Detection of pseudouridine modifications and type I/II hypermodifications in human mRNAs using direct, long-read sequencing

Sepideh Tavakoli[1‡], Mohammad Nabizadehmashhadtoroghi[2‡], Amr Makhamreh[1], Howard Gamper[4], Neda K. Rezapour[3], Ya-Ming Hou[4], Meni Wanunu[1, 3], and Sara H. Rouhanifard[1#]

[1]*Dept. of Bioengineering, Northeastern University, Boston, MA*

[2]*Dept. of Mechanical Engineering, Northeastern University, Boston, MA*

[3]*Dept. of Physics, Northeastern University, Boston, MA*

[4]*Dept. of Biochemistry and Molecular Biology, Thomas Jefferson University, Philadelphia, PA*

[‡]*These authors contributed equally to this work.* [#]*Corresponding author.*

Abstract

Enzymatic modifications to mRNAs have the potential to fine-tune gene expression in response to environmental stimuli. Notably, pseudouridine-modified mRNAs are more resistant to RNase-mediated degradation, more responsive to cellular stress, and have the potential to modulate immunogenicity and enhance translation *in vivo*. However, the precise biological functions of pseudouridine modifications remain unclear due to the lack of sensitive and accurate mapping tools. We developed a semi-quantitative method for high-confidence mapping of pseudouridylated sites on mammalian mRNAs via direct long-read nanopore sequencing. A comparative analysis of a modification-free transcriptome reveals that the depth of coverage and intrinsic errors associated with specific k-mer sequences are critical parameters for accurate basecalling. We adjust these parameters for high-confidence U-to-C basecalling errors that occur at pseudouridylated sites, and benchmark against sites that were previously identified in human rRNA or mRNA using biochemical methods. Using our method, we uncovered new pseudouridylated sites, many of which fall in k-mers that are known targets of pseudouridine synthases. Sites identified by U-to-C base calling errors were validated using 1000-mer synthetic RNA controls bearing a single pseudouridine in the center position. Our validation approach demonstrates that while U-to-C basecalling error occurs at the site of pseudouridylation, this basecalling error is systematically under-called at the pseudouridylated sites. We use our method to discover mRNAs with up to 7 unique sites of pseudouridine modification. Our pipeline allows direct detection of low- and high-occupancy pseudouridine modifications on native RNA molecules from nanopore sequencing data without resorting to RNA amplification, chemical reactions on RNA, enzyme-based replication, or DNA sequencing steps.

Introduction

Enzyme-mediated RNA chemical modifications have been extensively studied on highly abundant RNAs such as transfer RNAs[1] and ribosomal RNAs[2], however, we now know that messenger RNAs are also targets of RNA modification. Although modifications occur to a lesser extent in mRNAs than other RNAs[3], these modifications potentially impact gene expression[4], RNA tertiary structure formation[5], or the recruitment of RNA-binding proteins[6]. Pseudouridine (psi) is synthesized from uridine converted *in vivo* by one of more than a dozen pseudouridine synthases identified to date[7]**.** It was the first discovered RNA modification[8] and represents 0.2-0.6% of total uridines in mammalian mRNAs[3]. Psi-modified mRNAs are more resistant to RNase-mediated degradation[9] and also have the potential to modulate splicing[10] and immunogenicity[11] and alter translation[12,13] *in vivo*. Further, psi modifications of RNAs are responsive to cellular stress, leading to increased RNA half-life[14,15]. Little is known about the biological consequences of pseudouridylation, except for a few well-studied cases. For example, defective pseudouridylation in cells leads to disease, including X-linked dyskeratosis

congenita, a degeneration of multiple tissues that severely affects the physiological maintenance of 'stemness' and results in bone marrow failure[16,17]. A critical barrier to understanding the precise biological functions of pseudouridylation is the absence of high-confidence methods to map psi-sites in mRNAs. Psi modifications do not affect Watson-Crick base pairing[18], thereby making them indistinguishable from uridine in hybridization-based methods. Additionally, the modification bears the same molecular weight as the canonical uridine, making it challenging to detect directly by mass spectrometry[19,20].

Psi is conventionally labeled using N-cyclohexyl-N′-b-(4-methylmorpholinium) ethylcarbodiimide (CMC), a reagent that modifies the N1 and N3 positions of psi, N1 of guanine, and the N3 of uridine[21]. Treatment with a strong base removes the CMC from all the sites except for the N3 position of psi. Recently, the use of an RNA bisulfite reaction was demonstrated for the specific labeling of psi[22]. Indirect chemical labeling of psi combined with next-generation sequencing[3,15,22] has yielded over 2,000 putative psi sites within mammalian mRNAs, but different methods identified different sites and the overlap is limited[23], pointing to a need for improved detection and quantification technology. Reliance on an intermediate chemical reaction (i.e., CMC or RNA bisulfite) can lead to false-positive or false-negative results due to incomplete labeling or stringent removal of reagent from the N1 position of psi[24]. Further, each of these methods relies on the amplification of a cDNA library generated from the chemically modified mRNAs, leading to potential false positives from biased amplification. Finally, since these methods rely on short reads, it is difficult to perform combinatorial analysis of multiple modifications on one transcript.

Recently, several studies have reported using nanopore-based direct RNA sequencing to directly read RNA modifications[25–30]. In these reports, ion current differences for different k-mer sequences (k = 5 nucleotides) as an RNA strand is moved through the pore suggest the presence of a modified RNA base. Detection of psi using nanopores was also confirmed for rRNAs[27], for the *Saccharomyces cerevisiae* transcriptome[28], and for viralRNAs[30], as indicated by a U-to-C base-calling error at various sequence sites. Algorithms for psi quantification have been produced[28,29] for various k-mers using combinatorial sequences that contain psi sites within close proximity as well as control RNAs containing many natural RNA modifications, also in close proximity (e.g., rRNA). While these control molecules allow many k-mers to be studied, the accuracy of quantifying psi occupancy at a given modified site can be highly dependent on the nucleotide sequence surrounding the modification. Moreover, sequence context is particularly important for the measurement of RNA molecules wherein the secondary structure can influence the kinetics of translocation as mediated by the helicase[31]. Control molecules for psi modification that match the transcriptome sequence beyond the context of the measured k-mer are more desirable than random sequences.

Here, we describe a nanopore-based method to accurately map psi modifications in a HeLa transcriptome by comparing the sequence alignment to identical negative controls without RNA modifications. We demonstrate that the number of reads and specific k-mer sequences are critical parameters for defining psi sites and for assigning significance values based on these parameters, enabling us to make high-confidence and conservative, binary identifications of psi modification sites, transcriptome-wide. Our approach recapitulates 198 previously annotated psi sites, 34 of which are detected by 3 independent methods, thus providing a "ground truth" list of psi modifications in HeLa cells. Our approach also reveals 1,691 putative psi sites that have not been reported previously. We show that these new sites tend to occur within k-mer sequences including the PUS7 and TRUB1 sequence motifs that were previously reported.

We validate the accuracy of our algorithm for detecting sites of psi modifications using ribosomal RNAs which have been comprehensively annotated by mass spectrometry and assigned 41/46 psi modification using our method. Additionally, we synthesized and analyzed five 1,000-mer synthetic RNA controls containing either uridine or psi at a known pseudouridylated position in the human transcriptome. This quantitative analysis revealed that U-to-C mismatch errors are systematically under-called for the detection of psi, thus enabling us to apply a basecalling error cutoff to identify 40 high-occupancy, hypermodified type I psi sites, which are likely to confer measurable phenotypes. We discovered that these sites tend to occur in k-mer sequences for which uridine and guanine precede the pseudouridylated site. In accordance with previous findings that show higher median psi-ratio for positions with the TRUB1 and the PUS7 sequence motifs as compared to the other k-mers[23].

Further, we identify 38 mRNAs with multiple high-confidence psi sites, which are confirmed by single-read analysis. Interestingly, we find mRNAs with up to 7 unique psi sites. Combined, this work reports a pipeline that enables direct identification and quantification of the psi modification on native mRNA molecules, without requiring chemical reactions on RNA or enzyme-based amplification steps. The long nanopore reads allow, in principle, the detection of multiple modifications on one transcript, which can shed light on cooperative effects on mRNA modifications as a mechanism to modulate gene expression.

Results

Nanopore analysis of an unmodified HeLa transcriptome generated by *in vitro* transcription

To identify putative sites of mRNA psi modifications, we extracted RNA from HeLa cells and prepared two libraries: The first (direct) library consists of native mRNAs, which contain both canonical uridine and naturally occurring uridine modifications. The second consists of an *in vitro* transcribed (IVT) mRNA control library in which polyadenylated RNA samples were reverse transcribed to cDNAs, which were then transcribed back into RNA *in vitro* using canonical nucleotides to ensure the absence of RNA modifications (**Fig. 1a**). Each library was sequenced on a separate MinION flowcell and basecalled using *Guppy 3.2.10*. Three direct RNA libraries produced an average of ~1.2 million poly(A) RNA strand reads, respectively, of which ~800,00 (read quality of 7), with a read average N50 length (defined as the shortest read length needed to cover 50% of the sequenced nucleotides) of 850 bases and a median length of ~670 bases (**Supplementary Fig. 1**). Similarly, two IVT libraries produced an average of 1.6 million passed the quality filter, with N50 of 890 and a median length of 710 bases. Alignment was performed using minimap2.17[32] and the reads for each library were subsequently aligned to the GRCh38 human genome reference.

Basecalling accuracy is used to identify psi modifications in RNA

To define differences between the IVT and direct libraries for psi detection, any source of error other than the psi modification itself must be minimized, including misalignments to the GRCh38 human genome reference. We minimized the chances of incorrect alignment by only considering the primary alignment of each read (i.e., the alignment with the highest mapping quality. Also, any read with a mapping quality score lower than 20 was discarded from the downstream analysis, because the probability of correct alignment was <99%. The second potential source of error is the presence of single-nucleotide polymorphisms (SNPs), whereby the base is different from the reference genome. We identified likely SNP sites based on an

equivalent U-to-C mismatch percentage in both the IVT and the direct RNA sequencing samples (**Supplementary Fig. 2**), whereas in the case of a modified RNA nucleotide, the mismatch percentage in the direct RNA sequencing sample was significantly higher relative to the one from IVT at the site of modification (**Supplementary Fig. 2**). The third source of error is a basecalling error, whereby the basecalling algorithm fails to identify the correct base. To assess the basecalling accuracy using the *Guppy 3.2.10* algorithm, we calculated the error in the IVT control sample by comparing the basecalling to the reference genome (**Fig. 1b**). Since the IVT control contains only the canonical RNA nucleotides, these errors were independent of RNA modifications. We confirmed that the basecaller could reliably identify unmodified and aligned nucleotides with an average error of 2.64%.

To confirm the quality of the IVT unmodified transcriptome, we compared the transcripts per million (TPM) for individual genes in the IVT and direct RNA libraries and found that the TPMs were very similar (R^2=0.96; **Figure 1c.** We also compared the distribution of read lengths for the IVT and direct RNA libraries and found that the samples were overlapping (**Figure 1d**); likewise, the coverage for individual transcripts was similar for IVT and direct RNA libraries (**Figure 1e**), thus supporting the use of the IVT library as an equivalent, unmodified transcriptome control.

Direct RNA nanopore sequencing identifies pseudouridine modifications in mRNA via systematic U-to-C base-calling errors

We then examined specific locations on human mRNAs that have been previously identified as psi sites by chemical-based methods (**Figure 1f**). We selected 5 genes as examples: *IDI1* (chr10:1044099)[3,14,22], *PARP4* (chr13:24426505)[3,22], *PSMB2* (chr1:35603333)[3,14,22], *MCM5* (chr22:35424407) [3,14], and *PABPC4* (chr1:39565149)[3,14], representing a range of different k-mers with a putative psi in the center nucleotide (GUUCA, GUUCA, GUUCG, UGUAG, and GUUCC respectively). A range of k-mers was chosen because specific k-mer sequences can influence the accuracy of base-calling (**Supplementary Fig. 3**). We detected a systematic U-to-C mismatch error at the reported psi site in duplicates of each gene by direct RNA sequencing (*IDI1* (chr10:1044099)*:* 96.06±1.16*%*, *PARP4* (chr13:24426505): 91.71 ±7.56%*, PSMB2* (chr1:35603333): 81.07 ±1.68%, *MCM5* (chr22:35424407): 54.82 ± 4.96%, *PABPC4* (chr1:39565149): 55.08 ±3.97%*).* We confirmed that the IVT samples maintained the standard base-calling error at each site (3.75%, 4.54%, 1.67%, 5.26% and 8.34% respectively; **Fig. 1c**).

Calculating the significance of U-to-C mismatch as a proxy for psi is dependent on mismatch percentage at a given site, the number of reads, and the surrounding nucleotides.

To further improve the use of the U-to-C mismatch error as a proxy for psi we needed to minimize the error that occurs from other factors. We observed that the base quality on sites that have 3 or fewer reads is low, relative to the rest of the population, which would create bias in the downstream analysis (**Fig. 2a**). To ensure sufficient coverage in both the direct RNA and IVT samples, we require a minimum of 7 reads represented from each biological replicate for *de novo* detection. One reason for the lower quality of these sites is their proximity to the start/end of the aligned section of their corresponding reads. It is common for the aligner to clip a few mismatched bases from the start/end of reads (known as “soft-clipping”). We show that up to 3 bases adjacent to the soft clipped site usually yield lower base quality, and thus are not reliable regions to obtain information from (**Supplementary Fig. 4**).

To further investigate these mismatch errors, we gathered the data for all the canonical uridine sites from our IVT control sample (>3 million uridine sites transcriptome-wide). For each of these

positions, we calculated the U-to-C mismatch percentage, the number of aligned reads, and analyzed the surrounding bases of each site (i.e., we tabulated their 5-mers for which the target uridine site falls in the center). As expected, higher error rates were observed among low coverage sites (**Fig. 2b**). Additionally, the surrounding bases of a site influenced the mismatch error (**Fig. 2c**). For example, uridine sites within the CUUUG k-mer, on average, showed a 10% mismatch error in the IVT reads, while uridine sites within the AAUCU k-mer had less than 0.4% average mismatch error. The average U-to-C mismatch of the specific k-mer in IVT is an important factor to be considered because it is essential to prevent a misinterpretation of the inherent error of a k-mer as a site of modification. Therefore, the significance of the U-to-C mismatch percentage of a site must be interpreted based on a combination of the mismatch percentage (in the direct RNA sample), the number of reads (in the direct RNA sample), and the average U-to-C mismatch error of the equivalent k-mer (derived from the IVT sample; **Figure 2d**, **Supplementary Methods**).

It is important to ensure that the targets are not selected based on errors from other sources such as single-nucleotide polymorphisms, basecalling, or alignment. In the cases that the IVT error at a specific position is higher than the average error of that k-mer, the mismatch error from the direct RNA reads is compared with error at the specific site rather than the average error of that k-mer in IVT. To account for standard basecalling errors, we compare the direct reads to the IVT replicate with the highest error at that specific site.

Benchmarking of algorithm for predicting psi sites ($p < 0.001$) on human rRNA.
Human rRNA has been extensively annotated using mass spectroscopy. To benchmark our algorithm for selecting sites of psi modification, we generated and analyzed direct rRNA and IVT rRNA libraries from HeLa cells. A total of 43 previously validated rRNA positions from the 18s and 5.8s subunits had sufficient coverage for analysis. Of these sites, 38 (88.4%) were detected as psi using our method ($p < 0.001$; **Figure 2e-g**, **Supplementary Table 1**). Additionally, we detected 72 targets that are not on the list of previously detected psi positions. Further inspection reveals that 10/72 of those positions exist in validated rRNA positions but modified as 5-methyl uridine. Most of the remaining positions (53/62) are within 4 bases of another modification in rRNA, thus highlighting a limitation of detection in regions with modifications in proximity (**Figure 2h-i**, **Supplementary Figure 5**, **Supplementary Table 1**).

Benchmarking of putative, psi sites ($p < 0.001$) against existing methods.
Previous studies have identified putative psi sites on human mRNA using biochemical methods including CMC[3,14,15] and RNA bisulfite[22] (**Fig. 3a-d**). To evaluate the validity of our nanopore-based method we generated a list of 334 putative psi positions which were previously annotated by one or more biochemical methods, and selected a subset of these 334 targets that produced at least 7 reads in our direct RNA nanopore sequencing. To assess the ability of our algorithm for identifying psi positions within our direct RNA libraries, we assign *p* values (**Figure 2d**) for each of the putative psi positions and found 232 positions with $p < 0.01$ (**Fig. 3e**). Among these positions, 198 sites were validated by one other method and 34 were validated by two or more methods[22] which we define as “ground truth” due to the coincidence of all three methods, i.e., 2+ methods and nanopore (**Fig. 3f, Supplementary Table 1**). For sites with sufficient coverage, our algorithm for determining psi positions from direct RNA nanopore libraries had the highest overlap with Pseudo-seq (87.8%), followed by RBS-seq (77.9%), and lowest with CeU seq (67.6%).

Additional analysis of the positions that were identified by 2 or more, independent biochemical methods revealed 5 additional positions that are covered by direct RNA sequencing but were not identified as having a psi by our algorithm (**Fig. 3f**). These positions include *COL4A2* (chr13:110512877), *RPL18A* (chr19:17862095), *CTSA* (chr20:45897784), and *SLC25A1* (chr22:19177964), each of which had a low mismatch error in direct RNA sequencing, and FAM168B (*chr2:131049504*) that had high error in the corresponding IVT control.

Detection of putative psi sites of mRNA *de novo* using direct RNA nanopore sequencing.
Next, we sought to apply our method for *de novo* detection of putative psi sites, transcriptome wide. The filtration of these sequences was critical to ensure that the list we produce is conservative. To minimize the inclusion of sites that appear due to random error, we calculated significance based on the higher error between two replicates of IVT. We also required that two out of three direct replicates have the $p \leq 0.01$ to be defined as a putative psi site. Using our algorithm, we detected 1691 putative psi sites ($p < 0.01$), including 730 positions with a p value cutoff of 0.001 (**Fig. 4a, Supplementary Table 3**). Gene ontology analyses (GO Molecular Function 2021) were performed on genes with $p < 0.001$ using enrichR website, showing that the "RNA binding" group has the highest normalized percentage of these genes which is similar with all the transcripts GO (**Supplementary Fig. 6**) (**Supplementary Table 6**).

Distribution of highly represented psi-containing k-mers in the human transcriptome.
We assessed the k-mer frequencies for putative, pseudouridylated targets detected *de novo* with $p < 0.001$ (**Fig. 4b**) and found that, as expected, UGUAG which is the motif for PUS7 binding[33] is the most highly represented k-mer and the GUUCN k-mer, the motif for TRUB1[23], is among the most frequently detected targets. To evaluate the sequence conservation of nucleotides within k-mers bearing a psi site in the center position, we plotted the sequencing logo and found that the surrounding positions do not show any nucleotide preference (**Fig. 4c**).

Distribution of psi sites on mRNA sequences.
We characterized the distribution pattern of psi modifications on mature mRNA transcripts and observed that around 60% of them were located on the 3' untranslated region (UTR) and 35% on the coding sequence (CDS), with very few targets detected in the 5' UTR (**Fig. 4d**). The limited detection of psi sites in the 5' UTR is likely due to the low observed coverage in the 5' end of the RNA (i.e., near the transcription start site and covering a majority of the 5' UTR in many cases; **Fig. 4e**). Low coverage in the 5' ends of RNA is expected since the enzyme motor releases the last ~12 nucleotides, causing them to translocate at speeds much faster than the limit of detection[25]. Compared to the rest of the transcript, there is also a sharp drop in coverage at the tail end of the 3' UTR (near the transcription termination site, **Fig. 4e**). Interestingly, we found one example of a putative psi modification within a transcription stop site: *GAGE2A* (chrX:49596731).

We calculated the distance of each putative psi from the closest splice site for high confidence psi sites ($p < 0.001$). Prior to extracting the distance of the nearest splice junction for each target, the RNA isoform analysis tool, FLAIR[34], was used to bin the reads comprising high confidence pseudouridylated targets into their respective dominant isoform. Overall, targets in the 3' UTRs are separated from a splice site by a longer distance relative to targets in coding sequences (CDS) (**Figure 4f**). Considering the significant discrepancy in sequence length between CDS and 3' UTR, we observed a higher correlation between the splice distance of CDS-positioned targets and CDS length as compared to 3' UTR-positioned targets (**Supplementary Figure 6**).

U-to-C mismatch error from synthetic RNA controls with a site-specific psi is systematically under called for psi percentage

To verify that our algorithm reliably detects psi sites *de novo*, and to explore the quantitative accuracy of the U-to-C mismatch error as a proxy for pseudouridylation, we constructed five 1,000-mer synthetic mRNAs bearing a pseudouridine at the nanopore detected site (**Fig. 5a**). These controls were designed to recapitulate the 1,000-mer sequence flanking a naturally occurring psi in the human transcriptome. Two of the chosen targets (*PSMB2*; chr1:35603333[3,14,22] and *MCM5;* chr22:35424407[3,14]) were detected from two or more previous methods and the other three targets (*MRPS14;* chr1:175014468, *PRPSAP1;* chr17: 76311411, and *PTTG1P*; chr21:44849705) were detected *de novo* using the U-to-C mismatch error and our *p-value* cutoff. For each site, we constructed 1,000-mer RNA transcripts where the center uridine position was replaced with psi and ran these synthetic controls through the nanopore directly and measured the U-to-C mismatch error for each. If the mismatch error were a perfect proxy for psi, we expected to see 100% U-to-C mismatch in these synthetic controls. In contrast, we observed 38.17% U-to-C mismatch error for *PSMB2*, 32.16% for *MCM5*, 69.64% for *PRPSAP1*, 69.35% % for *MRPS14,* and 30.08% for *PTTG1P* (**Fig. 5b**). These results indicate a systematic under-calling of psi using our algorithm.

Pseudouridylated targets with >40% U-to-C mismatch error are classified as having type I hypermodification

We define hypermodification type I as a specific site within a transcript in which at least every other copy has a psi modification. We, therefore, reasoned that a 40% mismatch error was an appropriate cutoff because the base caller is systematically under-calling the psi. We further reasoned that 40% is at a maximum, representing half-modified transcripts. From our *de novo* psi detection analysis, we identified 40 unique sites of hypermodification type I including *AK2* (chr1: 33014553), *IDI1*(chr10:1044099), *GTF3C2*(chr2:196789267), *RHBDD2* (chr7:75888787)*, HSPD1 (chr2:197486726)* that show close to 100% mismatch error (**Supplementary Table 4**).

To assess the sequence conservation of nucleotides within k-mers bearing a psi in the center position, we selected all unique pseudouridylated sites with U-to-C mismatch error above 40% (**Supplementary Table 4**). We found that the -1 position shows a strong preference for uridine and the -2 position shows a strong preference for guanidine. This preference pattern becomes more significant as the mismatch percentage increases (**Fig. 5c**). The +1 position shows a strong preference for cytidine especially at sites with higher than 80% U-to-C mismatch error.

We then assessed the k-mer frequencies for psi targets detected *de novo* with U-to-C mismatch error at greater than 40% (**Fig. 5d**) as well as k-mer frequencies for psi targets detected de novo with an error less than 40% (**Fig. 5e**). We found that the GUUCN k-mer, the motif for TRUB1, represents most of the targets (30/105 sites around 29%). The k-mer UGUAG, the motif for PUS7 binding, was also detected (5/105 sites around 4.8%). In contrast, k-mer UGUAG (13/712, 1.8%), GUUCN, and all others occurred at a similar frequency as the most abundant “not hypermodified” targets (15/712, 2.1%). Indeed, sequence-specific recognition by TRUB1 is demonstrated by the observation of the highest pseudouridylation frequencies of its k-mer relative to the k-mer recognized by PUS7 and k-mers recognized by other enzymes.

We assessed the location of putative psi modifications on the transcript and found that type I hypermodified sites are biased towards 3’ UTRs, which is the same as sites that are not

hypermodified (**Supplementary figure 7**). No significant difference was observed in the splice distance of type I hypermodified sites between sites in the 3' UTR and those in CDS regions of mRNA when compared to "not hypermodified" sites (**Supplementary Figure 7**).

Messenger RNAs with more than one psi site are classified as having type II hypermodification

We define hypermodification type II as the mRNAs that can be pseudouridylated at two or more positions (**Fig. 6a**). Using only the sites with a high probability of psi modification (*p-value* <0.001), we identified 104 mRNAs pseudouridylated at 2 unique positions, 27 with 3 positions, 4 with 4 positions, 5 with 5 positions, 1 with 6 positions and 1 mRNA with 7 positions (**Fig. 6b**). For the mRNAs that are pseudouridylated at 2 positions, we plotted the mismatch error of the first and second sites of modification and found no correlation between the mismatches (R = 0.039; **Fig. 6c**) although this percentage is highly k-mer dependent. To determine if genes with 2 sites of pseudouridylation have the tendency to occur on the same read, we plotted every read for two mRNAs (*ATP5MPL* and *SLC2A1)* and labeled each site using the called base (canonical U or C indicating the presence of a pseudouridine; **Fig 6d**). We observed that these mismatches could happen on the same read or only on one read. For example, mismatch percentages for *SLC2A1* are 68.5% in position 1 (chr1:42926727) and 48.1% in position 2 (chr1:42926879) (31% on both, 54% on only one of them, 15% on none). Similarly, mismatch percentages for *ATP5MPL* are 12.6% in position 1 (chr14:103912536), 38.4% in position 2 (chr14:103912631) (7% mismatches on both sites, 37% on only one site, and 56% no mismatches). We plotted the distribution of psi type II hypermodification across the body of the transcript for trtanscripts bearing 3, 4 and 7 putative psi modifications and found a slight clustering of sites in the 3' UTR but the sites were relatively spread out across the transcript body suggesting independent enzymatic events.

Discussion

We have shown here that systematic U-to-C basecalling errors detected during direct nanopore sequencing of transcriptomes can serve as indicators for psi modifications, provided that the total number of reads, as well as the systematic error associated with the specific canonical (unmodified) k-mer, are considered. We provide a foundation for identifying psi sites with high confidence based on two approaches. In the first approach, U-to-C mismatch errors in native transcriptomes are compared against a corresponding, unmodified transcriptome as a negative control to eliminate standard basecalling errors that occur in canonical bases, and in addition, we uniquely weigh transcriptome wide average U-to-C errors in k-mers to minimize false positives due to low coverage. In the second, we use a set of long synthetic RNA control molecules with precisely positioned psi modifications, which aided in our discovery of systematic under-calling of psi modifications, pointing to limitations of basecalling-guided RNA modification detection algorithms. Our approaches are distinct from the ELIGOS algorithm because the average U-to-C error of unmodified k-mers is considered, enabling the analysis of low coverage sites that may show as significant error due to random nanopore basecalling error. Additionally, we determine individual psi sites by the exclusive presence of U-to-C mismatches rather than including all other substitutions, deletions, and insertions at a given site; this significantly reduces false positives in psi detection.

We demonstrate that this method for identifying psi sites can faithfully reproduce sites that were detected by CMC and bisulfite-based next-generation sequencing platforms. Importantly, we produce a "ground truth" list (198 mRNA positions detected by nanopore and one additional

method and 34 mRNA positions detected by nanopore and two additional methods) with validated, psi modifications in HeLa cells (**Supplementary table 1**) --a conservative list of putative targets to make the study of psi biology in cells more accessible. This work has also resulted in a comprehensive list of 1,691 novel sites of putative psi modification, which may be used by the field.

Interestingly, our algorithm for determining psi positions from direct RNA nanopore libraries had the highest overlap with Pseudo-seq (87.8%), followed by RBS-seq (77.9%), and lowest with CeU seq (67.6%). Among the methods that we used to validate our data, Pseudo-seq shows the highest overlap between the detection targets. However, some targets that the other methods detected were not detected by our method. For example, we found 334 sites of psi modification that were detected by 1 or more biochemical methods, but we only call 232 of these by nanopore sequencing. We suggest that artifacts from CMC labeling may account for this, including incomplete CMC adduct removal from unmodified uridines, reverse-transcriptase read through of CMC-modified psi sites or uneven amplification of low-occupancy psi-sites. Another potential reason for the differences could be batch differences between cell lines leading to differential occupancy at a given moment of mRNA extraction. On the other hand, our method of psi detection using nanopores also has limitations including limited coverage of individual genes, thus leading to many false negatives, and unpredictable basecalling in the presence of other modifications as shown in our rRNA data in **Figure 2**. We also observed several targets that were detected by our nanopore method that were not detected by other methods. While we are confident that these sites are modified due to differences between the native RNA versus the IVT control, and likely psi, we cannot rule out the possibility of other uridine modifications.

Previous studies have demonstrated the importance of long-range interactions[30] for the accurate calling of psi modifications by direct RNA sequencing, and a dwell time signature for a subset of psi-sites. The dwell time signature has the potential to increase the number of psi-identifications *de novo*, however, low coverage in transcriptome-wide sequencing runs precludes this analysis at this time. To account for the contributions of long-range interactions, we have validated our method by analysis of five synthetic 1,000-mers, each containing a site-specific psi found within a natural target sequence in the human transcriptome. We find that the U-to-C basecalling error systematically under-calls the psi modification. Based on this finding, we defined psi hypermodification type I as sites that have >40% U-to-C mismatch error. We also define hypermodification type II as mRNAs bearing multiple psi modification sites in a specific transcript. Finally, we show for the first time that psi modification can occur up to 7 times on a single transcript.

A fully quantitative measure of psi occupancy at a given site would require high-coverage sequencing runs of a comprehensive set of every possible, psi-containing k-mer within its natural sequence context (an estimated 13 nucleotides surrounding the modified site). While similar controls have previously been generated[28,30], all uridines were modified in those studies and consequently, these are not the ideal controls for detection of single psi modifications within the natural sequence contexts. Additionally, ribosomal RNA-based controls, which contain highly conserved modification sites, are not ideal controls for mRNA modifications because the spatial distribution of modifications in rRNA is very different than the sparse spatial distribution of modifications in mRNA. Although preparation of such a large set of control molecules is not feasible for any single laboratory, it is increasingly apparent that such a set would resolve remaining ambiguities in psi detection through nanopore sequencing. Although our method is

semi-quantitative, the synthetic controls that we have generated demonstrate that the basecalling error is reliable in the calling of psi at a given site. By setting a cutoff of 40% U-to-C mismatch for a given site we conservatively draw a list of high-confidence sites that are pseudouridylated with high occupancy, and thus, have a higher likelihood of leading to a measurable phenotype in the cell and conferring a functional impact on the cellular physiology.

Our work provides a powerful foundation for detection and analysis of psi modifications on mRNAs with sequence specificity and single-molecule resolution. Future work should include an expansion of synthetic controls and training of a new basecaller to improve our ability to quantify RNA modifications.

Methods

Cell culture:

HeLa cells were cultured in Dulbecco's modified Eagle's medium (Gibco, 10566024), supplemented with 10% Fetal Bovine Serum (FB12999102, FisherScientific) and 1% Penicillin-Streptomycin (Lonza,17602E). To extract sufficient poly-A RNA, three confluent, 10cm dishes were used for each experiment.

Total RNA extraction and Poly(A) RNA isolation:

The total RNA extraction protocol was performed using a method that is the combination of total RNA extraction using TRIzol (Invitrogen,15596026) and PureLink RNA Mini Kit (Invitrogen, 12183025). Cell types were washed with 3 ml ice-cold PBS. 2 ml of TRIzol was added to each 10cm dish and incubated at room temperature for 5 min. Every 1 ml of lysed cells in TRIzol was transferred to a LoBind Eppendorf tube and vortexed for 30 sec. 200 μl chloroform (Acros Organics,423555000) was added to each tube and mixed by shaking for 15 sec and incubated at room temperature for 3 min. Then the samples were centrifuged at 12000 XG for 15 min at 4°C. 0.4 ml of aqueous supernatant is transferred to a new LoBind Eppendorf tube and an equal volume of 70% ethanol is added to the solution followed by vortexing. In the following steps, PureLink RNA Mini Kit (Invitrogen, 12183025) and the protocol are performed according to the manufacturer's recommended protocol. Briefly, the solution is transferred to a pure link silica spin column and flow-through was discarded (every two microtubes were loaded on one column). The columns were washed with 0.7 ml of wash buffer I once and then with 0.5 ml wash buffer II twice. The total RNA was eluted using 50 ul nuclease-free water. The RNA concentration was measured using a NanoDrop 2000/2000c Spectrophotometer.
NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490L) is used to select poly(A) mRNA. The protocol is followed according to the manufacturer's protocol. The only modification was pooling 5 samples and performing the experiment in microtubes instead of PCR tubes. 15 samples (3 microtubes) were used in each experiment to get enough Poly-A RNA product. The products were eluted from the NEBNext polyA magnetic isolation (NEB, E7490S) in tris buffer. The three samples were pooled and ethanol precipitated to get to the concentration that is required for the sequencing step.

In vitro transcription, capping, and polyadenylation

cDNA-PCR Sequencing Kit (SQK-PCS109) kit was used for reverse transcription and strand-switching. Briefly, VN primer (VNP) and Strand-Switching Primer (SSP) were added to 50 ng poly-A RNA. Maxima H Minus Reverse Transcriptase (Thermo scientific, EP0751) was used to produce cDNA. IVT_T7_Forward and reverse primers were added to the product and PCR

amplified using LongAmp Taq 2X Master Mix (NEB, M0287S) with the following cycling conditions: Initial denaturation 30 secs @ 95 °C (1 cycle), Denaturation 15 secs @ 95 °C (11 cycles), Annealing 15 secs @ 62 °C (11 cycles), Extension 15 min @ 65 °C (11 cycles), Final extension 15 mins @ 65 °C (1 cycle), and Hold @ 4 °C. 1 µl of Exonuclease 1 (NEB, M0293S) was added to each PCR product and incubated at 37C for 15 min to digest any single-stranded product, followed by 15 min at 80C to inactivate the enzyme. Sera-Mag beads (9928106) were used according to the Manufacturer's protocol to purify the product. The purified product was then in vitro transcribed using "HiScribe T7 High yield RNA Synthesis Kit (NEB, E2040S) and purified using Monarch RNA Cleanup Kit (NEB, T2040S). The product was eluted in nuclease-free water and poly-A tailed using E. coli Poly(A) Polymerase (NEB, M0276). The product was purified once again using an RNA Cleanup Kit and adjusted to 500 ng polyA RNA in 9 ul NF water to be used in the Direct RNA library preparation.

For rRNA IVT, total RNA was poly-A tailed using E. coli Poly(A) Polymerase (NEB, M0276) and purified using RNA Cleanup kit (NEB, T2040S) then poly-A selected using NEBNext polyA magnetic isolation (NEB, E7490S). 50 ng of the poly-A tailed total RNA was the in vitro transcription according to the above protocol.

Synthetic sequence design

We constructed four synthetic 1,000-mer RNA oligos, each with a site-specifically placed k-mer. Two versions of each RNA were prepared, one with 100% uridine and the other with 100% psi at the central position of the k-mer. The uridine-containing RNAs were prepared byT7 transcription from G-block DNAs (synthesized by Integrated DNA Technologies), whereas the psi-containing RNAs were prepared by ligation of left and right RNA arms (each 500 nts in length) to a 15-mer RNA bearing a psi in the central position (synthesized by GeneLink). A T7 promoter sequence with an extra three guanines was added to all the DNA products to facilitate *in vitro* transcription. In addition, a 10 nt region within 30 nt distance of ψ was replaced by a barcode sequence to allow parallel sequencing of the uridine- and psi-containing samples. Finally, each left arm was transcribed with a 3' HDV ribozyme that self-cleaved to generate a homogeneous 3'-end. Full-length RNA ligation products were purified using biotinylated affinity primers that were complementary to both the left and right arms.

Direct RNA library preparation and sequencing

The RNA library for Direct RNA sequencing (SQK-RNA002) was prepared following the ONT direct RNA sequencing protocol version DRCE_9080_v2_revH_14Aug2019. Briefly, 500 ng poly-A RNA or poly-A tailed IVT RNA was ligated to the ONT RT adaptor (RTA) using T4 DNA Ligase (NEB, M0202M). Then the product is reverse transcribed using SuperScript™ III Reverse transcriptase (Invitrogen, 18080044). The product was purified using 1.8X Agencourt RNAClean XP beads, washed with 70% ethanol and eluted in nuclease-free water. Then the RNA: DNA hybrid ligated to RNA adapter (RMX) and purified with 1X Agencourt RNAClean XP beads and washed twice with wash buffer (WSB) and finally eluted in elution buffer (ELB). The FLO-MIN106D was primed according to the manufacturer's protocol. The eluate was mixed with an RNA running buffer (RRB) and loaded to the flow cell. MinKnow (19.12.5) was used to perform sequencing. Three replicates were from difference passages and different flow cells were used for each replicate. For Direct rRNA library preparation, total RNA was poly-A tailed using E. coli Poly(A) Polymerase (NEB, M0276) and purified using RNA Cleanup kit (NEB, T2040S) following up with the above protocol.

Base-calling, alignment, and signal intensity extraction

Multi-fast5s were base-calling real-time by guppy (3.2.10) using the high accuracy model. Then, the reads were aligned to the genome version hg38 using minimap 2 (2.17) with the option ''-ax splice -uf -k14''. The sam file was converted to bam using samtools (2.8.13). Bam files were sorted by "samtools sort" and indexed using "samtools index" and visualized using IGV (2.8.13). The bam files were sliced using "samtools view -h -Sb" and the signal intensities were extracted using "nanopolish eventalign".

Gene ontology and sequencing logo analysis:

Gene ontology (GO) analysis of Molecular Function 2021 was performed using enrichR website [35–37]. The sequence motifs are generated by kpLogo website [38].

Modification detection and analysis

A summary of the base calls of aligned reads to the reference sequence is obtained using the *Rsamtools* package. Mismatch frequency is then calculated for a list of verified pseudouridine sites. We observe that U-to-C mismatch frequency shows a better separation between the modified (IVT) and (potentially) modified (Direct) samples (refer to the scatter plots from SI, talk about the p-value from t-test that will be included for each panel in the caption).

We know from our control sample that U-to-C mismatch frequency depends on both the molecular sequence and coverage (**Fig 2. a, b, and c**). Therefore, the significance of an observed mismatch percentage at each site is calculated accordingly and via the following equation:

$$p(N, N_{mm,dseq}, p_0) = \sum_{N_{mm}=N_{mm,dseq}}^{N} \frac{N}{N_{mm}} \times p_0^{N_{mm}} \times (1-p_0)^{N-N_{mm}},$$

where the significance of the mismatch frequency at each U site is calculated using the sequence-dependent expected error and the read coverage at that site.

Statistical analysis

All experiments were performed in multiple, independent experiments, as indicated in the figure legends. All statistics and tests are described fully in the text or figure legend.

Code availability

Scripts for all analyses presented in this paper, including all data extraction, processing, and graphing steps are freely accessible at https://github.com/RouhanifardLab/PsiNanopore.git.

Data availability

All raw and processed data used to generate figures and representative images presented in this paper are available at https://www.ncbi.nlm.nih.gov/biosample/22863220.

Supplementary Information

Supplementary figures and tables can be found at the following link:
https://www.dropbox.com/sh/psxk6ux89t4jhyd/AABaP44eGOts6CZOq_8UhwS4a?dl=0

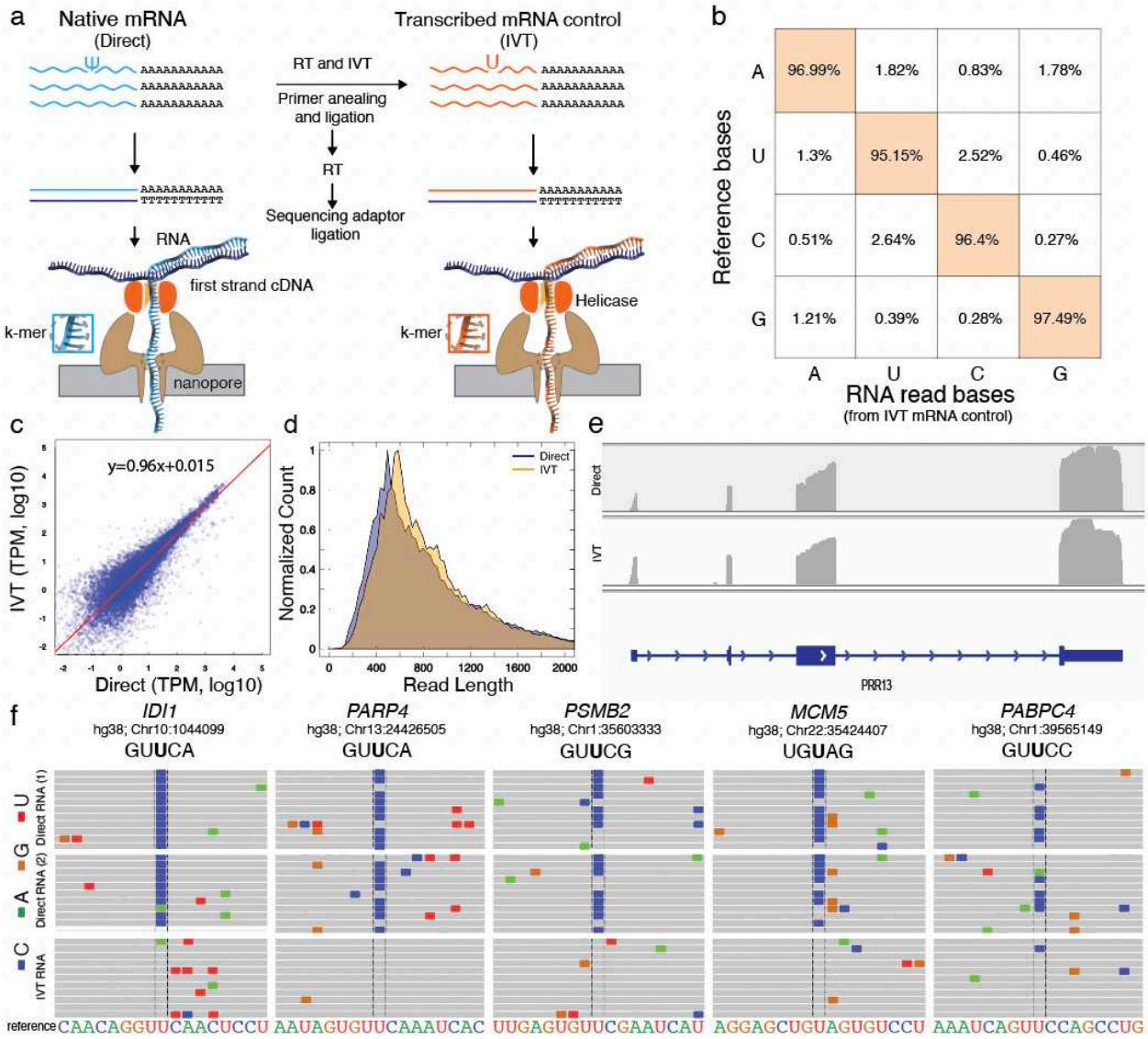
Acknowledgments

S.H.R acknowledges support from a Seed Networks Award from the Chan Zuckerberg Initiative CZF2019-002424 and NIH 5R01HG011087-02. M.W acknowledges support from NIH R01HG10087 and Oxford Nanopore Technologies. Y.H. acknowledges support from NIH GM011120.

Author Contributions

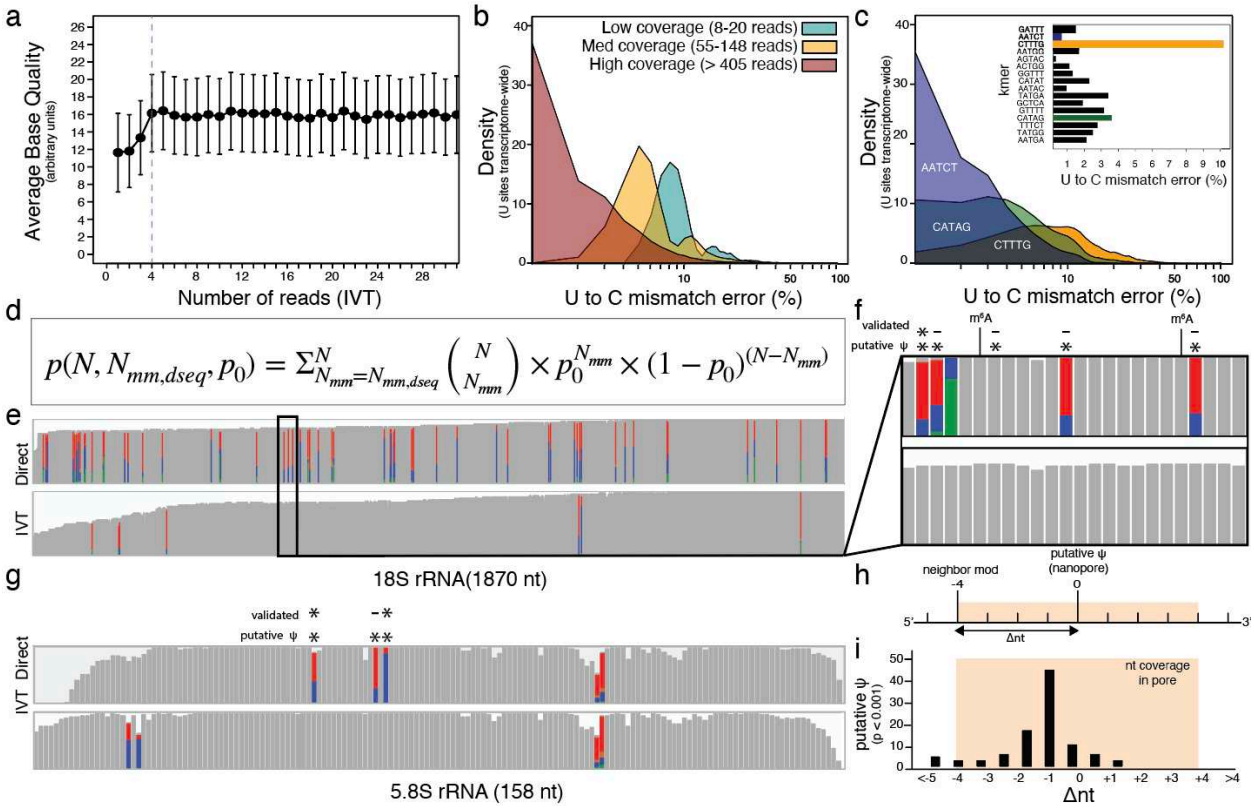
ST, MW and SHR conceived of the research. ST designed and performed the experiments. ST, MN, AM, and NR analyzed the data with guidance from MW and SHR. HG designed and synthesized synthetic RNA controls with guidance from YH. ST wrote the paper with guidance from SHR.

Figure 1:



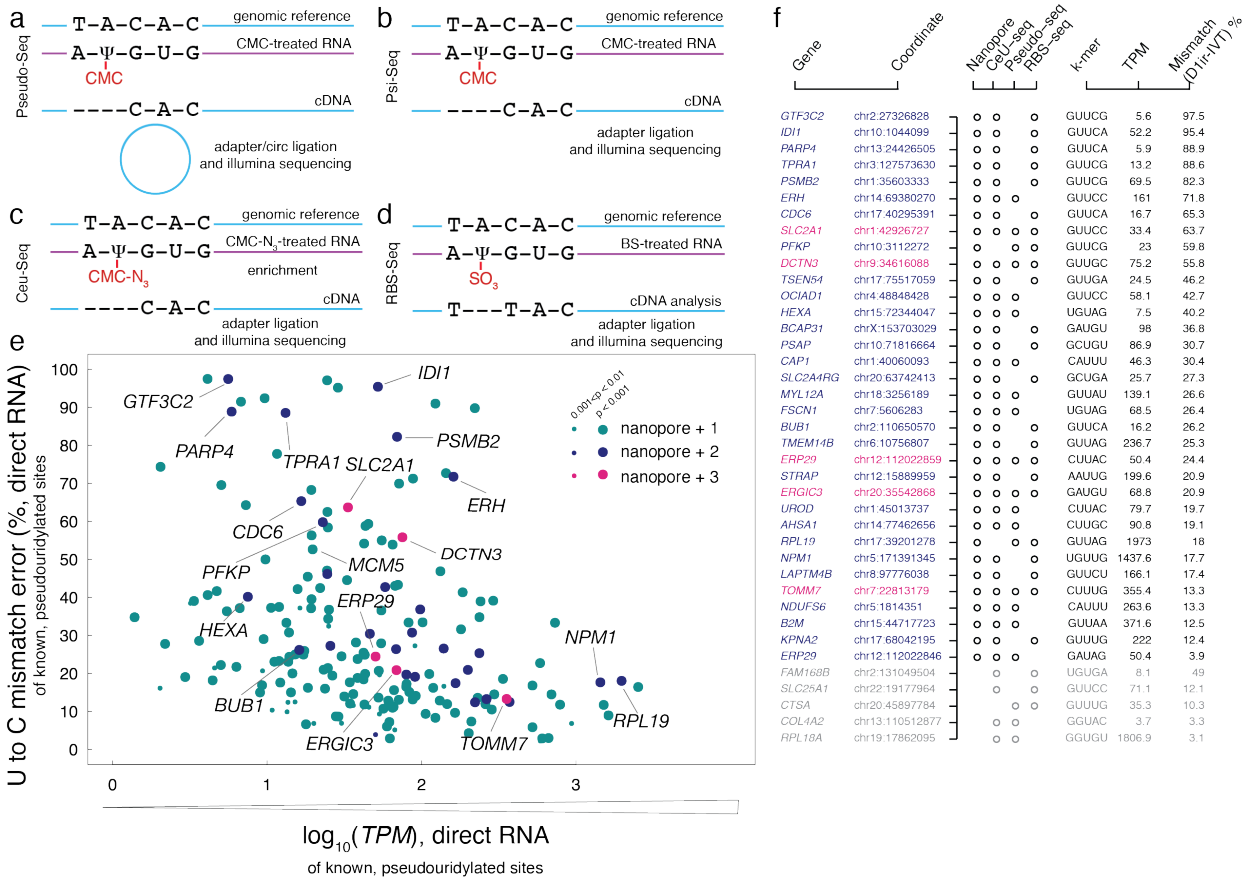
Nanopore native poly(A) RNA sequencing pipeline to identify psi-modified sites. a. Library preparation for Nanopore sequencing of native poly(A)-containing mRNAs (direct) and sequencing of in vitro transcribed (IVT) control. b. The accuracy of called bases of *in vitro* transcribed (IVT) control samples. The x-axis shows bases that are called nanopore reads and the y axis is the base identity from the reference sequence at the same position that the nanopore reads are aligned to. c. $\log_{10}$(TPM) of direct vs the $\log_{10}$(TPM) of IVT. d. Normalized count of different read lengths for direct reads (blue) vs IVT reads (orange). e. IGV snapshot of *PRR13* in direct (top) and IVT (bottom). f. Representative snapshot from the integrated genome viewer (IGV) of aligned nanopore reads to the hg38 genome (GRCh38.p10) at the pseudouridylated positions positions that have been validated by previous methods. Miscalled bases are shown in colors. Genomic reference sequence is converted to sense strand and shown as RNA for clarity.

Figure 2:



Systematic analysis of basecalling accuracy and quantification of the significance. a, average base quality for different numbers of reads using IVT reads. b, Distribution of U-to-C mismatch percentage for three populations, based on the read coverage. c, Distribution of U-to-C mismatch percentage for three populations, based on 5-mers. d, Quantification of the significance of a site based on U-to-C mismatch percentage, read coverage, and the sequence of the 5-mer of the site. e, IGV snapshot of 18S rRNA for Direct (Up) and IVT (down) f, the callout of a part of 18S rRNA region g, IGV snapshot of 5.8S rRNA for Direct (Up) and IVT (down) h, The schematic figure in which delta nt is the distance to the putative modification position. i, The histogram that shows the number of detected psi position by our method with different delta nt that shows the distance to the closest modification.

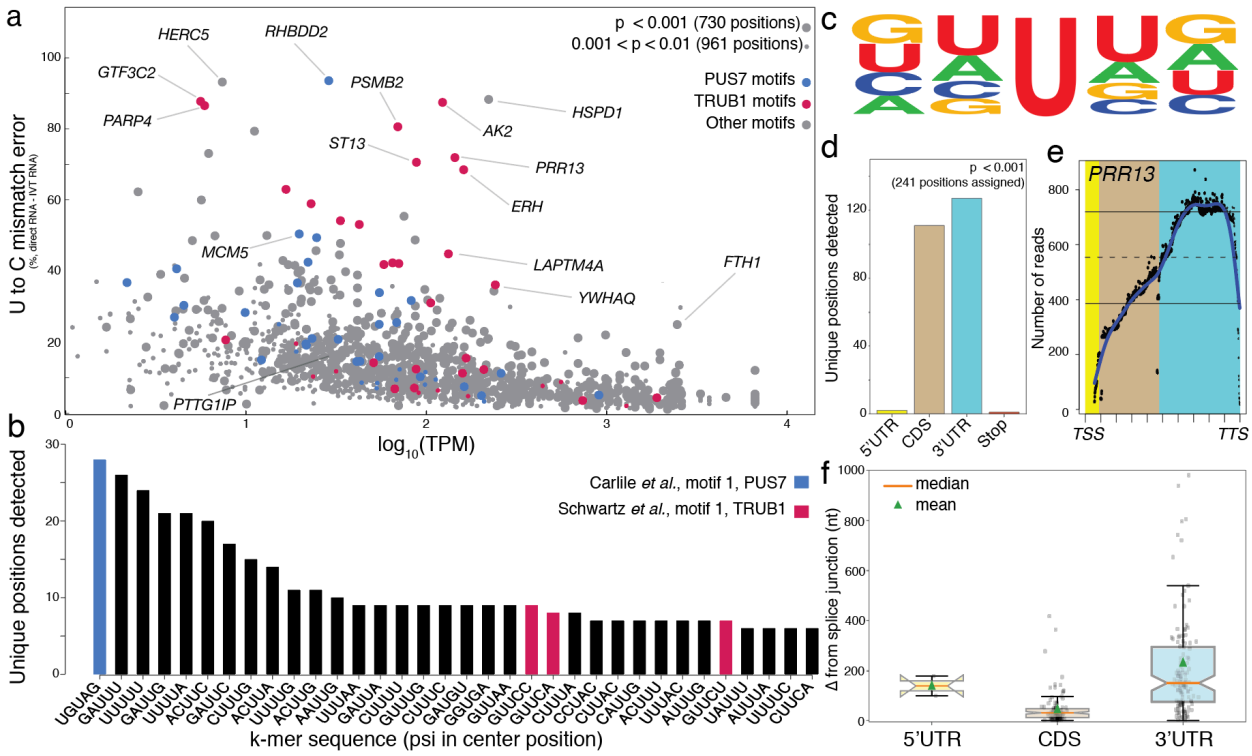
Figure 3



Gene	Coordinate	Nanopore	CeU-seq	Pseudo-seq	RBS-seq	k-mer	TPM	Mismatch (DTir-IVT) %
GTF3C2	chr2:27326828	o	o		o	GUUCG	5.6	97.5
IDI1	chr10:1044099	o	o		o	GUUCA	52.2	95.4
PARP4	chr13:24426505	o	o		o	GUUCA	5.9	88.9
TPRA1	chr3:127573630	o	o		o	GUUCG	13.2	88.6
PSMB2	chr1:35603333	o	o		o	GUUCG	69.5	82.3
ERH	chr14:69380270	o	o	o		GUUCC	161	71.8
CDC6	chr17:40295391	o	o		o	GUUCA	16.7	65.3
SLC2A1	chr1:42926727	o	o	o	o	GUUCC	33.4	63.7
PFKP	chr10:3112272	o		o	o	GUUCG	23	59.8
DCTN3	chr9:34616088	o	o	o	o	GUUGC	75.2	55.8
TSEN54	chr17:75517059	o	o		o	GUUGA	24.5	46.2
OCIAD1	chr4:48848428	o	o	o		GUUCC	58.1	42.7
HEXA	chr15:72344047	o	o	o		UGUAG	7.5	40.2
BCAP31	chrX:153703029	o	o		o	GAUGU	98	36.8
PSAP	chr10:71816664	o	o		o	GCUGU	86.9	30.7
CAP1	chr1:40060093	o	o	o		CAUUU	46.3	30.4
SLC2A4RG	chr20:63742413	o	o		o	GCUGA	25.7	27.3
MYL12A	chr18:3256189	o	o	o		GUUAU	139.1	26.6
FSCN1	chr7:5606283	o	o	o		UGUAG	68.5	26.4
BUB1	chr2:110650570	o	o		o	GUUCA	16.2	26.2
TMEM14B	chr6:10756807	o	o		o	GUUAG	236.7	25.3
ERP29	chr12:112022859	o	o	o	o	CUUAC	50.4	24.4
STRAP	chr12:15889959	o	o		o	AAUUG	199.6	20.9
ERGIC3	chr20:35542868	o	o	o	o	GAUGU	68.8	20.9
UROD	chr1:45013737	o	o	o		CUUAC	79.7	19.7
AHSA1	chr14:77462656	o	o	o		CUUGC	90.8	19.1
RPL19	chr17:39201278	o		o	o	GUUAG	1973	18
NPM1	chr5:171391345	o	o		o	UGUUG	1437.6	17.7
LAPTM4B	chr8:97776038	o	o		o	GUUCU	166.1	17.4
TOMM7	chr7:22813179	o	o	o	o	CUUUG	355.4	13.3
NDUFS6	chr5:1814351	o	o	o		CAUUU	263.6	13.3
B2M	chr15:44717723	o	o	o		GUUAA	371.6	12.5
KPNA2	chr17:68042195	o	o		o	GUUUG	222	12.4
ERP29	chr12:112022846	o	o	o		GAUAG	50.4	3.9
FAM168B	chr2:131049504		o		o	UGUGA	8.1	49
SLC25A1	chr22:19177964		o		o	GUUCC	71.1	12.1
CTSA	chr20:45897784			o	o	GUUUG	35.3	10.3
COL4A2	chr13:110512877		o	o		GGUAC	3.7	3.3
RPL18A	chr19:17862095		o	o		GGUGU	1806.9	3.1

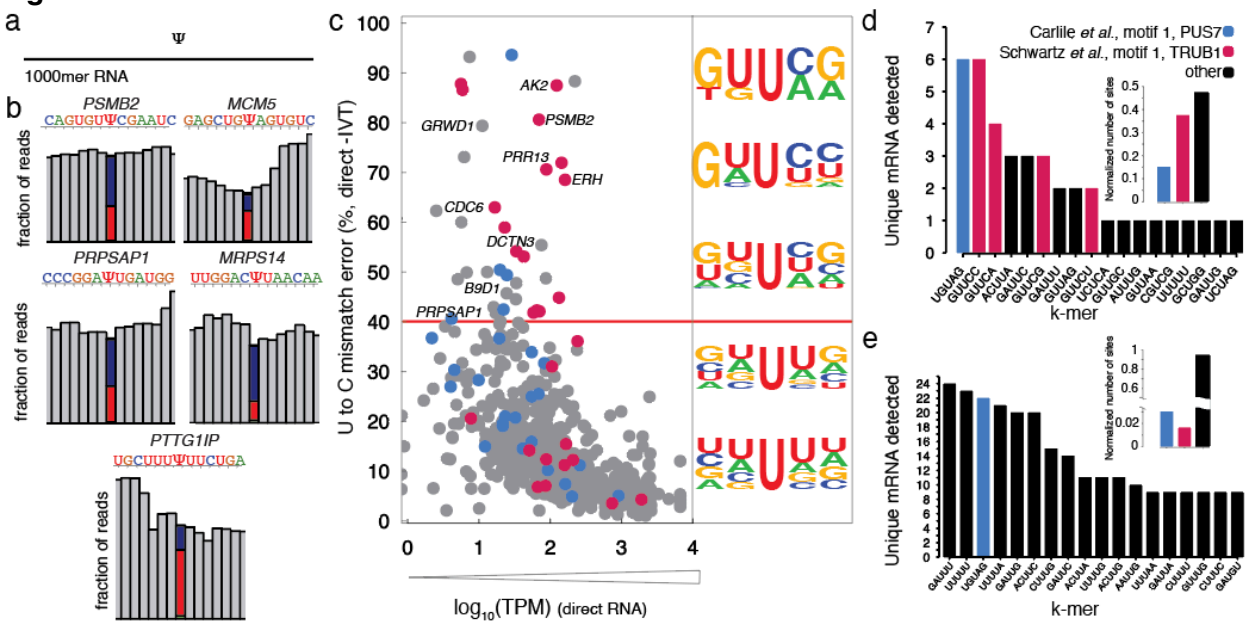
Previously detected psi modifications in the human transcriptome are validated by nanopore sequencing. a. The schematic workflow of the CMC-based methods that have detected psi modification in the human transcriptome. a. Pseudo-Seq, b. Ψ-Seq, c. CeU-Seq, and d. modified bisulfite sequencing (RBS-Seq). e. U-to-C mismatch error (%) or the merged replicates of direct RNA of known psi sites versus the log10(TPM) of merged direct RNA sequencing replicates. All targets shown are picked up by nanopore method and are validated by at least one previous method. green: validated by one previous method, blue: validated by two previous methods, magenta: validated by three previous methods, and orange: validated by four previous methods. f. The annotation of the genes containing a reported psi modification by two or more previous methods. The ones validated by nanopore sequencing with a high confidence value (*p* of both replicates < 0.001) (black) and not validated by our nanopore method (Grey).

Figure 4:



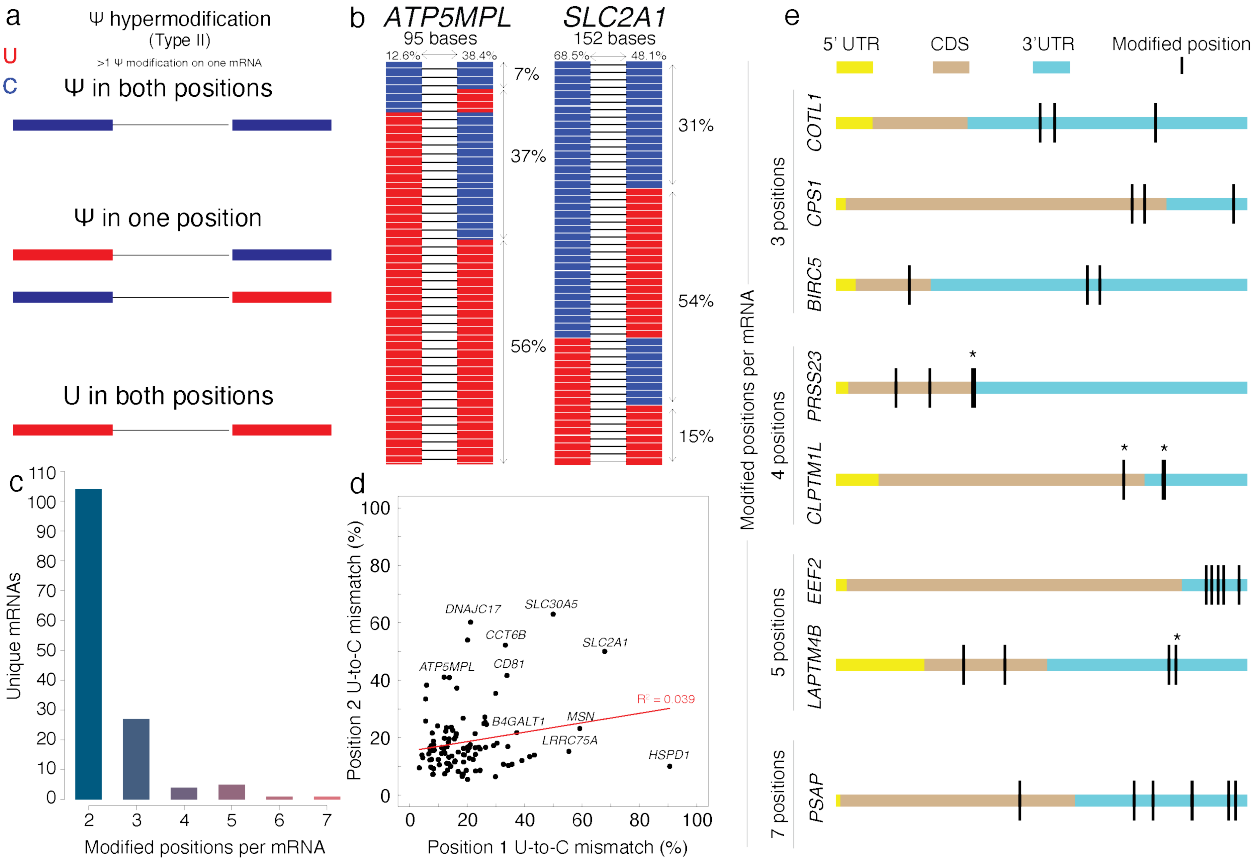
Nanopore sequencing detects the psi modification *de novo* and validates targets detected by previous methods. a. the U-to-C mismatches detected by nanopore sequencing versus the -log10(TPM) of merged direct RNA.large dot: the detected targets identified by the significance factor of two out of three replicates lower than 0.001, smaller dot: the detected targets identified by the significance factor of two out of three replicates lower than 0.01, blue: Targets with PUS7 motif, red: Targets with TRUB1 motif, and grey: Targets with the motifs other than PUS7 or TRUB1. b. The k-mer frequency of the most frequently detected targets with higher confidence. c. The sequence motif across the detected psi modification for all detected k-mers generated with kplogo[38]. d. The distribution of detected psi sites in the 5' untranslated region (5' UTR), 3' untranslated region (3' UTR), and coding sequence (CDS). e. The read depth of the reads aligned to PRR13 versus the relative distance to the transcription start site (TSS) and transcription termination site (TTS). f. The distance from the nearest splice junction of the sites detected in the 5'UTR, 3'UTR, or CDS after reads were assigned to a dominant isoform using FLAIR [34].

Figure 5:



Assessment of the ability of nanopore sequencing to detect psi sites in the human transcriptome using synthetic 1,000-mer RNA oligos. a. A pair of 1,000-mer synthetic RNA oligos were designed, one containing 100% uridine and the other containing 100% psi in a sequence that recapitulates the natural occurrence of psi in the human transcriptome. b. The frequency histograms of 13 nucleotides surrounding the detected psi position in the middle of a k-mer in 4 different mRNAs: *PSMB2*, *MCM5*, *PRPSAP1,* and *MRPS14, and PTTG1IP*. c. The U-to-C mismatches of the detected psi position for merged replicates of direct RNA seq versus -log10(significance). The targets with U-to-C mismatch of higher than 40% are defined as hypermodified type 1. The sequence motifs for different mismatch ranges are shown. d. K-mer frequency is shown for hypermodified type I and "not hypermodified" psi sites with the highest occurrence. e. Distribution of U-to-C mismatches higher than 40% in mRNA regions.

Figure 6:



Type II hypermodification is defined as the mRNA targets that contain two or more psi positions. a. Schematic figure of hypermodified type II which contains 2 psi positions. b. The histograms of hypermodified type II positions contain 2 to 7 psi nucleotides. c. The U-to-C mismatch of the position 1 versus position 2 of the hypermodified target contains two detected psi positions. d. Two examples of hypermodified type II with two detected psi positions indicating mismatch in a single read for the reads that cover both positions. e. Examples of the hypermodified type II with three or more psi positions distributed across each gene.

References

1. Roundtree, I. A., Evans, M. E., Pan, T. & He, C. Dynamic RNA Modifications in Gene Expression Regulation. *Cell* **169**, 1187–1200 (2017).
2. Taoka, M. *et al.* Landscape of the complete RNA chemical modifications in the human 80S ribosome. *Nucleic Acids Res.* **46**, 9289–9298 (2018).
3. Li, X. *et al.* Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. *Nat. Chem. Biol.* **11**, 592–597 (2015).
4. Mellis, I. A., Gupte, R., Raj, A. & Rouhanifard, S. H. Visualizing adenosine-to-inosine RNA editing in single mammalian cells. *Nat. Methods* **14**, 801–804 (2017).
5. Spitale, R. C. *et al.* Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519**, 486–490 (2015).
6. Wang, X. *et al.* N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* **505**, 117–120 (2014).
7. Hunter, S. *et al.* InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* **40**, D306-12 (2012).
8. Cohn & Elliot, W. Nucleoside-5′-Phosphates from Ribonucleic Acid. *Nature* 483–484 (1951).
9. Anderson, B. R. *et al.* Nucleoside modifications in RNA limit activation of 2′-5′-oligoadenylate synthetase and increase resistance to cleavage by RNase L. *Nucleic Acids Res.* **39**, 9329–9338 (2011).
10. Price, A. M. *et al.* Direct RNA sequencing reveals m6A modifications on adenovirus RNA are necessary for efficient splicing. *bioRxiv* 865485 (2019) doi:10.1101/865485.
11. Karikó, K., Buckstein, M., Ni, H. & Weissman, D. Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA.

Immunity **23**, 165–175 (2005).

12. Anderson, B. R. *et al.* Incorporation of pseudouridine into mRNA enhances translation by diminishing PKR activation. *Nucleic Acids Res.* **38**, 5884–5892 (2010).
13. Eyler, D. E. *et al.* Pseudouridinylation of mRNA coding sequences alters translation. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 23068–23074 (2019).
14. Schwartz, S. *et al.* Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell* **159**, 148–162 (2014).
15. Carlile, T. M. *et al.* Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature* **515**, 143–146 (2014).
16. Kirwan, M. & Dokal, I. Dyskeratosis congenita, stem cells and telomeres. *Biochim. Biophys. Acta* **1792**, 371–379 (2009).
17. Agarwal, S. *et al.* Telomere elongation in induced pluripotent stem cells from dyskeratosis congenita patients. *Nature* **464**, 292–296 (2010).
18. Charette, M. & Gray, M. W. Pseudouridine in RNA: what, where, how, and why. *IUBMB Life* **49**, 341–351 (2000).
19. Mengel-Jørgensen, J. & Kirpekar, F. Detection of pseudouridine and other modifications in tRNA by cyanoethylation and MALDI mass spectrometry. *Nucleic Acids Res.* **30**, e135 (2002).
20. Addepalli, B. & Limbach, P. A. Mass spectrometry-based quantification of pseudouridine in RNA. *J. Am. Soc. Mass Spectrom.* **22**, 1363–1372 (2011).
21. Ho, N. W. & Gilham, P. T. Reaction of pseudouridine and inosine with N-cyclohexyl-N'-beta-(4-methylmorpholinium)ethylcarbodiimide. *Biochemistry* **10**, 3651–3657 (1971).
22. Khoddami, V. *et al.* Transcriptome-wide profiling of multiple RNA modifications simultaneously at single-base resolution. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 6784–6789 (2019).

23. Safra, M., Nir, R., Farouq, D., Vainberg Slutskin, I. & Schwartz, S. TRUB1 is the predominant pseudouridine synthase acting on mammalian mRNA via a predictable and conserved code. *Genome Res.* **27**, 393–406 (2017).

24. Li, X., Ma, S. & Yi, C. Pseudouridine: the fifth RNA nucleotide with renewed interests. *Curr. Opin. Chem. Biol.* **33**, 108–116 (2016).

25. Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).

26. Liu, H. *et al.* Accurate detection of m6A RNA modifications in native RNA sequences. *Nat. Commun.* **10**, 4079 (2019).

27. Smith, A. M., Jain, M., Mulroney, L., Garalde, D. R. & Akeson, M. Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. *PLoS One* **14**, e0216709 (2019).

28. Begik, O. *et al.* Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-021-00915-6.

29. Huang, S. *et al.* Interferon inducible pseudouridine modification in human mRNA by quantitative nanopore profiling. *Genome Biol.* **22**, 330 (2021).

30. Fleming, A. M., Mathewson, N. J., Howpay Manage, S. A. & Burrows, C. J. Nanopore Dwell Time Analysis Permits Sequencing and Conformational Assignment of Pseudouridine in SARS-CoV-2. *ACS Cent. Sci.* (2021) doi:10.1021/acscentsci.1c00788.

31. Pyle, A. M. Translocation and unwinding mechanisms of RNA and DNA helicases. *Annu. Rev. Biophys.* **37**, 317–336 (2008).

32. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

33. Carlile, T. M. *et al.* mRNA structure determines modification by pseudouridine synthase 1. *Nat. Chem. Biol.* **15**, 966–974 (2019).

34. Tang, A. D. *et al.* Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun.* **11**, 1438 (2020).

35. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).

36. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90-7 (2016).

37. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Curr Protoc* **1**, e90 (2021).

38. Wu, X. & Bartel, D. P. kpLogo: positional k-mer analysis reveals hidden specificity in biological sequences. *Nucleic Acids Res.* **45**, W534–W538 (2017).