. Even though several new classifiers are currently being developed and published by the community and are leading to ever better and more reliable predictions (Cruciani et al. 2023; Teng et al. 2024) the best strategy for ONT would still be to also provide a user-friendly solution for clinical demands or questions from fundamental research without requiring end users to rely exclusively on community developments.

Here, significant progress has been made with RNA004 and the new m6A model from ONT. We are aware that there is a number of new m6A callers such as xron or m6ABasecaller, that might produce different results, but testing them entirely would have been outside the scope of this paper (Teng et al. 2024; Cruciani et al. 2023). For the standard model, we could confirm that PromethION level throughput for m6A calling can be supported directly in what we would consider to be a mock experiment for later use in routine clinical practice. Peripheral blood is the most frequent type of

sample in routine diagnostics and less invasive than a culture made from fibroblasts for example. With this, we hope to provide an interesting sample to the community for the new RNA004 chemistry besides pure cell line samples.

M6A in human samples is a modification that is ideally suited for evaluation with nanopore devices. The modifier is present on over 7000 human genes, with a preferred location near the STOP codon (Dominissini et al. 2012). Standard dRNA-seq ligates an adapter to the 3' end only. This means, that the 3' end is statistically more likely to be sequenced than the 5' end. Similarly, modifications that are close to the 3' end, such as m6A, are covered well. Additionally, the high frequency of occurrence means that meaningful metagene information is available after the first batches of reads, as we could demonstrate during the sequencing of the peripheral blood sample in this paper. Likewise, the mismatch pattern was stable for the Ψ in HEK293T samples after approximately 500 seconds (RNA004) and the rest of the sequencing period would technically have been not necessary. For future applications this signifies that barcoding, read-until sequencing (when the nanopore device stops automatically once condition is fullfilled) or a combination of both can be used to reduce either sequencing time or cost, depending on the limiting factor.

This opens up very intriguing possibilities for bench to bedside translational research. As has been proven with the DNA chemistries and cancer research, nanopore devices can detect methylation differences, copy number deviations, fusion genes and cancer gene hallmark mutations. Combined with the real time data output of a nanopore sequencer same-daydiagnostics are feasible, even rapid assessment within hours during surgery, overall reducing total turnaround time and improving patient outcomes (Euskirchen et al. 2017; Djirackor et al. 2021; Patel et al. 2022; Vermeulen et al. 2023). With the new chemistry RNA004 similar research becomes feasible for routine RNA diagnostics applications. Theoretically, modifications can be measured at the same time as gene counts, smaller RNA pieces, structural abberations, SNP calls and the future may even bring targeted resequencing of the same molecule. Nanopore sequencing has the possibility to reverse the current. This has been used to resequence one molecule multiple times, enhancing quality even if the molecule is ultra rare in the suspension. For example in 2022, Zhang et al. ligated a streptavidin stopper at the 5' end of DNA molecules. The stopper prevented the molecule from fully passing to the trans side of the flow cell, a reversal of the current allowed the molecule to be resequenced. On top of that, one may recover the sample from the cis side of the flow cell via a library recovery protocol. These features make the technology very attractive for multi-omics studies and biomarker discovery of either the cell free fraction or inside of extracellular vesicles, additionally Alzheimer research or the epigenetics of aging on the mRNA level (Jonkhout et al. 2017; Jörg et al. 2024). We might even argue that nanopore based dRNA-seq - as opposed to classical RNA seq and follow-up experiments - is the most ethical choice in an animal model system for some instances. For example when direct RNA sequencing is planned but multi-omics would be an option for the future if either the pure differential expression was inconclusive or if it pointed to a methylation disturbance.

Taken together, the new chemistry is much improved and directly useable for clinical research and real-time applications. The larger user-base that is expected to follow will hopefully lead to more support for still missing modification models, diagnostics grade computational pipelines or barcoding.

Material and Methods

Sample description

Three different sample sources were employed during this manuscript and sequenced a total of nine times respectively. The Universal Human Reference RNA (UHRR) was purchased from Thermo Fisher Scientific, catalogue number QS0639. The human biosample was taken from a healthy volunteer after written informed consent was obtained.

HEK293T cells pseudouridine

HEK293T cells were transfected with artifical snoRNAs as well as eGFP and mCherry. The snoRNAs have been designed to target a premature stop codon within the eGFP and mCherry transcripts at nucleotide position 115 and 556 respectivly. In Case A both mRNAs are targeted for

pseudouridinylation where as in Case B the mCherry mRNA is preferably targeted owing for a decrease in eGFP pseudouridinylation (Schartel et al., 2024).

Direct RNA library preparation for the cell line samples and RNA002/RNA004 chemistries

For direct RNA library preparation, we utilized either the old chemistry Direct RNA sequencing chemistry (ONT, SQK-RNA002) or the updated direct RNA sequencing chemistry (ONT, SQK-RNA004) following the manufacturer's protocol. Briefly, 100 ng of poly(A)-tailed RNA or 1000 ng of total RNA was adjusted to 9 µl with nuclease-free water. To this RNA sample, 3 µl of NEBNext Quick Ligation Reaction Buffer (NEB, B6058), 1 µl RT Adapter (RTA) (ONT), and 1.5 µl T4 DNA Ligase 2M U/ml (NEB, M0202) were added, resulting into a total volume of 15 µl. The reaction was mixed by pipetting and incubated for 10 minutes at room temperature. Next, the reverse transcription master mix was prepared by mixing 9 µl nucleasefree water, 2 µl 10 mM dNTPs, 8 µl 5x first-strand buffer (Thermo Fisher Scientific), and 4 µl 0.1 M DTT. This master mix was added to the RNA sample containing the RT adapter-ligated RNA along with 2 µl SuperScript III reverse transcriptase. The reaction was incubated at 50 C for 50 minutes, followed by 70 C for 10 minutes, and then brought to 4 C. 72 µl RNAClean XP beads (Beckman Coulter, A63987) were then added to the reaction, followed by incubation on a Hula mixer for 5 minutes at room temperature. Subsequently, the sample underwent two washes with 70% ethanol, and the DNA was eluted with 20 µl nuclease-free water. This eluted DNA was used in the adapter ligation reaction. For the adapter ligation reaction, 8 µl NEBNext Quick Ligation Reaction Buffer, 6 µl RNA Adapter (RMX for RNA002; RLA for RNA004), 3 µl nuclease-free water, and 3 µl T4 DNA Ligase were mixed with the 20 µl eluted DNA to a total volume of 40 µl. The reaction was incubated for 10 minutes at room temperature. After incubation, 20 µl RNAClean XP beads were added to the adapter ligation reaction, followed by incubation on a Hula mixer for 5 minutes at room temperature. The sample was then subjected to two washes with the Wash Buffer (WSB, ONT) using a magnetic rack. Following the washes, the pellet was resuspended in 41 µl (RNA002) or 33 µl (RNA004) of Elution Buffer (EB) and incubated at 37 C for 10 minutes in a Hula mixer to release long fragments from the beads. Finally, the eluate was cleared by pelleting the beads on a magnet, and the eluate was retained and transferred to a clean 1.5 ml tube. 1 µl of reverse-transcribed and adapted RNA was quantified using the Qubit fluorometer. For R9.4.1 PromethION sequencing (RNA002), 40 µl of the library was mixed with 35 µl nuclease-free water and 75 µl RRB and loaded into a R9.4.1 PromethION flow cell. For RNA chemistry PromethION sequencing (RNA004), 32 µl of library was mixed with 100 µl Sequencing Buffer (SB) and 68 µl Library Solution (LIS) and loaded into RNA chemistry PromethION flow cell.

IVT

The peripheral blood was obtained from a healthy volunteer. The RNA was extracted using the PAXgene Blood miRNA Kit from Qiagen according to the Manufacturer´s protocol, except the RNA was eluted in nuclease free water instead of the provided buffer. The RNA was characterized using the Bioanalyzer total RNA Nano Assay according to the Manufacturer´s protocol. The RNA had a concentration of 363ng/µl and a RIN value of 7,2. A globin depletion was performed with the GLOBINclear™-Human Kit from Thermofisher Scientific (AM1980) according to the Manufacturer´s protocol. The globin depletion was carried out four times. The total input amount of RNA of the globin depletion was 20µg, the total output amount was 11µg globin depleted RNA. The concentration was measured using the Qubit RNA HS Assay from Thermofisher Scientific. 2µg of RNA was stored for later use in the direct RNA Run. 9µg of RNA was taken forward to the Poly(A) selection using NEBNext Poly(A) mRNA Magnetic Isolation Module according to the Manufacturer´s protocol. The Poly(A) enrichment was carried out three times. The total output of the Poly(A) enrichment was 23ng of mRNA measured with the same Bioanalyzer assay listed above. The sample had an average length of ~1kb. The sample concentration was measured again using the same Qubit assay listed above. The following reverse transcription (RT), PCR, in vitro transcription (IVT), polyadenylation and 5`capping was carried out according to the paper of Tavakoli et al. (2023). Following individual deviations were made: The IVT Primers used in the PCR had a final concentration in the reaction of 0,5µM per primer. The input amount used for the RT and PCR was 7,1ng of mRNA. The output amount was 905ng of cDNA, measured with the Qubit DNA HS Assay from Thermofisher Scientific. The IVT was carried out twice, each with an input of 126,7ng cDNA. The output was pooled, the final amount was 4,9µg RNA measured with the Qubit RNA HS assay. The library preparation was

performed using the direct RNA sequencing Kit (SQK-RNA004) from Oxford Nanopore Technologies. The library output was 167ng of RNA/cDNA hybrid measured with the Qubit DNA HS Assay. The library was loaded completely onto the PromethION RNA Flow Cell (FLO-PRO004RA).

Basecalling and alignment of RNA002 and RNA004 runs

The raw pod5 files from all five RNA004 sequencing runs were basecalled using Dorado v0.4.3 with the model rna004_130bps_sup@v3.0.1**.** The model allowed for direct calling of m6A modifications in the DRACH context using flag *--modified-bases m6A_DRACH.* Poly(A) tail lengths were also estimated by including flag *--estimate-poly-a*, after the tailfindr algorithm that was recently adopted by ONT. The basecalling of raw pod5 files from the four RNA002 sequencing runs was done with Dorado‘s high accuracy model for RNA002 namely rn002_70bps_hac@v3. Basecalled reads of all samples were then aligned to the primary assembly of the human reference genome hg38, downloaded from Gencode release 43 (*GRCh38.primary_assembly.genome.fa.gz*). Alignment was performed with Minimap2 v2.24 with the following settings: *-y --MD -ax splice -uf -k14.* The resulting BAM files were sorted and indexed using samtools v1.16.1. The HEK293 samples A and B were additionally mapped onto the EGFP and mCherry reference sequences for target modification detection analysis. The quality metrics of all sequencing runs and mappings were derived by NanoComp v1.23.1. The average base call quality, the alignment-based percent identity and the N50 read length were visualized in python3.8 using matplotlib v.3.8.3 and seaborn v.0.13.2.

Illumina Data

Comparative data from Illumina were obtained from the Sequencing Quality Consortium (SEQC)[cite/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4321899/]. We downloaded the Universal Human Reference RNA (UHRR) Illumina HiSeq 2000 subset of the study and processed the data according to best practices (GSE47774). For experimental details, we refer to SEQC.

Metagene analysis

The *pileup* command from ONT Modkit v0.2.4 was used to extract the methylation-specific reads from the alignments, . The metagene plots visualizing m6A density were created with Loreme v0.2.4 *plot-genes.* Plots were smoothed by adding flag *--smooth* to the command. Basic gene annotations for hg38 were downloaded from Gencode release 43 (*gencode.v43.basic.annotation.gff3*).

Annotation of genomic features

Reads were mapped to genomic features using featureCounts v.2.0.0. Basic gene annotation downloaded from Gencode v43 served as the annotation reference (*gencode.v43.basic.annotation.gff3*). Parameter *–L* was passed to featureCounts to account for long reads as input; *–s 0* to perform unstranded read counting. The format of the annotation file was explicitly specified with *-F 'GTF'.*This option is applicable for GTF and GFF.

Pseudouridine detection analysis on target positions

We took advantage of the property that pseudouridine modifications generate a mismatch from uridine to cytosine (U>C) during basecalling in direct RNA sequencing reads (Begik et al., 2021). After filtering for primary aligned reads by using the parameter *-F 3488* in *samtools view*, the position-specific basecalling error profiles of the vector-carrying samples were extracted including mismatches, deletions, and insertions for the EGFP and mCherry sequences using the *mpileup* function provided by samtools v1.16.1. *Mpileup* was executed with the parameters *-x –no-output-ins –no-output-del –no-output-ends*. The percentages of basecalling error patterns at each position were plotted for both biological samples sequenced with the RNA002 and RNA004 flowcells. To calculate the significance of U>C mismatches of different read depths, we applied SNP calling on filtered bam files underlying different numbers of reads (4000 reads * {1,10,20,...,100}) aligned to the target sequences EGFP and mCherry by using bcftools v1.18. The bcftools functions *mpileup* and *calls* were executed using the recommended parameters for Nanopore long-read sequencing *–-config ont* and *–-prior 0.01*, respectively. The „U>C mutation rate“ of significantly detected mutations was plotted for all RNA002 and RNA004 condition-specific samples. All plots were created in R v4.2.2 using the R package ggplot2
v3.4.4.

Homopolymer length

The distributions of the observed homopolymer lengths were obtained with Counterr v.0.1 using the default settings. The hg38 reference sequence from Gencode release 43 was given for the *-genome* flag. The returned numpy objects *hist_len_hp.npy* were visualized using matplotlib v.3.8.3 and seaborn v.0.13.2.

Transcript discovery

Novel and full-length transcripts discovery was performed with Bambu v.3.2.4. Available replicates of mapped sequencing data were merged into single BAM files using the *merge* function provided by samtools v.1.16.1. The merged BAM files were given to Bambu using hg38 reference sequences and gene annotations in GTF format obtained from Gencode v45 for each comparison. Bambu was run with default parameters, extracting novel and full-length transcript datasets. From the Bambu output, Venn diagrams and upset plots were created using matplotlib_venn v.0.11.10 and UpSetPlot v.0.9.0, respectively.

Differential methylation analysis

Differentially methylated loci (DML) between the two replicates from the UHRR and HEK293T samples were identified using DSS v.2.50.1 contained in Bioconductor release 3.18. The sequencing data containing m6A calls were prepared using the *pileup* function in modkit v.0.2.5. The resulting bedMethyl tables were subset to include only the genomic position, the valid coverage (N valid_cov), and the number of modified calls (Nmod) and split by forward and reverse strand for clear overlapping with the genome annotations in later steps. Sites with a coverage below five were removed. The DSS function DMLtest was run with smoothing enabled, followed by callDML with a p-value threshold of 1 to retain all sites for later filtering. Basic gene annotations from Gencode v43 were added to the resulting dataset using the intersect function in bedtools v.2.30.0. Data from both strands was merged and significantly differentially methylated loci were categorized based on a FDR threshold of 0.001 and an absolute log2-fold change larger than 1. Volcano plots were created with matplotlib v.2.2.3.

Counts of writer and eraser genes

Normalized counts of specific RNA methylation writer and eraser genes between UHRR and HEK293T samples were obtained with the Python implementation of DESeq2 v.0.4.7 using the featureCounts table described before as input. The normalized counts were subset to contain only those corresponding to the following genes: METTL16, METTL3, METTL14, WTAP, ZC3H13, RBM15, KIAA1429, FTO and ALKBH5. Counts were plotted using matplotlib v.3.8.3 and seaborn v.0.13.2.

Basecall time measurements

Unless otherwise noted, basecall time was measured via „time basecall_command.sh“ on an NVIDIA DGX-Station A100.

Mendeliome Counts

featureCounts v2.0.0 tables for genes > 0 or genes > 10 reads coverage were intersected with an in-house list of mendeliome genes using bedtools v2.27.1. Plots were generated in R using ggplot2 v3.4.4.

Gene Body coverage analysis of single genes

The gene body coverages for the genes XIST and MT-ND3 were extracted by filtering the aligned bam files for reads that map onto the respective gene regions using *samtools view*. The genebody_coverage.py function from RSeQC v5.0.2 was then used to extract the percentage of reads covering gene body regions divded into 100 percentiles. The gene body coverage results were plotted for both genes using the R package ggplot2 v3.4.4.

Data Availability

The data for this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under the accession number PRJEB74238 . The information in this section will be updated once the manuscript has been conditionally accepted.

Code Availability All code written in support of this publication is publicly available at https://github.com/CSGGroup- Mainz/RNA004-Manuscript.

Acknowledgements This work was supported by funding from ERC ADG MultiOrganelleDesign (EAL). This work was also partly funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [project number 439669440 TRR319 RMaP TP A05/C01/C03 to M.H.] T.B, C.H. and S.G. acknowledge funding from the Emergent AI Center funded by the Carl-Zeiss-Stiftung. S.S. and S.G. acknowledge funding from the Forschungsinitiative Rheinland-Pfalz and the ReALity initiative of the Johannes Gutenberg University Mainz. S.Sy. acknowledge the M3odel initiative from the Forschungsinitiative Rheinland-Pfalz for financial support.

References

Abebe JS, Price AM, Hayer KE, Mohr I, Weitzman MD, Wilson AC, Depledge DP. 2022. DRUMMER—rapid detection of RNA modifications through comparative nanopore sequencing. *Bioinformatics* **38**: 3113–3115.

Alfonzo JD, Brown JA, Byers PH, Cheung VG, Maraia RJ, Ross RL. 2021. A call for direct sequencing of full-length RNAs to identify all modifications. *Nat Genet* **53**: 1113–1116.

Begik O, Lucas MC, Pryszcz LP, Ramirez JM, Medina R, Milenkovic I, Cruciani S, Liu H, Vieira HGS, Sas-Chen A, et al. 2021. Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat Biotechnol* **39**: 1278–1291.

Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, Love MI, Göke J. 2023. Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nat Methods* **20**: 1187–1195.

Cruciani S, Delgado-Tejedor A, Pryszcz LP, Medina R, Llovera L, Novoa EM. 2023. *De novo*basecalling of m ^{6}A modifications at single molecule and single nucleotide resolution. http://biorxiv.org/lookup/doi/10.1101/2023.11.13.566801 (Accessed March 28, 2024).

Delaunay S, Helm M, Frye M. 2024. RNA modifications in physiology and disease: towards clinical applications. *Nat Rev Genet* **25**: 104–122.

Desrosiers R, Friderici K, Rottman F. 1974. Identification of Methylated Nucleosides in Messenger RNA from Novikoff Hepatoma Cells. *Proc Natl Acad Sci USA* **71**: 3971–3975.

Destefanis E, Avşar G, Groza P, Romitelli A, Torrini S, Pir P, Conticello SG, Aguilo F, Dassi E. 2021. A mark of disease: how mRNA modifications shape genetic and acquired pathologies. *RNA* **27**: 367–389.

Diensthuber G, Pryszcz L, Llovera L, Lucas MC, Delgado-Tejedor A, Cruciani S, Roignant J-Y, Begik O, Novoa EM. 2023. *Enhanced detection of RNA modifications and mappability with high-accuracy nanopore RNA basecalling models*. Genomics http://biorxiv.org/lookup/doi/10.1101/2023.11.28.568965 (Accessed March 25, 2024).

532 Djirackor L, Halldorsson S, Niehusmann P, Leske H, Capper D, Kuschel LP, Pahnke J, Due-Tønnessen BJ, Langmoen
533 IA, Sandberg CJ, et al. 2021. Intraoperative DNA methylation classification of brain tumors impacts neurosurgical
534 strategy. *Neuro-Oncology Advances* **3**: vdab149.

535 Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J,
536 Amariglio N, Kupiec M, et al. 2012. Topology of the human and mouse m6A RNA methylomes revealed by m6A-
537 seq. *Nature***485**: 201–206.

538 Euskirchen P, Bielle F, Labreche K, Kloosterman WP, Rosenberg S, Daniau M, Schmitt C, Masliah-Planchon J,
539 Bourdeaut F, Dehais C, et al. 2017.Same-day genomic and epigenomic diagnosis of brain tumors using real-time
540 nanopore sequencing. *Acta Neuropathol* **134**: 691–703.

541 Furlan M, Delgado-Tejedor A, Mulroney L, Pelizzola M, Novoa EM, Leonardi T. 2021. Computational methods for RNA
542 modification detection from nanopore direct RNA sequencing data. *RNA Biology* **18**: 31–40.

543 Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, et al.
544 2018. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* **15**: 201–206.

545 Hassani A, Walton S, Shah N, Abuduweili A, Li J, Shi H. 2022. Escaping the Big Data Paradigm with Compact
546 Transformers. http://arxiv.org/abs/2104.05704 (Accessed August 1, 2023).

547 Helm M, Motorin Y. 2017. Detecting RNA modifications in the epitranscriptome: predict and validate. *Nat Rev Genet* **18**:
548 275–291.

549 Jain M, Abu-Shumays R, Olsen HE, Akeson M. 2022. Advances in nanopore direct RNA sequencing. *Nat Methods* **19**:
550 1160–1164.

551 Jenjaroenpun P, Wongsurawat T, Wadley TD, Wassenaar TM, Liu J, Dai Q, Wanchai V, Akel NS, Jamshidi-Parsian A,
552 Franco AT, et al. 2021. Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids*
553 *Research* **49**: e7.

554 Jonkhout N, Tran J, Smith MA, Schonrock N, Mattick JS, Novoa EM. 2017. The RNA modification landscape in human
555 disease. *RNA***23**: 1754–1769.

556 Jörg M, Plehn JE, Kristen M, Lander M, Walz L, Lietz C, Wijns J, Pichot F, Rojas-Charry L, Wirtz Martin KM, et al. 2024.
557 N1-methylation of adenosine (m1A) in ND5 mRNA leads to complex I dysfunction in Alzheimer's disease. *Mol*
558 *Psychiatry*. https://www.nature.com/articles/s41380-024-02421-y (Accessed March 28, 2024).

559 Karikó K, Buckstein M, Ni H, Weissman D. 2005. Suppression of RNA Recognition by Toll-like Receptors: The Impact of
560 Nucleoside Modification and the Evolutionary Origin of RNA. *Immunity* **23**: 165–175.

561 Liu H, Begik O, Lucas MC, Ramirez JM, Mason CE, Wiener D, Schwartz S, Mattick JS, Smith MA, Novoa EM. 2019.
562 Accurate detection of m6A RNA modifications in native RNA sequences. *Nat Commun* **10**: 4079.

563 Lucas MC, Pryszcz LP, Medina R, Milenkovic I, Camacho N, Marchand V, Motorin Y, Ribas De Pouplana L, Novoa EM.
564 2024. Quantitative analysis of tRNA abundance and modifications by nanopore RNA sequencing. *Nat Biotechnol*
565 **42**: 72–86.

566 Maestri S, Furlan M, Mulroney L, Coscujuela Tarrero L, Ugolini C, Dalla Pozza F, Leonardi T, Birney E, Nicassio F,
567 Pelizzola M. 2024. Benchmarking of computational methods for m6A profiling with Nanopore direct RNA
568 sequencing. *Briefings in Bioinformatics* **25**: bbae001.

569 Motorin Y, Helm M. 2022. RNAnucleotide methylation: 2021 update. *WIREs RNA* **13**: e1691.

570 Parker MT, Knop K, Sherwood AV, Schurch NJ, Mackinnon K, Gould PD, Hall AJ, Barton GJ, Simpson GG. 2020.
571 Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification
572 eds. Y. Wan and C.S. Hardtke. *eLife* **9**: e49658.

Patel A, Dogan H, Payne A, Krause E, Sievers P, Schoebe N, Schrimpf D, Blume C, Stichel D, Holmes N, et al. 2022. Rapid-CNS2: rapid comprehensive adaptive nanopore-sequencing of CNS tumors, a proof-of-concept study. *Acta Neuropathol* **143**: 609–612.

Pratanwanich PN, Yao F, Chen Y, Koh CWQ, Wan YK, Hendra C, Poon P, Goh YT, Yap PML, Chooi JY, et al. 2021. Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. *Nat Biotechnol* **39**: 1394–1402.

Schartel L, Jann C, Wierczeiko A, et al. Selective RNA pseudouridinylation in situ by circular gRNAs in designer organelles. Published online July 19, 2024. doi:10.21203/rs.3.rs-4756705/v1

Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain S. 2019. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat Commun* **10**: 3359.

Stoiber M, Quick J, Egan R, Eun Lee J, Celniker S, Neely RK, Loman N, Pennacchio LA, Brown J. 2016. *De novo*Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. http://biorxiv.org/lookup/doi/10.1101/094672 (Accessed April 5, 2024).

Suzuki T. 2020. RNA Modifications in Health and Disease. *The FASEB Journal* **34**: 1–1.

Tavakoli S, Nabizadeh M, Makhamreh A, Gamper H, McCormick CA, Rezapour NK, Hou Y-M, Wanunu M, Rouhanifard SH. 2023. Semi-quantitative detection of pseudouridine modifications and type I/II hypermodifications in human mRNAs using direct long-read sequencing. *Nat Commun* **14**: 334.

Teng H, Stoiber M, Bar-Joseph Z, Kingsford C. 2024. Detecting m6A RNA modification from nanopore sequencing using a semi-supervised learning framework. http://biorxiv.org/lookup/doi/10.1101/2024.01.06.574484 (Accessed March 28, 2024).

Ugolini C, Mulroney L, Leger A, Castelli M, Criscuolo E, Williamson MK, Davidson AD, Almuqrin A, Giambruno R, Jain M, et al. 2022. Nanopore ReCappable sequencing maps SARS-CoV-2 5′ capping sites and provides new insights into the structure of sgRNAs. *Nucleic Acids Research* **50**: 3475–348 Mary McMahon9.

Vermeulen C, Pagès-Gallego M, Kester L, Kranendonk MEG, Wesseling P, Verburg N, De Witt Hamer P, Kooi EJ, Dankmeijer L, Van Der Lugt J, et al. 2023. Ultra-fast deep-learned CNS tumour classification during surgery. *Nature* **622**: 842–849.

Wan YK, Hendra C, Pratanwanich PN, Göke J. 2022. Beyond sequencing: machine learning algorithms extract biology hidden in Nanopore signal data. *Trends Genet***38**: 246–257.

Wierczeiko A, Pastore S, Mündnich S, Busch AM, Dietrich V, Helm M, Butto T, Gerber S. 2023. NanopoReaTA: a user-friendly tool for nanopore-seq real-time transcriptional analysis ed. A. Mathelier. *Bioinformatics* **39**: btad492.

Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, et al. 2019. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods* **16**: 1297–1305.

Yue Y, Liu J, Cui X, Cao J, Luo G, Zhang Z, Cheng T, Gao M, Shu X, Ma H, et al. 2018. VIRMA mediates preferential m6A mRNA methylation in 3′UTR and near stop codon and associates with alternative polyadenylation. *Cell Discov* **4**: 10.

Zhang K, Chen Y-J, Wilde D, Doroschak K, Strauss K, Ceze L, Seelig G, Nivala J. 2022. A nanopore interface for higher bandwidth DNA computing. *Nat Commun* **13**: 4904.

Zhong Z-D, Xie Y-Y, Chen H-X, Lan Y-L, Liu X-H, Ji J-Y, Wu F, Jin L, Chen J, Mak DW, et al. 2023. Systematic comparison of tools used for m6A mapping from nanopore direct RNA sequencing. *Nat Commun***14**: 1906.

616