mRNA psi profiling using nanopore DRS reveals cell type-specific pseudouridylation

Caroline A. McCormick[1], Yuchen Qiu[1], Oleksandra Fanari[1], Yifang Liu[1], Dylan Bloch[1], Isabel N. Klink[1], Michele Meseonznik[1], Miten Jain[1,2], Meni Wanunu[1,2], Sara H. Rouhanifard[1]
[1]*Dept. of Bioengineering, Northeastern University, Boston, MA*
[2]*Dept. of Physics, Northeastern University, Boston, MA*
Correspondence: s.rouhanifard@northeastern.edu

Abstract
Pseudouridine (psi) is one of the most abundant human mRNA modifications generated from the isomerization of uridine via psi synthases, including *TRUB1* and *PUS7*. Nanopore direct RNA sequencing combined with our recent tool, Mod-*p* ID, enables psi mapping, transcriptome-wide, without chemical derivatization of the input RNA and/or conversion to cDNA. This method is highly efficient for detecting changes in positional psi occupancies across cell types, which can inform our understanding of the impact on gene expression. We sequenced, mapped, and compared the positional psi occupancy across six immortalized human cell lines derived from diverse tissue types. We found that lung-derived cells have the highest proportion of psi, while liver-derived cells have the lowest. Further, among a list of highly conserved sites across cell types, most are TRUB1 substrates and fall within the coding sequence. Interestingly, we identify cell type-specific sites of psi modification in ubiquitously expressed genes. We validate these sites by ruling out single-nucleotide variants, analyzing current traces, and performing enzymatic knockdowns of psi synthases. Finally, we characterize sites with multiple psi modifications on the same transcript (hypermodification type II) and found that these can be conserved or cell type-specific. Among these, we discovered examples of multiple psi modifications within the same k-mer for the first time and analyzed the effect on current distribution. Our data support the hypothesis that motif sequence and the presence of psi synthase are insufficient to drive modifications and that cell type-specific trans-acting factors play a major role in driving pseudouridylation.

Introduction

Pseudouridine (psi) is an RNA modification resulting from the isomerization of uridine by the psi synthase family of enzymes (PUS). Psi is present in various RNAs, including transfer RNA (tRNA), ribosomal RNA (rRNA), noncoding RNA (ncRNA), and messenger RNA (mRNA). On mRNAs, psi impacts translational fidelity(1, 2), structural stability(3), and pre-mRNA processing(4), but the contribution of individual psi sites on mRNAs is unknown. To discern the role of psi in the human transcriptome, we need a comprehensive map of psi across human cell types. Previous studies have identified psi positions (5–7) and the relative psi-to-uridine fractions in mRNA (5, 6) across cell lines. The only differential analyses between different cell types primarily used mouse-derived tissue samples where psi positions were common and specific to cell types(5). However, these methods rely on either *N*-cyclohexyl-*N*′-(2-morpholinoethyl)carbodiimide methyl-*p*-toluenesulfonate (CMC) to block the progress of reverse transcriptase during library preparation(8, 9) or bisulfite detection which results in a systematic deletion signature during reverse transcription(5, 7, 10). Both methods rely on cDNA conversion and amplification, leading to amplification biases(11). Further, these methods are prone to false positives when the detected psi has neighboring uracil sites(5).

Nanopore Direct RNA Sequencing (DRS) is currently the only method to detect native human RNA without conversion to cDNA by passing the RNA through a biological pore and associating

an ionic current disruption with a corresponding local sequence. Because a direct strand of RNA is passed through the pore (i.e., not a converted cDNA strand), chemical modifications to the RNA nucleobases often produce aberrations in the expected ionic current disruption(12). DRS has been used to map several mRNA modifications to date, including psi(13–15), N^6-methyladenosine (m^6A)(16–21), and inosine (I)(22) across the human transcriptome. The presence of psi on a native RNA strand frequently leads to a systematic U-to-C basecalling error which we compare against an unmodified human transcriptome(23) in a method called Mod-*p* ID(24) to assign psi-sites across the transcriptome(23). Mod-*p* ID is a semi-quantitative technique as the frequency of the U-to-C basecalling error is sequence-dependent, and therefore, the occupancy of one site cannot be compared to that of a different site because of the U-to-C error. However, relative psi occupancies for the same site can be compared across cell lines or conditions(25). Furthermore, DRS has the potential to detect multiple modifications on a single transcript, providing a way to study multiple modifications on the same transcript(26).

Here, we apply Mod-*p* ID to perform a comparative, transcriptome-wide analysis of psi in six diverse, immortalized human cell lines: A549 (lung), HeLa (cervix), HepG2 (liver), Jurkat (T cells), NTERA (testes), SH-SY5Y (neuron-like). We examine how relative psi occupancies change across these cell types and investigate the relationship between relative psi occupancy and psi synthases. We also detect cell-type-specific expression of psi modifications in ubiquitously expressed transcripts. Finally, we detect transcripts bearing multiple modifications and show that double modifications can occur within the same k-mer. Our differential approach provides insight into specialization in RNA modifications across cell types transcriptome-wide.

Results

Comparative analysis of pseudouridine mRNA modifications from six human cell lines using DRS

To quantify transcriptome-wide differences in psi expression across human cell types, we isolated and sequenced using DRS poly-A selected RNA from six immortalized human cell lines: i) A549, alveolar basal epithelial carcinoma; ii) HeLa, cervical carcinoma; iii) HepG2, hepatocellular carcinoma; iv) Jurkat, T-cell leukemia; v) NTERA-2, human embryonic carcinoma derived from testicular cancer; and vi) SH-SY5Y, neuroblastoma (**Fig. 1a**). The primary alignments for each sample yielded the following number of reads: 2,362,999 (A549); 4,257,869 (HeLa); 2,407,221 (HepG2); 10,931,896 (Jurkat); 10,107,973 (NTERA); 6,064,224 (SH-SY5Y). Following basecalling and genomic alignment, we mapped putative psi positions across the transcriptome for each cell line using our previously described pipeline, which we will refer to as Mod-*p* ID(24), by comparing our direct libraries to an *in vitro* transcribed (IVT) library(13, 23). An IVT control derived from the same cell type (i.e., paired-IVT or p-IVT) is preferential for this analysis; however, we have previously demonstrated that a pan-IVT comprised of merged unmodified transcriptomes of different cell lines can enrich data sets where paired IVTs do not have sufficient coverage(23).

We used the following stringent filtering criteria to reduce the possibility of false positives in psi identification: 1) IVT control has ≥ 10 reads, 2) IVT error is ≤ 10%, 3) If the paired IVT dataset has <10 reads, we used the pan-IVT

dataset. For all sites that passed these criteria, we applied Mod-*p* ID to calculate *p*-values in the six cell lines (see **Methods**; **Figure 1b**).

Our first requirement for the direct library was that the putative site must have a significant difference in U-to-C basecalling error between the direct and IVT sample ($p < 0.001$). Next, we required the number of reads from the direct DRS library to be ≥ 10 for a given site. Finally, we required the U-to-C mismatch error derived from the direct library to be ≥ 10%. We identified putative psi sites for each cell line (**Fig 1c; Supplementary Table 1**). To find the relative ratio of putative sites for a cell line, we calculated the number of putative psi sites detected by Mod-*p* ID and normalized by dividing by the number of aligned genomic uridines for each cell line, obtaining 10.40×10^{-4} for A549 cells, 8.97×10^{-4} for HeLa cells, 4.46×10^{-4} for HepG2 cells, 9.80×10^{-4} for Jurkat cells, 10.85×10^{-4} for NTERA cells, and 5.37×10^{-4} for SH-SY5Y cells. Globally, A549 and NTERA cells had the highest relative ratio of putative sites across cell lines, while HepG2 and SH-SY5Y cells had the lowest (**Fig 1c**).

Expression profiling of PUS enzymes across cell types

We calculated the transcripts per million (TPM) for PUS enzyme mRNAs and compared the levels across cell lines (**Figure 1d**; **Supplementary Table 2**). We categorized putative psi positions within a known psi synthase motif to test the hypothesis that higher PUS enzyme levels can result in more psi sites in a given cell type. This analysis assumes that PUS mRNA levels across cell types are commensurate with their corresponding PUS enzyme levels. For the psi synthase *TRUB1,* we searched for the motif GUUCN(27); for *PUS7,* we searched for the motif UNUAR(28). We computed the proportion of Ψ positions with the motif k-mer divided by the total number of positions with the motif k-mer,

$$proportion\ of\ TRUB1\ psi\ positions\ = \frac{sum(genomic\ psi\ positions\ with\ GUUCN\ k-mer)}{total\ genomic\ positions\ with\ GUUCN\ k-mer}$$

$$proportion\ of\ PUS7\ psi\ positions\ = \frac{sum(genomic\ psi\ positions\ with\ UNUAR\ k-mer)}{total\ genomic\ positions\ with\ UNUAR\ k-mer}$$

and compared these proportions to the normalized RNA expression levels for *TRUB1* and *PUS7*. Interestingly, we found that NTERA cells have the highest proportion of psi sites within a PUS7 motif (**Figure 1e**), while SH-SY5Y cells have the lowest proportion of psi (**Figure 1e**). These findings are consistent with the high global levels of psi in NTERA cells and low levels in SH-SY5Y cells (**Figure 1c**). We also found that Jurkat cells have the highest proportion of psi within a TRUB1 motif, while NTERA cells have the lowest. We compared the proportion of psi within each of these motifs to the respective enzyme mRNA expression levels. Generally, we found no correlation between PUS expression levels in a given cell type and the total number of psi sites for that cell type (**Figure 1e**).

Comparison of conserved psi sites across six human cell lines

To compare the occupancy of individual psi positions across cell types, we first selected sites for which a psi modification is conserved (i.e., has been identified in every cell line) by Mod-*p* ID ($p < 0.001$). We then filtered these positions only to include those with at least 30 reads from the

direct library and a minimum of 10 IVT reads for all six cell lines. This filtration step, which ensured that the mRNAs were highly expressed across all cell types, produced a list of 70 psi positions (**Supplementary Table 3**).

Of these 70 sites, we applied additional filtering criteria, including a minimum of 30% U-to-C error in each cell line, to compile a final list of "housekeeping" psi sites (**Supplementary Table 3**), used here to denote psi sites that are abundant across all cell types. Cross-validation of our 20 detected "housekeeping" sites with other methods (i.e., siRNA knockdown of a *TRUB1* or *PUS7(25)*, CMC-mediated Illumina sequencing methods(6, 8), and RNA bisulfite labeling methods(5–7, 10, 29) produced a list of 17 orthogonally confirmed, conserved sites of psi modification (**Figure 2a**). We mapped 11 sites to the CDS and 6 to the 3' UTR (**Figure 2b**). Interestingly, the *TRUB1* motif (GUUCN) was the predominant k-mer appearing in 14 of our 17 highly conserved sites, each additionally validated by a *TRUB1* siRNA knockdown experiment (25). These sites are found within genes that perform essential functions across tissue types. However, a few common pathways were identified. *NIP7*, *DKC1,* and *PARP4* are involved in nucleic acid processing and maintenance, and *SLC30A5* and *SLC2A1* play a role in transport functions. Interestingly, the only PUS7 substrate on this "conserved" list was *RHBDD2*, which encodes for an intramembrane serine protease necessary for processing various proteins and maintaining cellular homeostasis.

Next, we calculated a similarity score for each psi position to compare across all six cell lines (**Figure 2a**). The conserved psi site located on *NIP7* (chr16:69342144) has the most similar positional occupancy ranging from 98% (NTERA) to 100% (A549, HepG2, and SH-SY5Y). *NIP7* (Nucleolar pre-rRNA processing protein) is involved in RNA binding activity for ribosomal biogenesis. We also observed in a previous study that this position on *NIP7* is relatively static in SH-SY5Y cells across multiple treatments(25). Inversely, the relative positional occupancy of the psi on *SLC2A1* (chr1:42926727) ranges from 32% U-to-C error (Jurkat) to 79% (HepG2). *SLC2A1* (solute carrier family 2 member 1) is a critical glucose transporter essential for tissue-specific energy delivery. While *SLC2A1* is present in both liver and T cells, its primary role in the liver is related to supporting glucose metabolism, whereas in T cells, *SLC2A1* is crucial for immune function and response. This psi site has also been found to have stable occupancy across different treatments for SH-SY5Y cells(25).

Relationship between transcript levels and positional psi occupancy

The conserved transcript *FKBP4* (chr12:2803909) has highly variable psi occupancy across the cell lines ranging from 38% (Jurkat cells) to 69% (A549 cells), as well as high variability in transcript expression levels ranging from 81±8 (SH-SY5Y cells) to 735±16 (NTERA cells) TPMs (**Figure 2a,c**). *FKBP4* (prolyl isomerase 4) stabilizes and manages steroid hormone receptors in both lung and T cells. FKBP4 affects inflammation and immune responses in the lung, while FKBP4 regulates cell differentiation and effector functions in T cells.

We observed several examples of an inverse relationship between mRNA levels and psi positional occupancy. In NTERA cells, *FKBP4* expression is the highest out of all the cell lines (349 TPM) but one of the lowest psi positional occupancy at 42%. Inversely, A549 cells have significantly lower *FKBP4* expression levels with 111±6.TPMs but show the highest psi positional occupancy at 69%. *SLC2A1* is also differentially expressed with variable psi

occupancy at chr1:42926727 across the cell lines. However, this case directly correlates transcript abundance and psi positional occupancy. HepG2 cells have the highest expression of *SLC2A1* at 210±9 TPMs and the highest U-to-C mismatch at 79We %, and Jurkat cells have the lowest expression and psi occupancy at 37±4. and 32%, respectively (**Figure 2a,c**).

Secondary structure modeling of the region surrounding psi position for TRUB1 targets
As mentioned earlier, most highly conserved targets were TRUB1 substrates (**Figure 2a**). Previous studies have analyzed the structural features standard to TRUB1 targets in a single cell line(27) and found that the stem and loop structure on mRNA within the TRUB1 motif is sufficient for mRNA pseudouridylation. We were interested in seeing if these sites were conserved across cell types because the sequences are excellent substrates for TRUB1 and if we could uncover more structural features by focusing on these sites. Using Mfold(30), we took a 21-mer sequence flanking the conserved and validated psi positions found within a TRUB1 motif (14 targets) and predicted the RNA secondary structure (**Figure 2d,e; Supplementary Table 4**). Additionally, we randomly selected 14 control sites with the same consensus motif but bearing no psi modification (**Figure 2f,g**).

In comparing the structural characteristics of these groups, we observed that the control sites exhibit a broader range of loop lengths, varying from 3 to 12 nucleotides, and a stem length ranging from 3 to 8 base pairs (**Figure 2f,g**). Conversely, the query sites demonstrate a narrower loop length range of 4 to 10 nucleotides and a shorter stem length, varying from 2 to 6 base pairs. Notably, internal loops are present in several control targets, including *MRPL30*, *AHCYL1*, *SRRT*, and *TPI1*, a feature absent in modified targets. Furthermore, the distance from the k-mer to the stem's 5' or 3' end is more extended in unmodified structures, ranging from 1 to 8 nucleotides, compared to 1 to 5 nucleotides in modified structures (**Figure 2f,g**).

Site-specific psi presence or absence can be cell-type dependent.
The unique presence or absence of a modification site in a cell type implies some form of epitranscriptome-level specialization for that cell type. We, therefore, were interested to see whether psi sites were present at given positions only in specific cell lines, even if the RNA expression levels were conserved. We constructed a table of Mod-*p* ID detected and orthogonally confirmed psi sites where the modification was detected in 3 of the 6 cell lines (**Figure 3a; Supplementary Table 3**). For example, *DAZAP1* (chr19:1434893) was only present in A549, HeLa, and HepG2 while *ITFG1 (*chr16:47311407*)* was present only in A549, HepG2 and SH-S75Y cells. Interestingly, for this group of modified sites, we found no positions within this category that were hypermodification type 1(13); all identified sites were <40% U-to-C error. We compared all aligned reads' raw ionic current traces at the psi position, revealing that the known psi presence affects the current (**Figure 3b**).

The unique absence of a psi site for a single cell line may implicate cell type-specific functionality. We relaxed the criteria to include positions with at least 30 direct reads in every cell line and 10% psi positional occupancy for the detected site in 5 out of 6 cell lines (**Supplementary Table 3**). We detected 38 putative psi sites with a unique absence in a single cell line with Mod-*p* ID, 12 of which are orthogonally confirmed psi positions (**Figure 3c**). All 12 sites are present in A549 and HeLa cells, indicating no unique absence of psi at any detected position in these two cell lines. Inversely, SH-SY5Y was only found to be present in 6 psi

positions, making it the cell line with the highest occurrence of uniquely absent sites. Four of the six psi positions present in SH-SY5Y are found on transcripts (*CDC42* (chr1:22092496), *WDR45* (chrx:49075653), *YWHAQ* (chr2:9584978), and *SNAP29* (chr22:20870467)) related to neuron function directly in the cellular activities or through processes crucial for neuronal health and activity. Each of these 12 positions was orthogonally validated to be psi (**Figure 3c**). The position with the largest difference between the two cell lines was within *H2AZ2* (chr7:448355564), which is hypermodified (type I) in HepG2 cells and not modified in SH-SY5Y cells while being highly expressed in both. This gene encodes a variant of the histone H2A complex, plays a role in synaptic plasticity in neurons, and plays a role in liver development. We do not know which PUS enzyme is responsible for this modification. However, this position was orthogonally confirmed by two other methods as psi.

To further confirm the cell-type-specific absence of psi, we overlaid the per-read ionic current intensity distribution ionic current traces for all aligned reads to the psi position (**Figure 3d**). As expected, at chr2:9584978 (*YWHAQ*), we observed a single, smooth distribution in HepG2 cells and the pan-IVT, whereas, in the other five cell lines, there is an additional peak to the left of the expected distribution peak, indicating a difference in the physical properties of the sequences (i.e., thus leading to a change in the ionic current disruption) corresponding to the observed absence of psi (**Figure 3d**).

Likewise, the unique presence of psi positions in a cell line may further provide information on the cell-type-specific functionality of psi. Of the 63 orthogonally confirmed psi sites using Mod-*p* ID present in only one cell line with at least 10% relative positional occupancy, seven sites have a *TRUB1* motif, and 12 have a *PUS7* motif. Only 4 of the orthogonally confirmed psi sites have above a 20% U-to-C error signature: *DRG2* (chr17:18107480), *MRPL46* (chr14: 88467298), *SLC25A1* (chr22:19177964), and *IPO7* (chr11:9445686). Additionally, we identified 14 putative psi sites using Mod-*p* ID, which have >30% U-to-C errors (**Figure 3e**). For all sites of unique modification presence, we ruled out the presence of a single-nucleotide variant in these cell lines by Sanger sequencing on the corresponding gDNA (**Figure 3f; Supplementary Figure S1**). We wanted to confirm that these sites deviated from the expected ionic current, which indicates a probable modification. We observed that the pan-IVT trace and the other five cell lines have similar ionic current trace distributions and a visible deviation of the cell line distribution with the modification (**Figure 3g**).

Type II hypermodification of transcripts can be variable or shared across cell lines.
Type II psi hypermodification has been previously defined as transcripts with more than one psi modification(24). To assess the type II hypermodification status across cell types, we filtered to require ten direct reads, and the paired-IVT had ten reads at the query position (**Supplementary Table 5**). We detected up to 5 modified positions per transcript for A549 cells, HeLa cells, HepG2 cells, Jurkat cells, and NTERA cells. We detected the presence of many transcripts with two modifications in all cell lines. A549 cells, NTERA cells, and Jurkat cells consistently had the highest detected type II hypermodified transcripts. In contrast, SH-SY5Y cells had the lowest number of hypermodified transcripts as we only detected up to 3 positions on one transcript (**Figure 4a**). This was surprising because SH-SY5Y cells have 6.1M aligned reads, while A549 cells only have 2.4M reads.

PSAP encodes the prosaposin gene and is one example of a transcript with multiple positions orthogonally confirmed (chr10:71816374(7, 10), 71816664(5–7, 10, 29), 71817048(10), and 71819612(10)) and detected at varying positional occupancy with our Mod-*p* ID method (**Figure 4b**). The encoded protein is essential for the proper functioning of the lysosome. Of the positions identified, chr10:71819612 is in the CDS, while the other three are in the 3' UTR. Neither of these confirmed sites has a known PUS motif. All 4 psi locations are present in HeLa cells, 3 in A549 cells, 2 in NTERA, 1 in HepG2, and 0 in Jurkat and SH-SY5Y. The positional occupancies for detected psi sites are relatively low, with a maximum of 33% in HepG2 at chr10:71819612. Mod-*p* ID also identified a 5th putative position at chr10:71816662 in HeLa and NTERA cells. This additional putative psi site falls within the same k-mer of chr10:71816664, a putative double modification (**Figure 4b**).

LAPTM4B (chr8:97776038, 97776040) is another example of a double modification, where chr8:97776038 has a TRUB1 k-mer, GUUCU (**Figure 4c**), which has been orthogonally confirmed with both PRAISE(10) and CeU-seq(6). CMC and bisulfite-based methods for psi detection cannot detect two neighboring modifications. Thus, there are no known examples of double modifications. Double modifications can also potentially disrupt current traces by direct RNA sequencing. Still, we explored the possibility of a double modification to see how this might affect the current.

We performed a strand analysis of all six cell lines' combined reads (pan-direct) to see if the double modification occurred simultaneously in both positions on a given read. A read must span both positions and only contain a U or C to be included in the analysis. We separated the 1,823 reads into four groups: i) no modification (1,619 reads), ii) U-to-C error in the first position and an unmodified second position (76 reads), iii) an unmodified first position and U-to-C error in the second position (124 reads), and iv) double modification (4 reads). We wanted to visualize the ionic current distribution at these two positions, specifically when the nucleotide of interest is in the middle position of the nanopore, for each grouping. We observe changes in the distribution pattern corresponding to each grouping. Overall, this suggests that singly-modified and doubly-modified populations may exist within the same k-mer, but the mechanism remains unknown (**Figure 4d**).

Discussion

Our recently developed technique, Mod-*p* ID, enables the mapping of psi across the human transcriptome using nanopore DRS and an IVT control to define baseline signals. Here, we compared the relative psi occupancy at specific positions across different cell types, including A549, HeLa, HepG2, Jurkat, NTERA, and SH-SY5Y cells, and found that A549 cells had the highest percentage of putative psi sites relative to the library size (**Figure 1**). This observation is unexpected because our analysis of PUS enzyme levels consistently shows lower levels of PUS transcripts in this cell line. This suggests that factors beyond enzyme levels contribute to psi modification and that human lung cells may be more prone to pseudouridylation.

We defined "conserved psi sites" as sites that have high expression of mRNA (defined by >30 reads) and >30% U-to-C error in a single position within all cell lines (**Figure 2**). By limiting the analysis to sites with more than 30 reads, we restrict the analysis to "housekeeping genes" to focus on differences in psi rather than gene expression. Interestingly, most conserved psi

positions that fit our filtering criteria were substrates for the TRUB1 enzyme, indicating that TRUB1 targets are the most conserved across human cell lines. Among these sites, we observed that conserved psi sites can have both consistent and variable occupancy, and we assigned a similarity score to quantify this. The conserved psi sites with the most similar relative occupancy (i.e., similarity score between 0 and 10) are also close to 100% psi occupancy. For example, the occupancies for all cell lines at the psi position on *NIP7* (chr16:69342144) are above 97.6%, while the occupancies of the psi position on *SLC2A1* (chr1:42926727) vary from 40-90% depending on cell type. Both *NIP7* and *DKC1* are involved in ribosome biogenesis. *NIP7* is implicated in pre-rRNA processing(31), while *DKC1* assembles and modifies ribosomal RNA (rRNA) within the nucleolus. While *NIP7* does not possess pseudouridine synthase activity, it interacts with proteins such as *DKC1*, catalyzing the pseudouridylation of specific rRNA residues, which is crucial for ribosome function and stability. Among the sites with lower similarity scores, *PLD3* (chr19:40378040) was found at the lowest occupancy (30%) in SH-SY5Y cells associated with Alzheimer's disease and may play a role in neuronal function and lipid metabolism. The predictive secondary structure modeling of canonical nucleotide sequences for conserved *TRUB1* psi sites showed that all genetically encoded uridines are within hairpin loops, consistent with previous analysis(27)(**Figure 2**). Nucleotides located in a hairpin loop are inherently more accessible to enzymes such as *TRUB1,* which is the predominant PUS enzyme modifying mRNA(27). We observed deviations in the length of the loop and stem for sites that did not have a detected psi to support this hypothesis.

We also observed positions susceptible to pseudouridylation in specific cell lines while remaining unmodified in others (**Figure 3**). We consistently observe lower positional occupancy, typically ranging from 10% to 50%, at psi sites detected in two, three, four, or five distinct cell lines. This makes it possible that the modification fell below the detection limit for some cell lines. We carefully checked for cell line-specific SNVs that could be misinterpreted as a modification for these sites and excluded these from the analysis. To understand the reason for the U-to-C basecalling error, we analyzed the raw signal data features to detect anomalies in signal patterns. The presence of a modification causes a visible shift of the ionic current intensity distribution from that of the corresponding canonical nucleotide at a given position(15). We observe this shift even at a lower relative positional occupancy level for the absence of a modification site in a single cell line (**Figure 3f, i**). Conversely, the unique presence of a putative psi position in a cell line can suggest its functional importance and regulatory role in that cell type. *DRG2* (chr17:18107480) is a cell type-specific psi modified by PUS7 detected in HeLa cells. DRG2 helps to regulate cell growth and differentiation. Disruption in *DRG2* expression in HeLa cells has been shown to alter microtubule dynamics, resulting in cell proliferation changes(32).

Hypermodification type II on a single transcript varies between cell lines (**Figure 4**). While one cell line may have psi sites at multiple positions on the transcript, another may exhibit hypermodification at different positions on the same transcript. A special case of type II hypermodification is a double modification where two modifications occur within a k-mer so that both modifications will exist in the nanopore at the same time. For this type of biological pore, approximately 11 nucleotides will fit within the region interacting with the pore, and modifications that are close to each other can potentially impact the detected signal of the other. It is generally accepted that mRNA modifications are more scarce than modifications in rRNAs

and tRNAs. However, Illumina-based technologies cannot discern neighboring modifications, so there was no previous evidence that satellite modifications occur on mRNAs. We found a few rare examples of these modifications. For example, we observed double modifications in *LAPTM4B* in A549, HeLa, and HepG2 cell lines at chr8:97776038 and chr8:97776040 (**Figure 4**). The visible differences in the KDE of merged ionic traces may indicate that the detection of double modification is not a false positive. chr8:97776038 has been orthogonally confirmed(6, 10). TRUB1 siRNA knockdown in HeLa cells showed the U-to-C error disappear in both chr8:97776038 and chr8:97776040 compared to both the native HeLa mRNA and scrambled control. We ruled out the presence of an SNV in all cell lines for these sites by Sanger sequencing on the corresponding gDNA (**Supplementary Figure S1**).

Overall, these studies advance our knowledge of site-specific psi distribution across the human body and push forward the types of RNA modification analysis that can be accomplished using DRS (specifically combinatorial analysis). As new chemistries and flow cells are developed, the biological insights we derive from comparative analysis across cell types will not change; they should only improve sensitivity and resolution. Importantly, these studies will act as a springboard, prompting focused biological analyses to discern the complex biological roles of psi in the human body.

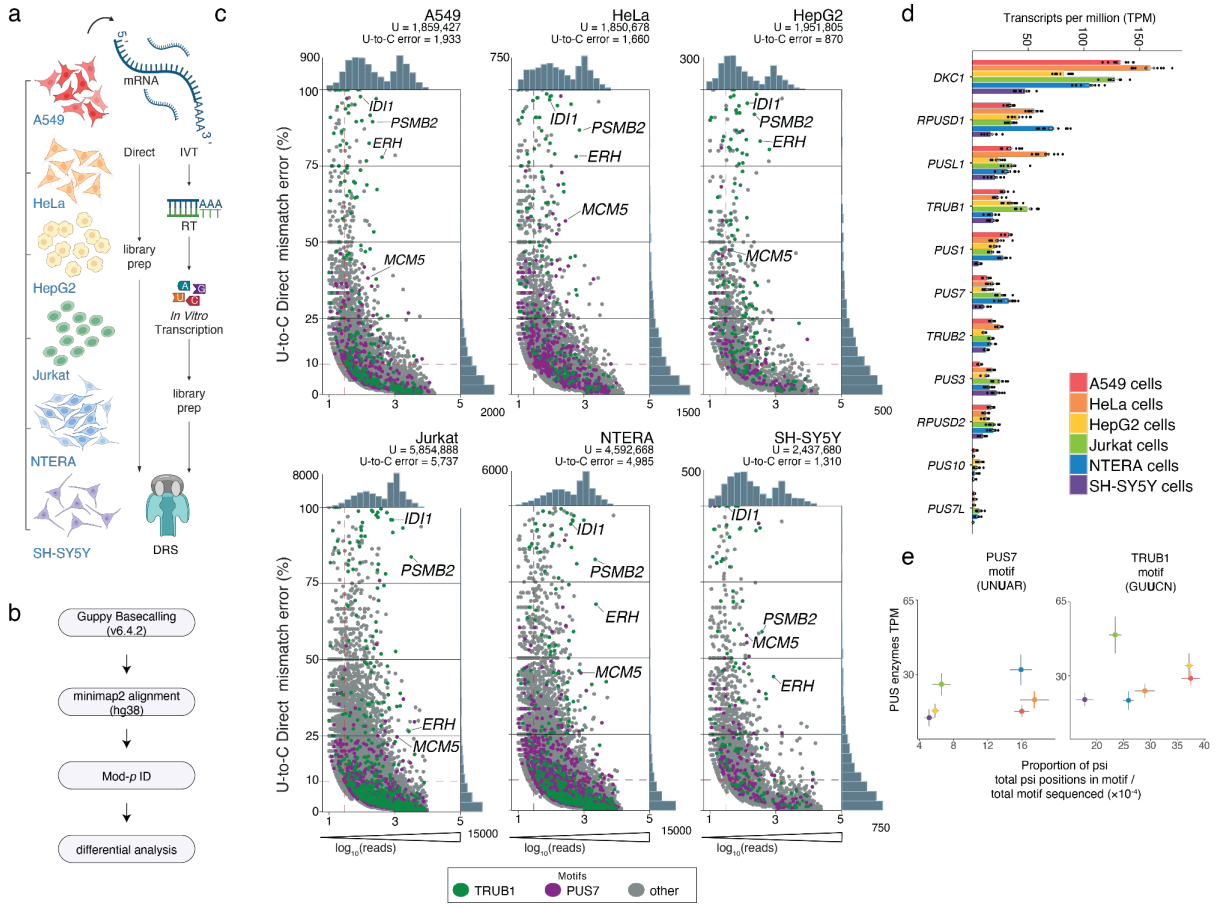

Figure 1. Individual cell line putative psi identification by Mod-*p* ID and PUS expression analysis. **a)** Experimental summary of A549, HepG2, HeLa, Jurkat, NTERA, and SH-SY5Y direct and IVT library prep and DRS. **b)** overview of computational pipeline following DRS. **c)** Number of total direct reads versus the relative occupancy of putative (% mm) psi positions detected by Mod-*p* ID ($p < 0.001$) for each cell line. Grey points do not have an identified motif; colors indicate k-mer motifs for *PUS7* (purple) and *TRUB1* (green). Red dashed lines indicate baseline requirements for a position included in downstream analysis. **d)** TPMs of PUS enzymes. Individual points are from the computational resampling of the dataset. Significance from (***$p<0.001$) one-way ANOVA. **e)** Fraction of the total number of reads with a U-to-C mismatch in a *TRUB1* or *PUS7* k-mer divided by the total number of *TRUB1* or *PUS7* k-mers observed versus the TPMs of the PUS enzyme and standard error of the mean from computational resampling.

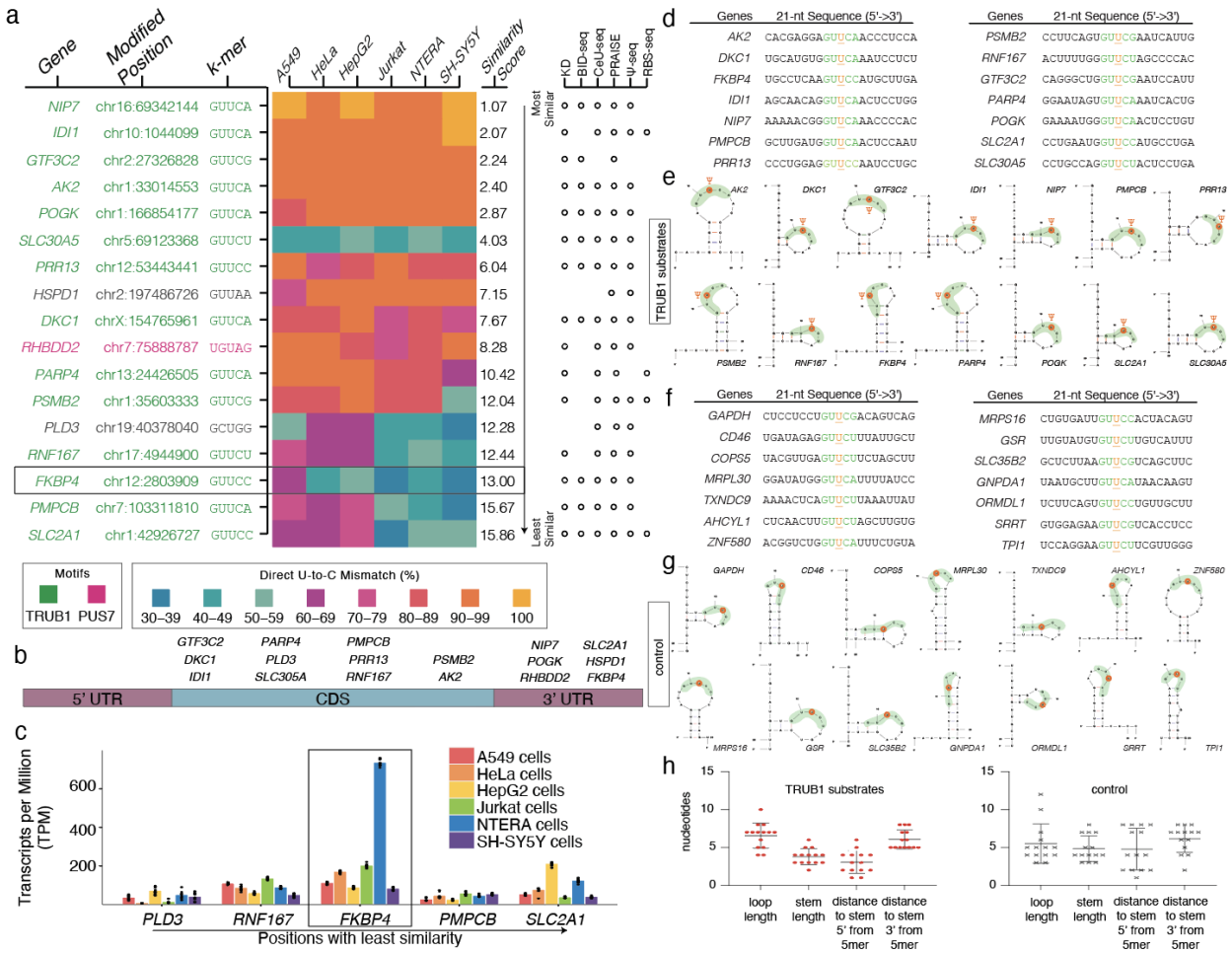

Figure 2. Conserved psi sites across 6 human cell lines. **a**) Heatmap of relative psi occupancy for sites identified by Mod-p ID[17] with ≥30 reads and ≥30% U-to-C mismatch for each cell line. Colors indicate that the psi site has a TRUB1 (green) or PUS7 (pink) k-mer motif. Similarity scores indicate the standard deviation of the mm% between the six cell lines. (right) Orthogonal methods confirming the psi site. **b**) Transcript location of psi sites. **c**) TPMs of transcripts bearing conserved psi positions with the least similarity for each cell line. **d**) 21-nt sequence of the conserved TRUB1 psi sites. **e**) Predictive secondary structure modeling of canonical 21-nt sequences of the conserved TRUB1 psi sites. **f**) 21-nt sequence of randomly selected unmodified uridine positions **g**) Predictive secondary structure modeling of canonical 21-nt sequences of the unmodified uridines **h**) Comparison of sequence properties between 14 TRUB1 substrates and 14 control sequences. The bars show the median value with standard error.

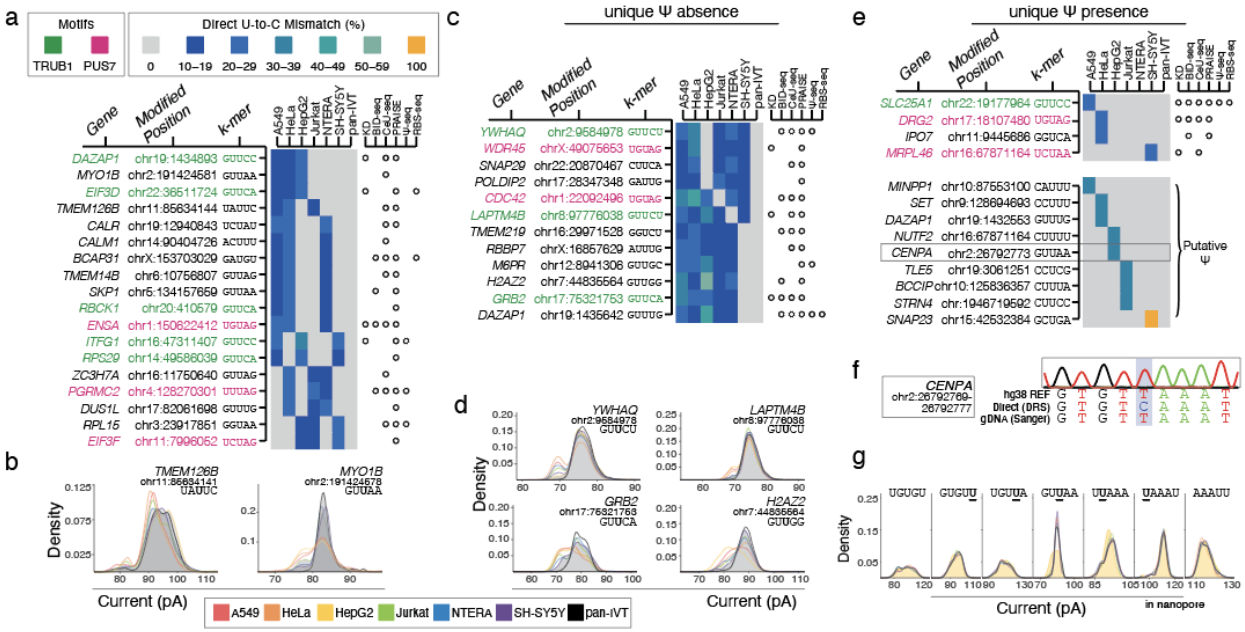

Figure 3. Cell type-specific expression of psi. **a**) Heatmap of relative psi occupancy for sites detected in 3 cell lines identified by Mod-*p* ID and confirmed by orthogonal methods. Colors indicate that psi site has a TRUB1 (green) or PUS7 (pink) k-mer motif. **b**) KDE ionic current traces of cell-type split presence/absence of psi site. Colors indicate traces for each cell line. The cell line with no psi detected is a grey-shaded distribution. **c**) Heatmap of the unique absence of a cell line identified by Mod-p ID and confirmed by orthogonal methods. Colors indicate that psi site has a TRUB1 (green) or PUS7 (pink) k-mer motif. **d**) KDE ionic current traces of cell-type specific absence of psi site. Colors indicate traces for each cell line. The cell line with no psi detected is a grey-shaded distribution. **e**) Heatmap of cell type-specific unique presence of putative psi sites' relative occupancy. **f**) Sanger sequencing of cell type-specific putative psi positions *CENPA* **g**) KDE ionic current traces of *CENPA* for +/-3 positions. Colors indicate traces for each cell line. The cell line detected to have the putative psi is shaded with its corresponding color. pan-IVT trace is in black.

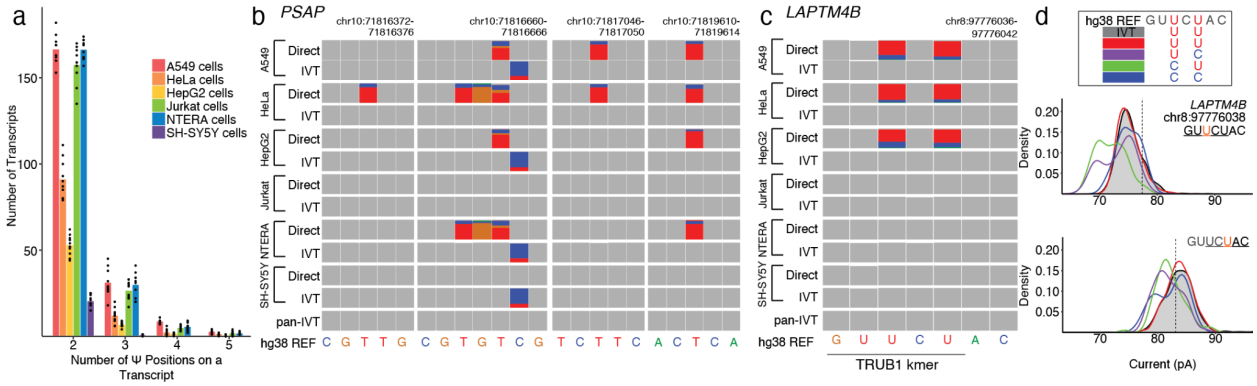

Figure 4 Type II hypermodification analysis. **a)** Number of hypermodified type II transcripts with 2, 3, 4, and 5 putative psi sites. **b)** Variable type II hypermodification across cell lines for the prosaposin gene, *PSAP*. Location on the transcript of three putative psi sites and IGV snapshots of each position's k-mer for the six-cell line's direct and IVT library and pan-IVT. **c)** Special case of type II hypermodification: double modification where multiple putative psi positions are detected within a k-mer on *LAPTM4B*. IGV snapshot of each position's k-mer for the six-cell line's direct and IVT library and pan-IVT. The first detected modification is found in *TRUB1* motif k-mer. **d)** Merged cell line KDE ionic current trances of the two double mod putative psi sites for all modification combinations for the two sites in different colors. The ionic trace for the pan-IVT is shaded in grey.

Code Availability
All code used in this work is publicly available at
https://github.com/RouhanifardLab/PanHumanPsiProfiling

Data Availability
The data that support this study are available from the corresponding author upon reasonable request. Sequences were aligned to genome version hg38.p10. Unless otherwise stated (*See Accessing Publicly Available Data Sets*), all FASTQ files and Fast5 raw data generated in this work are publicly available in NIH NCBI SRA under the BioProject accession PRJNA1108269. Due to the size of the supporting data bam files and Mod-*p* ID outputs are hosted on Dropbox at https://www.dropbox.com/scl/fo/p939edmv71o5cwl9by5x0/AFbhroxm3LIyaGY_kzgODHM?rlkey=nd6lgzg3usahxwiq8d3k2vbj6&st=uss6f43a&dl=0

Online Methods

Accessing Publicly Available Data Sets. Nanopore DRS raw fast5s for all IVTs were sourced from NIH NCBI SRA BioProject accession PRJNA947135(23). Nanopore DRS raw fast5s for native HeLa replicates were sourced from NIH NCBI SRA BioProject accession PRJNA777450(13). Nanopore DRS raw fast5s for native SH-SY5Y replicates were sourced from NIH NCBI SRA BioProject accession PRJNA1092333(25).

Cell culture. HeLa, HepG2, A549, and NTERA-2 cells were cultured in Dulbecco's modified Eagle's medium (Gibco, 10566024); SH-SY5Y cells were cultured in 1:1 EMEM: F12. All media was supplemented with 10% Fetal Bovine Serum (FB12999102, FisherScientific) and 1% Penicillin-Streptomycin (Lonza,17602E). Cells were cultured at 37°C with 5% CO_2 in 10 cm tissue culture dishes until confluent.

Total RNA extraction and Poly(A) selection. Total RNA extraction from cells and Poly(A) selection was performed using a previously established protocol(24). Six confluent 10 cm cell culture dishes were washed with ice-cold PBS, lysed with TRIzol (Invitrogen,15596026) at room temperature, and transferred to an RNAse-free microcentrifuge tube. Chloroform was added to separate the total RNA in the aqueous supernatant from the organic phase following centrifugation. The aqueous supernatant was extracted and transferred to a new RNAse-free microcentrifuge tube. An equal volume of 70% absolute ethanol was added. PureLink RNA Mini Kit (Invitrogen, 12183025) was used to purify the extracted total RNA following the Manufacturer's protocol. Total RNA concentration was measured using the Qubit™ RNA High Sensitivity (HS) assay (Thermo Fisher, Q32852). Poly(A) selection was performed using NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB, E7490L) according to the Manufacturer's protocol. The isolated Poly(A) selected RNA was eluted from the beads using Tris buffer. The poly(A) selected RNA concentration was measured using the same Qubit™ assay listed above.

***In vitro* transcription and polyadenylation.** The protocol for IVT, capping, and polyadenylation is described previously(24).

Sequencing, Basecalling, and Alignment Procedure. Cell line replicates were prepared for Nanopore Direct RNA sequencing (DRS) following the ONT

SQK-RNA002 kit protocol, including the reverse transcription step. RNA sequencing on the MinION and PrometheION platforms was performed using ONT R9.4.1 (FLO-MIN106) and (PRO-002). DRS runs were basecalled with Guppy v6.4.2 using the high accuracy model and the default basecalling quality score filter of Q ≥7(35). Basecalled read replicates for a cell line were merged and then aligned with minimap2 to the GRCh38.p10 reference genome for downstream analysis.

Genomic DNA Extraction and Sanger Sequencing. To analyze putative SNVs, we performed Sanger sequencing on genomic DNA (gDNA). gDNA extraction was performed using a Monarch Genomic DNA Purification Kit (NEB, T3010S) following the Manufacturer's protocol. PCR primers were designed to amplify ~300 nt regions surrounding the detected Ψ positions using Primer-BLAST with default settings (**Supplementary Table 6**). Using the Manufacturer's protocol, the PCR reaction was set up with Q5 polymerase (NEB, M0491L). Thermocycling conditions were as follows: initial denaturation at 98 °C for 30 seconds, 25 cycles of 98 °C for 10 seconds, then 63 °C for 20 seconds and 72 °C for 15 seconds, final extension at 72 °C for 2 minutes, and holding at 10 °C. PCR products were purified using a Monarch PCR & DNA Cleanup Kit (NEB, T1030S) following the Manufacturer's protocol. The concentration of eluted DNA was determined using a Nanodrop. The purified PCR products were imaged on a 2% agarose TBE gel to confirm specific amplification. Samples were sent to Quintara Biosciences for SimpliSeq™ Sanger sequencing. Results are shown in **Supplementary Figure S1**.

Bootstrapping. Because we compared MinION and PromethION runs, which are substantially different sizes, we performed computational resampling to generate ten computational replicates for the native mRNA with 1.2M direct reads, each randomly sampled from the base called FASTQ files for each cell line. To account for the reduction in the total number of reads, we relaxed the threshold to 10 direct reads at a position.

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

Acknowledgments. S.H.R. and M.W. acknowledge support from NIH R01HG011087 and NIH R01HG012856 as well as support through an Opportunity Fund by the Technology Development Coordinating Center at Jackson Laboratories (NHGRI federal award no. U24HG011735).

Author Contributions

C.A.M. and S.H.R. conceived of the research. C.A.M. designed the experiments. C.A.M., Y.Q., O.F., D.B., and I.N.K. performed experiments and sequencing runs. C.A.M., Y.Q., Y.L. analyzed the data with guidance from S.H.R., M.J. and M.W. C.A.M. wrote the paper with guidance from S.H.R.

REFERENCES

1. Guzzi,N., Muthukumar,S., Cieśla,M., Todisco,G., Ngoc,P.C.T., Madej,M., Munita,R., Fazio,S., Ekström,S., Mortera-Blanco,T., *et al.* (2022) Pseudouridine-modified tRNA fragments repress aberrant protein synthesis and predict leukaemic progression in myelodysplastic syndrome. *Nat. Cell Biol.*, **24**, 299–306.

2. Jack,K., Bellodi,C., Landry,D.M., Niederer,R.O., Meskauskas,A., Musalgaonkar,S., Kopmar,N., Krasnykh,O., Dean,A.M., Thompson,S.R., *et al.* (2011) rRNA pseudouridylation defects affect ribosomal ligand binding and translational fidelity from yeast to human cells. *Mol. Cell*, **44**, 660–666.

3. Kierzek,E., Malgowska,M., Lisowiec,J., Turner,D.H., Gdaniec,Z. and Kierzek,R. (2014) The contribution of pseudouridine to stabilities and structure of RNAs. *Nucleic Acids Res.*, **42**, 3492–3501.

4. Martinez,N.M., Su,A., Burns,M.C., Nussbacher,J.K., Schaening,C., Sathe,S., Yeo,G.W. and Gilbert,W.V. (2022) Pseudouridine synthases modify human pre-mRNA co-transcriptionally and affect pre-mRNA processing. *Mol. Cell*, **82**, 645–659.e9.

5. Dai,Q., Zhang,L.-S., Sun,H.-L., Pajdzik,K., Yang,L., Ye,C., Ju,C.-W., Liu,S., Wang,Y., Zheng,Z., *et al.* (2023) Quantitative sequencing using BID-seq uncovers abundant pseudouridines in mammalian mRNA at base resolution. *Nat. Biotechnol.*, **41**, 344–354.

6. Li,X., Zhu,P., Ma,S., Song,J., Bai,J., Sun,F. and Yi,C. (2015) Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. *Nat. Chem. Biol.*, **11**, 592–597.

7. Khoddami,V., Yerra,A., Mosbruger,T.L., Fleming,A.M., Burrows,C.J. and Cairns,B.R. (2019) Transcriptome-wide profiling of multiple RNA modifications simultaneously at single-base resolution. *Proc. Natl. Acad. Sci. U. S. A.*, **116**, 6784–6789.

8. Carlile,T.M., Rojas-Duran,M.F., Zinshteyn,B., Shin,H., Bartoli,K.M. and Gilbert,W.V. (2014) Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature*, **515**, 143–146.

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

10. Zhang,M., Jiang,Z., Ma,Y., Liu,W., Zhuang,Y., Lu,B., Li,K., Peng,J. and Yi,C. (2023) Quantitative profiling of pseudouridylation landscape in the human transcriptome. *Nat. Chem. Biol.*, **19**, 1185–1195.

11. Aird,D., Ross,M.G., Chen,W.-S., Danielsson,M., Fennell,T., Russ,C., Jaffe,D.B., Nusbaum,C. and Gnirke,A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.

12. Workman,R.E., Tang,A.D., Tang,P.S., Jain,M., Tyson,J.R., Razaghi,R., Zuzarte,P.C., Gilpatrick,T., Payne,A., Quick,J., *et al.* (2019) Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods*, **16**, 1297–1305.

13. Tavakoli,S., Nabizadeh,M., Makhamreh,A., Gamper,H., McCormick,C.A., Rezapour,N.K., Hou,Y.-

M., Wanunu,M. and Rouhanifard,S.H. (2023) Semi-quantitative detection of pseudouridine modifications and type I/II hypermodifications in human mRNAs using direct long-read sequencing. *Nat. Commun.*, **14**, 334.

14. Huang,S., Zhang,W., Katanski,C.D., Dersh,D., Dai,Q., Lolans,K., Yewdell,J., Eren,A.M. and Pan,T. (2021) Interferon inducible pseudouridine modification in human mRNA by quantitative nanopore profiling. *Genome Biol.*, **22**, 330.

15. Begik,O., Lucas,M.C., Pryszcz,L.P., Ramirez,J.M., Medina,R., Milenkovic,I., Cruciani,S., Liu,H., Vieira,H.G.S., Sas-Chen,A., *et al.* (2021) Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat. Biotechnol.*, **39**, 1278–1291.

16. Liu,H., Begik,O., Lucas,M.C., Mason,C.E., Schwartz,S., Mattick,J.S., Smith,M.A. and Novoa,E.M. Accurate detection of m6A RNA modifications in native RNA sequences. 10.1101/525741.

17. Leger,A., Amaral,P.P., Pandolfini,L., Capitanchik,C., Capraro,F., Miano,V., Migliori,V., Toolan-Kerr,P., Sideri,T., Enright,A.J., *et al.* (2021) RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nat. Commun.*, **12**, 7198.

18. Hendra,C., Pratanwanich,P.N., Wan,Y.K., Goh,W.S.S., Thiery,A. and Göke,J. (2022) Detection of m6A from direct RNA sequencing using a multiple instance learning framework. *Nat. Methods*, **19**, 1590–1598.

19. Pratanwanich,P.N., Yao,F., Chen,Y., Koh,C.W.Q., Wan,Y.K., Hendra,C., Poon,P., Goh,Y.T., Yap,P.M.L., Chooi,J.Y., *et al.* (2021) Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. *Nat. Biotechnol.*, **39**, 1394–1402.

20. Jenjaroenpun,P., Wongsurawat,T., Wadley,T.D., Wassenaar,T.M., Liu,J., Dai,Q., Wanchai,V., Akel,N.S., Jamshidi-Parsian,A., Franco,A.T., *et al.* (2021) Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Res.*, **49**, e7.

21. Lorenz,D.A., Sathe,S., Einstein,J.M. and Yeo,G.W. (2020) Direct RNA sequencing enables m6A detection in endogenous transcript isoforms at base-specific resolution. *RNA*, **26**, 19–28.

22. Nguyen,T.A., Heng,J.W.J., Kaewsapsak,P., Kok,E.P.L., Stanojević,D., Liu,H., Cardilla,A., Praditya,A., Yi,Z., Lin,M., *et al.* (2022) Direct identification of A-to-I editing sites with nanopore native RNA sequencing. *Nat. Methods*, **19**, 833–844.

23. McCormick,C.A., Akeson,S., Tavakoli,S., Bloch,D., Klink,I.N., Jain,M. and Rouhanifard,S.H. (2023) Multicellular, IVT-derived, unmodified human transcriptome for nanopore direct RNA analysis. *bioRxiv*, 10.1101/2023.04.06.535889.

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

25. Fanari,O., Tavakoli,S., Akeson,S., Makhamreh,A., Nian,K., McCormick,C.A., Qiu,Y., Bloch,D., Jain,M., Wanunu,M., *et al.* (2024) Probing enzyme-dependent pseudouridylation using direct RNA sequencing to assess neuronal epitranscriptome plasticity. *bioRxiv*, 10.1101/2024.03.26.586895.

26. Huang,S., Wylder,A.C. and Pan,T. (2024) Simultaneous nanopore profiling of mRNA m6A and

pseudouridine reveals translation coordination. *Nat. Biotechnol.*, 10.1038/s41587-024-02135-0.

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

28. Giambruno,R., Zacco,E., Ugolini,C., Vandelli,A., Mulroney,L., D'Onghia,M., Giuliani,B., Criscuolo,E., Castelli,M., Clementi,N., *et al.* (2023) Unveiling the role of PUS7-mediated pseudouridylation in host protein interactions specific for the SARS-CoV-2 RNA genome. *Mol. Ther. Nucleic Acids*, **34**, 102052.

29. Schwartz,S., Bernstein,D.A., Mumbach,M.R., Jovanovic,M., Herbst,R.H., León-Ricardo,B.X., Engreitz,J.M., Guttman,M., Satija,R., Lander,E.S., *et al.* (2014) Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell*, **159**, 148–162.

30. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.

31. Morello,L.G., Hesling,C., Coltri,P.P., Castilho,B.A., Rimokh,R. and Zanchin,N.I.T. (2011) The NIP7 protein is required for accurate pre-rRNA processing in human cells. *Nucleic Acids Res.*, **39**, 648–665.

32. Dang,T., Jang,S.H., Back,S.H., Park,J.W. and Han,I.-S. (2018) DRG2 Deficiency Causes Impaired Microtubule Dynamics in HeLa Cells. *Mol. Cells*, **41**, 1045–1051.