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

Enzyme-mediated chemical 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 modification on mRNAs remain unclear due to the lack of sensitive and accurate tools for mapping. We developed a semi-quantitative method for mapping pseudouridylated sites with high confidence directly on mammalian mRNA transcripts via direct RNA, long-read nanopore sequencing. By analysis of a modification-free transcriptome, we demonstrate that the depth of coverage and intrinsic errors associated with specific k-mer sequences are critical parameters for accurate base-calling. We adjust these parameters for high-confidence U-to-C base-calling errors that occur at pseudouridylated sites, which are benchmarked against sites that were identified previously by biochemical methods. We also uncovered new pseudouridylated sites, many of which fall on genes that encode RNA binding proteins and on uridine-rich k-mers. Sites identified by U-to-C base calling error were verified using 1000mer synthetic RNA controls bearing a single pseudouridine in the center position, demonstrating that 1. the U-to-C base-calling error occurs at the site of pseudouridylation, and 2. the basecalling error is systematically under-calling the pseudouridylated sites. High-occupancy sites with >40% U-to-C basecalling error are classified as sites of hyper modification type I, whereas genes with more than one site of pseudouridylation are classified as having type II hyper modification which is confirmed by single-molecule analysis. We report the discovery of mRNAs with up to 7 unique sites of pseudouridine modification. Here we establish an innovative pipeline for direct identification, quantification, and detection of pseudouridine modifications and type I/II hypermodifications on native RNA molecules using long-read sequencing 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]; 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[2], these modifications potentially impact gene expression[3], RNA tertiary structures[4], or the recruitment of RNA-binding proteins[5]. Pseudouridine (psi) is synthesized from uridine converted *in vivo* by one of more than a dozen pseudouridine synthases identified to date[6]**.** It was the first discovered RNA modification[7] and represents 0.2-0.6% of total uridines in mammalian mRNAs[2]. Psi-modified mRNAs are more resistant to RNase-mediated degradation[8] and also have the potential to modulate splicing[9] and immunogenicity[10] and enhance translation[11] *in vivo*. Further, psi modifications of RNAs are responsive to cellular

stress, leading to increased RNA half-life[12,13]. 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[14,15]. 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[16], 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[17,18].

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[19]. Treatment with a strong base removes the CMC from all of 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[20]. Indirect chemical labeling of psi combined with next-generation sequencing[2,13,20] has yielded over 2,000 putative psi sites within mammalian mRNAs, but different methods identified different sites that have a limited overlap[21], pointing to a need for improved 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[22]. 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 report using nanopore-based direct RNA sequencing [23,24,25,26] to directly read RNA modifications. In these reports, ion current differences for different k-mer sequences (k = 5 bases) as an RNA strand is moved through the pore hint at the presence of a modified RNA base. Detection of psi using nanopores was also confirmed for rRNAs[25], for the *Saccharomyces cerevisiae* transcriptome[27], and for viralRNAs[28], as indicated by a U-to-C base-calling error at various sequence sites. Analysis of human cell transcriptomes using this method is challenging because of the high input mRNA requirements (>500 ng). Further, quantifying the occupancy at a given modified site depends largely on the nucleotide sequence surrounding the modification, and requires controls that match the transcriptome sequence beyond the context of the measured k-mer (k = 5). The 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[29]. Nonetheless, an accurate pipeline for assessing mRNA modifications in human transcriptomes could guide our understanding of the roles of these modifications in regulating gene expression.

Here, we describe a nanopore-based method to accurately map psi modifications on a HeLa transcriptome by comparing them 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. Our approach recapitulates 122, previously annotated psi sites, thus providing a "ground truth" list of psi modifications that have been validated by independent methods. Our approach also reveals 1,942 putative psi sites that have not previously been reported. We show that these new sites tend to occur within transcripts that encode RNA binding proteins and in uridine-rich

k-mer sequences including the PUS7[30] and TRUB[12] sequence motifs that were previously reported.

We validate the accuracy of the U-to-C mismatch error as a proxy for psi modifications by analysis of 4 synthetic RNA 1000-mers. Each of these synthetic oligos contains 100% uridine or 100% psi at a known pseudouridylated site in the human transcriptome. The analysis reveals that U-to-C mismatch errors are systematically under-called for the detection of psi. Using a base-calling error cutoff, we identify 105 high-occupancy, hypermodified psi sites, which are likely to confer a measurable phenotype. 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[21].

Finally, we identify 38 mRNAs with more than two high-confidence psi sites, which are confirmed by single-molecule 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. Further, the long-read lengths of the nanopore method allow 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

Previous studies have shown that psi modifications can be detected using direct RNA nanopore sequencing[24,27] by monitoring the statistics of uridine basecalling errors in an ensemble of reads from similar transcripts. However, the accuracy of basecalling errors as a proxy measurement for psi modification has yet to be determined in the context of native human mRNA sequences. Lack of this information has precluded the ability to obtain a precise transcriptome-wide map of psi-modifications in a sample. A critical negative control for this analysis is to generate an identical mRNA library with no RNA modifications[23]. To measure the effects of mRNA psi modifications, we extracted RNA from HeLa cells and prepared two libraries: The first consists of native mRNAs (Direct) which contain both canonical uridine and naturally occurring uridine modifications. The second consists of an *in vitro* transcribed mRNA control (IVT) 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[23] (**Fig. 1a**). Each library was sequenced on a separate Minion flowcell and basecalled using *Guppy 3.2.10*. Two direct runs produced 848,000 and 1,002,813 poly(A) RNA strand reads, respectively, of which 724,000 and 821,879 reads passed quality filters (read quality of 7), with a read N50 length (defined as the shortest read length needed to cover 50% of the sequenced nucleotides) of 834 ± 50.09 bases and a median length of 624.3 ± 39.3 bases (**Supplementary Fig. 1**). Similarly, IVT runs produced 1,822,844 reads of which 1,330,412 passed the quality filter, with N50 of 854 and a median length of 666 bases. Alignment was performed using minimap2.17[31] and the reads for the 1st replicate (573,547), 2nd replicate (659,727), and IVT (1,007,597) were subsequently aligned to the GRCh38 human genome reference.

Utilizing basecalling accuracy to identify psi modifications in RNA

To explore differences between the IVT and Direct RNA samples 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 a wrong alignment by only considering the primary alignment of each read (i.e., the alignment with the highest mapping quality; **Supplementary Fig. 2**). Also, any read with a mapping quality score lower than 20 was disregarded for the downstream analysis, because the probability of the alignment being correct was lower than 99%. Further, this cutoff choice allowed us to retain the maximum number of reads without observing significant mismatch error.

The second source of a mismatch 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. 3**), 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. 4**).

The third and most significant source of error is erroneous basecalling, 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 considered to be independent of RNA modifications. We confirmed that the basecaller could reliably identify unmodified and aligned nucleotides with an average error of 2.64%.

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. We selected 5 genes as examples: *IDI1* (chr10:1044099)[2,12,20], *PARP4* (chr13:24426505)[2,20], *PSMB2* (chr1:35603333)[2,12,20], *MCM5* (chr22:35424407) [2,12], and *PABPC4* (chr1:39565149)[2,12], 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-mer sequences was chosen because specific k-mer sequences can influence the accuracy of base-calling (**Supplementary Fig. 5**). 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**).

Alterations to the distribution of k-mer current signals correspond to mRNA psi modification for positions with >80% U-to-C mismatch error.

To investigate if psi sites may be detected by ionic current distribution, we extracted the current traces using Nanopolish[32] and systematically analyzed the ionic signal from the direct RNA library for the "representative subset" of biochemically validated psi sites (**see Methods**). We observed a shift in the current distribution of most targets (**Fig. 1d**, **Supplementary Fig.7).** Of the targets with >80% U-to-C mismatch error, we observed a visible shift in the current

distribution for 7 out of 8 targets when compared to the respective unmodified, IVT control samples (**Supplementary Fig. 6**). However, we did not observe this alteration in current distribution for targets with a U-to-C mismatch error of <80%. For example, for the psi site on *MCM5* (chr22:35424407,k-mer: UGUAG) (**Fig. 1c, d**), the U-to-C mismatch error is around 50%, and the current distribution for the direct RNA read and IVT controls do not show any obvious differences. We also observed that the current distribution could happen at a few nucleotides away from the modified nucleotides (*RHBDD2* (chr7:75888787), k-mer: UGUAG, **Supplementary Fig. 6**). Overall, the U-to-C mismatch is a more reliable indicator of putative psi sites than current distribution analysis.

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**). 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. 8**).

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. Therefore, the significance of the U-to-C mismatch percentage of a site must be interpreted based on a combination of the mismatch percentage, number of reads, and the surrounding nucleotides in the k-mer. Analysis of the significance of a U-to-C mismatch at a given position (**Fig. 2d**) showed that, regardless of the sequence, the significance of a mismatch frequency is an ascending function of coverage. Also, low-error k-mers yield higher significance (**Fig. 2e**). For example, higher significance values are calculated for the AAUCU k-mer (low-error sequence, blue), compared to CUUUG k-mer (high-error sequence, orange). To ensure that the targets are not selected based on the mismatches from other sources like single-nucleotide polymorphisms, basecalling, or alignment, we consider both the IVT mismatch percentage and k-mer based error in the calculation of significance.

Benchmarking of putative, psi sites with high significance against existing methods. Previous studies have identified putative psi sites on human mRNA using biochemical methods including CMC[2,13, 12] and RNA bisulfite[20] (**Fig. 3a-d**). We compared the accuracy of using the U-to-C mismatch error with our significance calculation from direct RNA nanopore sequencing in identifying psi sites using 759 validated mRNA targets. We selected these targets, which were previously annotated by one or more biochemical methods and also produced at least 7 reads by nanopore sequencing. Of these, 686 were validated by one other method and 73 sites were validated by two or more methods[20](**Supplementary Table 1**). To assess the significance of

each of these sites based on the number of reads and k-mer, we plotted the *p* values of the validated positions versus the U-to-C mismatch error for the psi site with $p < 0.01$ (**Fig. 3e**).

Of the 73 validated targets, 69 of them contained a higher U-to-C mismatch error in the direct RNA reads than in IVT reads (94.5%), indicating that the mismatch error corroborates well with existing methods. We defined 53 of the 73 targets as "ground truth" since they have sufficient coverage; these sites constitute ~72.6% of the target sites (**Fig. 3f**).

We benchmarked against 4 independent methods, CeU-seq[2], Pseudo-seq[13], Ψ-seq[33], and RBS-seq[20], and found that the protein-coding targets detected by Pseudo-seq have the highest overlap with nanopore detection, showing an overlap of 26/60 (~43.3%; **Supplementary table 1**). The substantial overlap may reflect the fact that both methods probe RNA from HeLa cells. Differences with other methods include the use of other human cell lines[2,33] that may have differential expression of psi sites as well as occupancy, and the inclusion of an enrichment step[2] that has the potential to unevenly amplify very low occupancy sites.

Detection of putative psi sites of mRNA *de novo* using direct RNA nanopore sequencing.
Next, we sought to apply our significance cutoff ($p < 0.01$) for *de novo* detection of putative, pseudouridylated sites, transcriptome-wide. To ensure that the source of the mismatch error is from the direct RNA read, we selected targets with significant mismatch error in the direct RNA reads ($p < 0.01$) and low mismatch error in IVT ($p > 0.1$). To confirm that the *p-value* cutoff was correct in excluding single-nucleotide polymorphisms (SNPs), we extracted genomic DNA and performed Sanger sequencing on a few selected targets with high mismatch error ($p < 0.001$) in IVT (control). We confirmed that high mismatch error in both IVT and direct is indicative of SNPs and also the presence of error-prone k-mers (**Supplementary Fig 4**). Using our algorithm, we detected 2064 putative psi sites ($p < 0.01$), including 817 positions with a *p-value* cutoff of 0.001 for both replicates (**Fig. 4a, Supplementary Table 2**). Gene ontology analyses (GO Molecular Function 2021) were performed on genes with $p < 0.001$ using enrichR website [34–36], showing that the "RNA binding" group has the highest normalized percentage of these genes (**Fig. 4b**) (**Supplementary Table 5**).

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. 4c**) and found that, as expected, UGUAG which is the motif for PUS7 binding[30] and GUUCN k-mer, the motif for TRUB1[21], are among the most frequently detected targets. To assess the sequence conservation of nucleotides within k-mers bearing a psi site in the center position, we grouped all of the highly represented k-mers and found that the +1 and -1 positions had a higher preference for the uridine nucleotide however the +2 and -2 positions do not show any nucleotide preference (**Fig. 4d**).

Distribution of psi sites on mature 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 coding sequence (CDS), with very few targets detected in the 5' UTR (**Fig. 4e**). The limited detection of psi sites in the 5' UTR is due to the low coverage that is observed 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). 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[23]. 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. 4f**).

We then calculated the distance of the detected psi target from the splice site was calculated for high confidence targets. Prior to extracting the distance of the nearest splice junction for each target, the RNA isoform analysis tool, FLAIR[37], 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) (**Fig. 4g**). Taking into account 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 (**Fig. 4h**) as compared to 3' UTR-positioned targets (**Fig. 4i**).

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 is reliably detecting psi sites *de novo* and to explore the quantitative accuracy of the U-to-C mismatch error as a proxy for pseudouridylation, we constructed 4, 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[2,12,20] and *MCM5;* chr22:35424407[2,12]) were detected from two or more previous methods and the other two targets (*MRPS14;* chr1:175014468 and *PRPSAP1;* chr17: 76311411) were detected *de novo* using the U-to-C mismatch error and our *p-value* cutoff. We constructed an unmodified 1,000-mer (100% uridine) as well as a 1,000-mer, where the center uridine position was replaced with psi (100% psi). For each gene, we ran the 0 and 100% modified versions 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 77.6% U-to-C mismatch error for *PSMB2*, 57.1% for *MCM5*, 62.0% for *PRPSAP1* and 78.1% for *MRPS14* an average of 2.4% U-to-C mismatch error for the unmodified control samples in the same positions (**Fig. 5b**). The results are indicative of a systematic under-calling of psi based on U-to-C mismatch error.

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 and at 40%, representing half-modified transcripts is at a maximum. From our *de novo* psi detection analysis, we identified 105 unique sites of hypermodification type I including *POGK (chr1:166854177), GTF3C3(chr2:196789267)*, *NIP7(chr16:69342144)*, *IDI1(chr10:1044099)*, *RHBDD2 (chr7:75888787)* 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 above 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 >40% (**Fig. 5d**) and found that the GUUCN k-mer, the motif for TRUB1[21] represents the most targets (30/105 sites around 29%). The k-mer UGUAG, the motif for PUS7 binding[30], 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%). These latter k-mers were unique and were not motif to a specific enzyme. Possibly, they are recognized for pseudouridylation through a secondary structure that they reside in by enzymes such as PUS1[30] but not by enzymes that recognize a specific motif. 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[21].

Using the results from the analysis in **Fig. 4e-i,** we found that type I hypermodified sites are biased towards 3' UTRs, which is the same as sites that are not hypermodified (**Fig. 5e**). Out of the 105 type I hypermodified sites found, 71 were assigned to an annotated isoform with high confidence. 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 (**Fig. 5e**).

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 on 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 each individual 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, *SLC2A1* has a 68.5% mismatch in position 1 (chr1:42926727) and 48.1% mismatch in position 2 (chr1:42926879) (31% on both, 54% on only one of them, 15% on none). Similarly, *ATP5MPL* has 12.6% mismatch in position 1 (chr14:103912536) and 38.4% mismatch in position 2 (chr14:103912631), 7 % on both, 37% on only one of them, 56% on none.

Discussion

We have shown here that systematic U-to-C basecalling error from direct RNA nanopore sequencing of transcriptomes can serve as a proxy for detecting the presence of psi at a given position, although the total number of reads as well as the systematic error associated with the specific canonical (unmodified) k-mer must be taken into account. Prior to this work the transcriptome-wide identification of psi sites in human mRNA was based primarily on CMC modification of psi sites, which had not been independently tested using non CMC-based methods. Here, we provide a foundation for identifying psi sites with high confidence using both a critical, unmodified transcriptome as a negative control that distinguishes standard basecalling errors that occur in unmodified k-mers, in combination with a set of synthetic

controls that demonstrate the limitations of the current basecalling algorithms for calling RNA modifications.

We demonstrated 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 of 122 mRNA positions with psi modifications in HeLa cells that have been validated by multiple, independent methods--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 novel sites of psi modification, which often occur on U-rich k-mer sequences and typically on genes that encode RNA binding proteins.

Among the methods that we used to validate our data, Pseudo-seq shows the highest overlap between the detection targets. However, more than half of the targets that the Pseudo-seq method detected were not detected by our method. We conjecture that several 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. The only way to address this is with a quantitative method for defining the occupancy of psi at a given site. On the other hand, our conservative method might lead to some false negative targets. 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.

We have validated our method by analysis of four 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. We show for the first time that the 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). Similar controls have previously been generated[27,28], however, 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. Although preparation of such a large set of control molecules is not feasible for any single laboratory, it is more and more 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, we conservatively draw a list of sites that are pseudouridylated with high frequency 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 analysis and mapping of psi modifications on mRNAs with 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, FisherSientific) 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.

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. Two replicates were from difference passages and different flow cells were used for each replicate.

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 [34–36]. 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

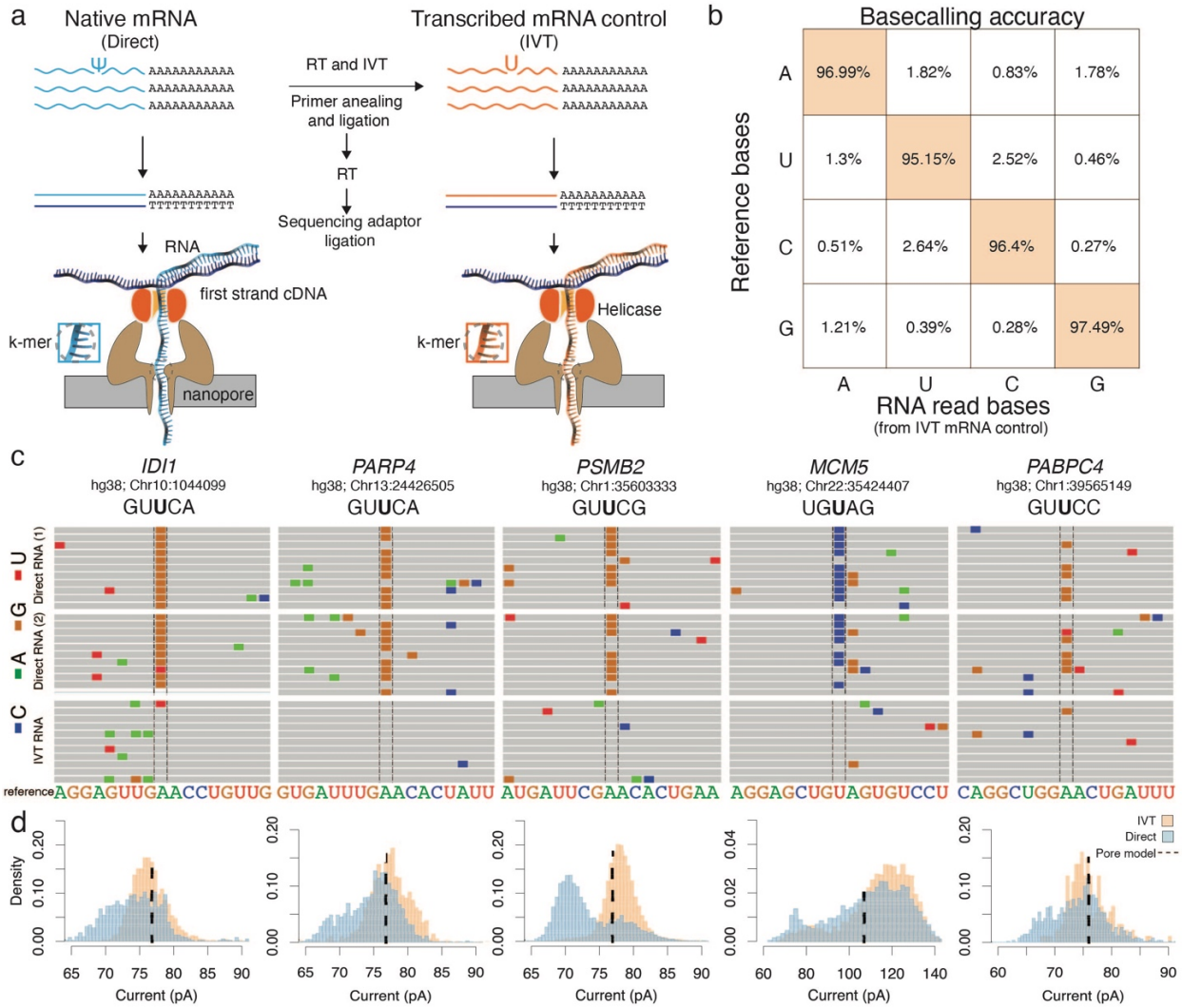
Acknowledgments

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

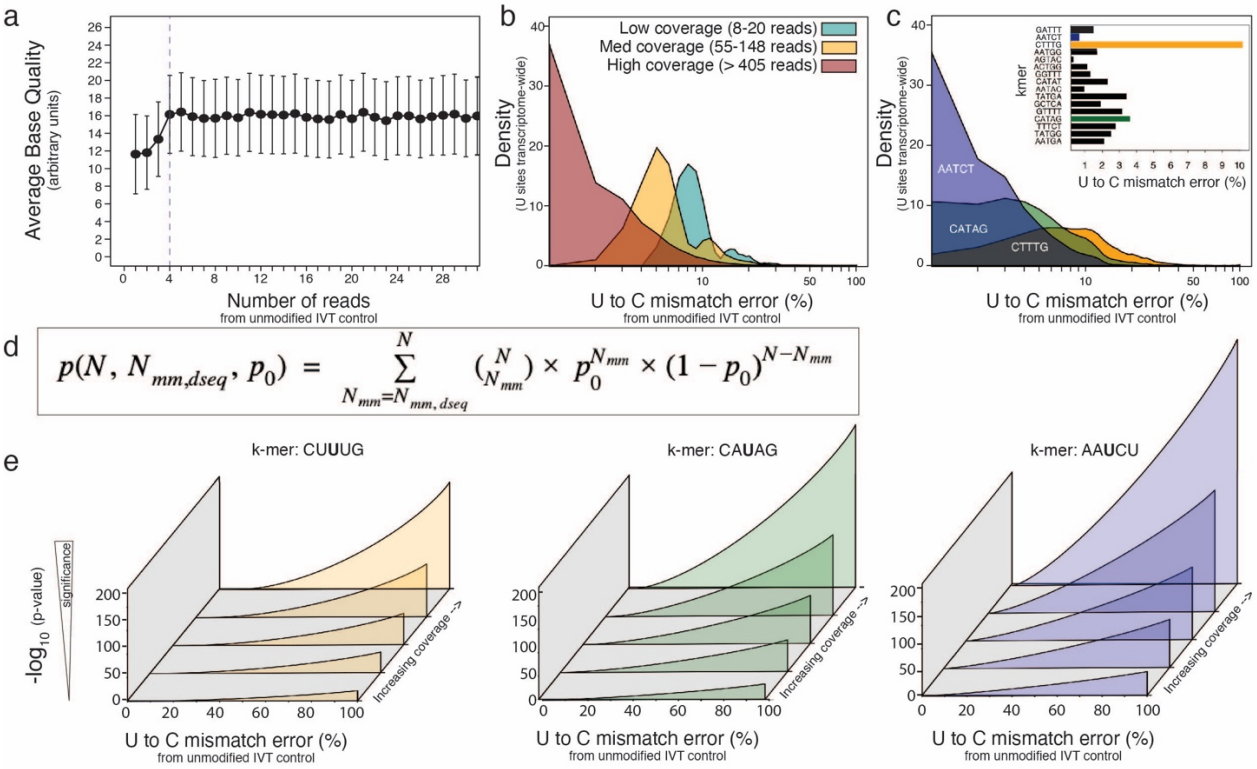
ST, MW and SHR conceived of the research. ST designed and performed the experiments. ST, MN, AM, and NR analyzed the data with guidance from MW and SHR. HG designed and synthesized synthetic RNA controls with guidance from YH. ST wrote the paper with guidance from SHR.

Figure 1:



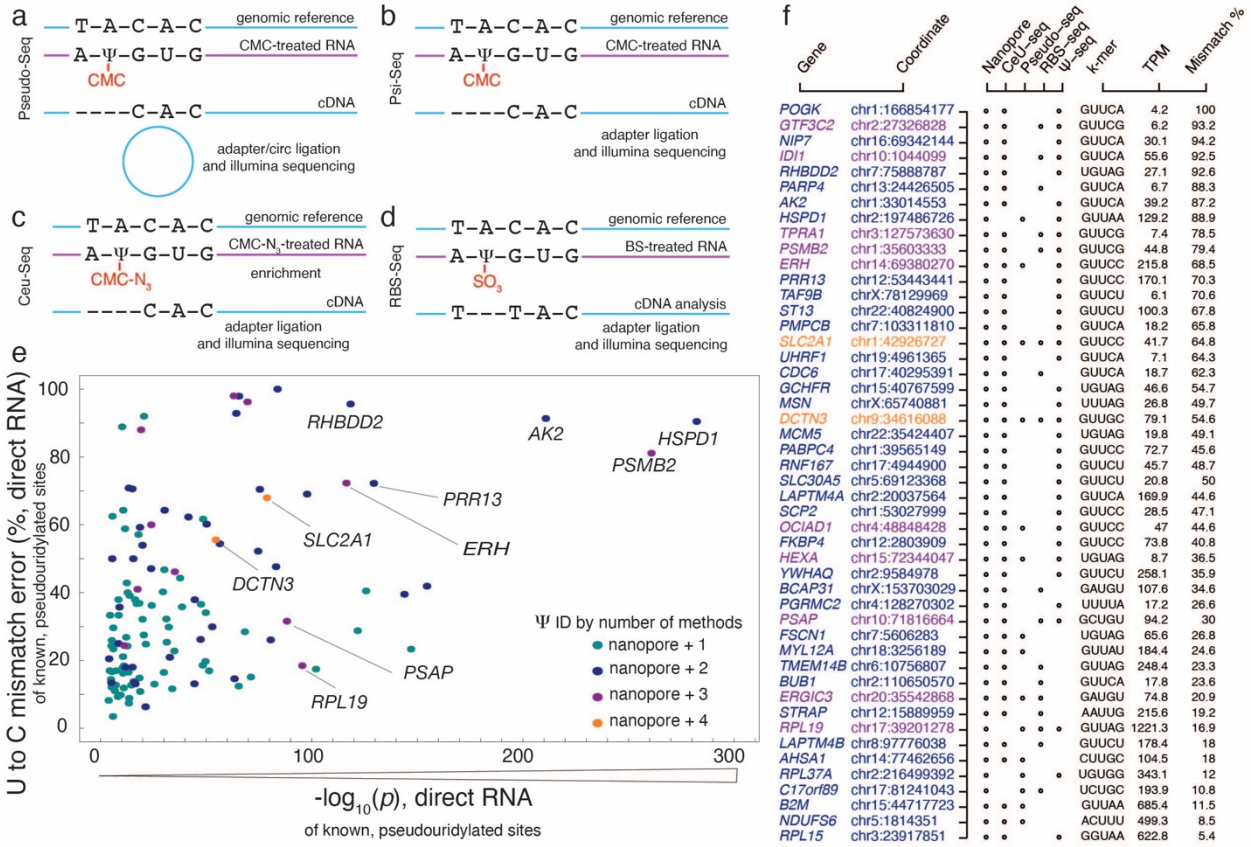
Nanopore native poly(A) RNA sequencing pipeline to identify psi-modified sites. a. Library preparation for Nanopore sequencing of native poly(A)-containing mRNAs (direct) and sequencing of in vitro transcribed (IVT) control. b. The accuracy of called bases of *in vitro* transcribed (IVT) control samples. 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, Representative snapshot from the integrated genome viewer (IGV) of aligned nanopore reads to the hg38 genome (GRCh38.p10) at the positions that have been validated by previous methods. Miscalled bases are shown in colors. As *IDI1*, *PARP4*, *PSMB2*, and *PABPC4* are aligned to the negative strand, the mismatches are shown as G instead of C.d, Comparing the ionic current signals of the Direct RNA sequencing (blue) and IVT control (orange) of the same targets in panel c. Dashed line indicates the model k-mer.

Figure 2:



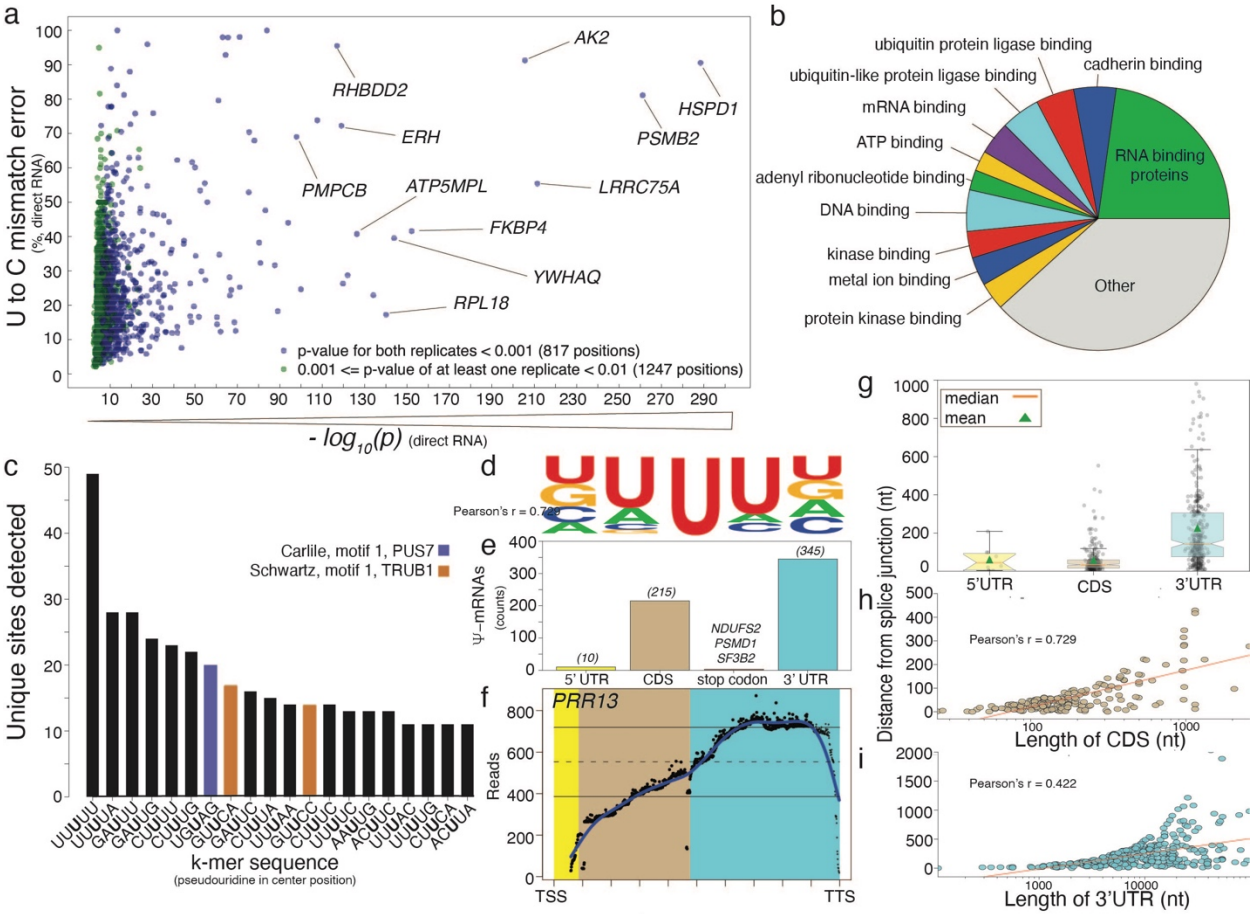
Systematic analysis of basecalling accuracy and quantification of the significance. a, average base quality for different numbers of reads using IVT reads. b, Distribution of U-to-C mismatch percentage for three populations, based on the read coverage. c, Distribution of U-to-C mismatch percentage for 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, Significance for different U-to-C mismatch percentages, and read coverages, for a high-error k-mer (CUUUG), medium-error k-mer (CAUAG) and a low-error k-mer (AAUCU).

Figure 3



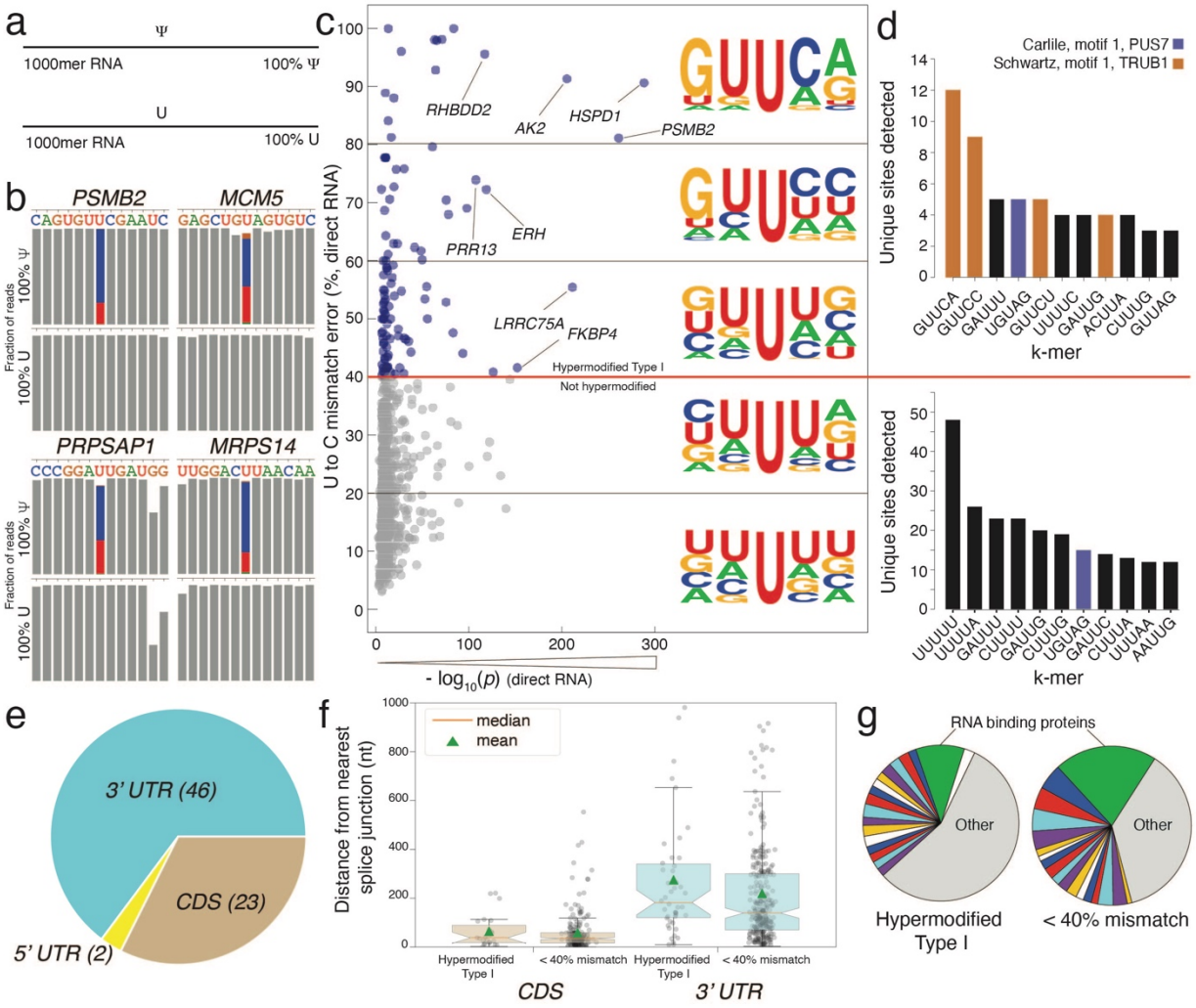
Previously detected psi modifications in the human transcriptome are validated by nanopore sequencing. a. The schematic workflow of the CMC-based methods that have detected psi modification in the human transcriptome. a. Pseudo-Seq, b. Ψ-Seq, c. CeU-Seq, and d. modified bisulfite sequencing (RBS-Seq). e. U-to-C mismatch error (%) or the merged replicates of direct RNA of known psi sites versus the -log10(significance) 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 validated by two or more previous methods and also detected by nanopore sequencing with a high confidence value (p of both replicates < 0.001).

Figure 4:



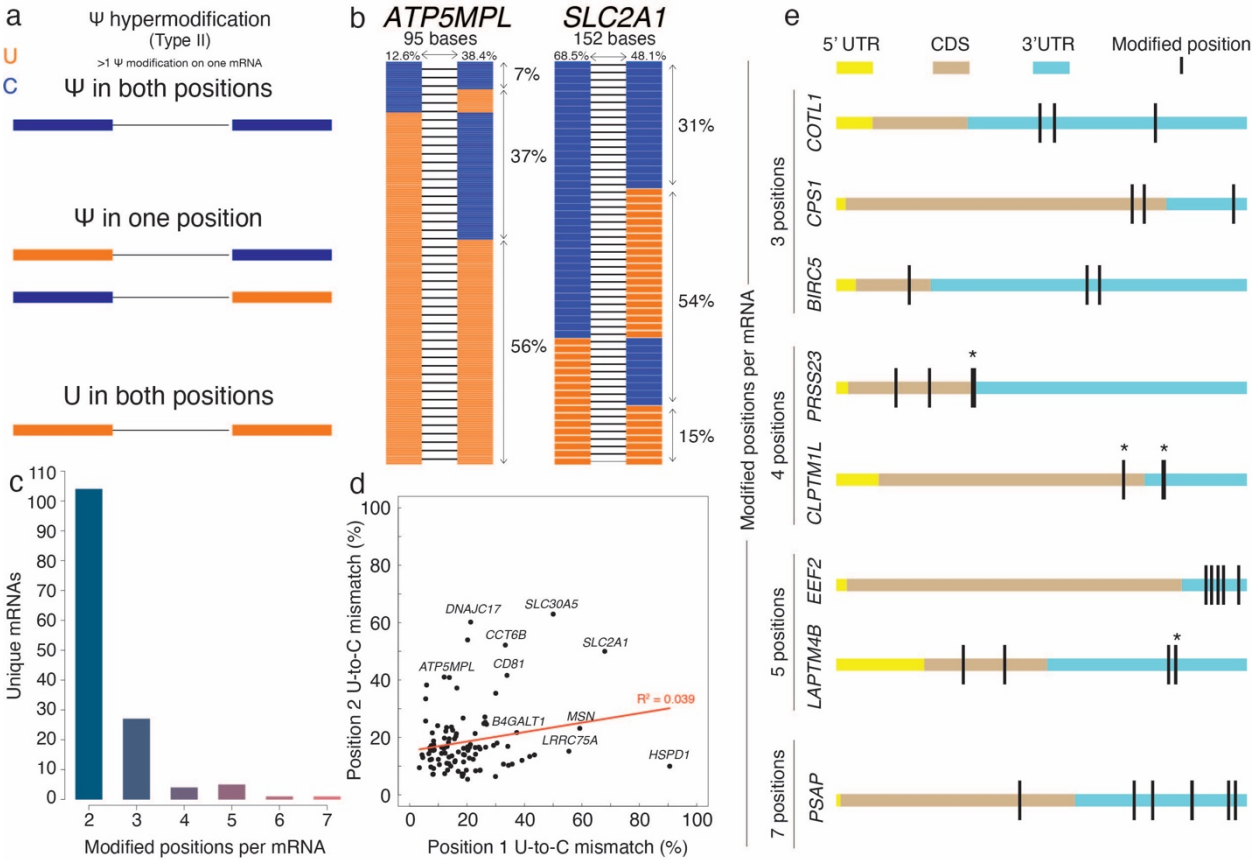
Nanopore sequencing detects the psi modification *de novo* and validates targets detected by previous methods. a. the U-to-C mismatches detected by nanopore sequencing versus the -log10(significance) of merged direct RNA. green: the detected targets identified by the significance factor of both replicates lower than 0.01, or the significance factor of at least one replicate is equal to or lower than 0.001, blue: The detected targets that have a significance factor of lower than 0.001 in both replicates (higher confidence). b. The gene ontology (GO) analysis of Molecular Function 2021 for the gene annotations that contain higher confidence detected pseudouridylation. The analysis is performed using enrichR website[34–36]. c. The k-mer frequency of the most frequently detected targets with higher confidence. d. The sequence motif across the detected psi modification for the most frequently detected k-mers generated with kplogo [38]. e. The distribution of detected psi sites in the 5' untranslated region (5' UTR), 3' untranslated region (3' UTR), and coding sequence (CDS). f. The read depth of the reads aligned to PRR13 versus the relative distance to the transcription start site (TSS) and transcription termination site (TTS). g. 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 [37]. h. Correlation of splice distance of targets located on a CDS region of their respective dominant isoform versus the full length of that particular CDS. i. Correlation of the distance between the nearest splice site and targets located on the 3'UTR region versus the full length of that particular 3'UTR.

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*. 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. f. Comparison of psi site occurance near splice sites between hypermodified sites and not hypermodified psi sites. g. Gene ontology (GO) analysis of Molecular Function 2021 for the genes in hypermodified type I and not hypermodified extracted from the enrichR website[34–36].

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. Li, X. *et al.* Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. *Nat. Chem. Biol.* **11**, 592–597 (2015).
3. 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).
4. Spitale, R. C. *et al.* Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519**, 486–490 (2015).
5. Wang, X. *et al.* N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* **505**, 117–120 (2014).
6. Hunter, S. *et al.* InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* **40**, D306–12 (2012).
7. Cohn & Elliot, W. Nucleoside-5′-Phosphates from Ribonucleic Acid. *Nature* 483–484 (1951).
8. 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).
9. 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.
10. 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).
11. Anderson, B. R. *et al.* Incorporation of pseudouridine into mRNA enhances translation by diminishing PKR activation. *Nucleic Acids Res.* **38**, 5884–5892 (2010).

12. Schwartz, S. *et al.* Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell* **159**, 148–162 (2014).
13. Carlile, T. M. *et al.* Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature* **515**, 143–146 (2014).
14. Kirwan, M. & Dokal, I. Dyskeratosis congenita, stem cells and telomeres. *Biochim. Biophys. Acta* **1792**, 371–379 (2009).
15. Agarwal, S. *et al.* Telomere elongation in induced pluripotent stem cells from dyskeratosis congenita patients. *Nature* **464**, 292–296 (2010).
16. Charette, M. & Gray, M. W. Pseudouridine in RNA: what, where, how, and why. *IUBMB Life* **49**, 341–351 (2000).
17. 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).
18. Addepalli, B. & Limbach, P. A. Mass spectrometry-based quantification of pseudouridine in RNA. *J. Am. Soc. Mass Spectrom.* **22**, 1363–1372 (2011).
19. Ho, N. W. & Gilham, P. T. Reaction of pseudouridine and inosine with N-cyclohexyl-N'-beta-(4-methylmorpholinium)ethylcarbodiimide. *Biochemistry* **10**, 3651–3657 (1971).
20. 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).
21. 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).
22. Li, X., Ma, S. & Yi, C. Pseudouridine: the fifth RNA nucleotide with renewed interests. *Curr.*

Opin. Chem. Biol. **33**, 108–116 (2016).

23. Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).
24. Liu, H. *et al.* Accurate detection of m 6 A RNA modifications in native RNA sequences. *Nat. Commun.* **10**, 1–9 (2019).
25. 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).
26. Begik, O. *et al.* Decoding ribosomal RNA modification dynamics at single molecule resolution. *Cold Spring Harbor Laboratory* 2020.07.06.189969 (2020) doi:10.1101/2020.07.06.189969.
27. 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.
28. 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.
29. Pyle, A. M. Translocation and unwinding mechanisms of RNA and DNA helicases. *Annu. Rev. Biophys.* **37**, 317–336 (2008).
30. Carlile, T. M. *et al.* mRNA structure determines modification by pseudouridine synthase 1. *Nat. Chem. Biol.* **15**, 966–974 (2019).
31. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
32. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).

33. Lovejoy, A. F., Riordan, D. P. & Brown, P. O. Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in S. cerevisiae. *PLoS One* **9**, e110799 (2014).

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

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

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

37. 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).

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