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

Caroline A. McCormick[1], Michele Meseonznik[1], Yuchen Qiu[1], Oleksandra Fanari[1], Mitchell Thomas[1], Yifang Liu[1], Dylan Bloch[1], Isabel N. Klink[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 via psi synthases, including *TRUB1* and *PUS7*. Nanopore direct RNA sequencing combined with our recently developed tool, Mod-*p* ID, enables psi mapping, transcriptome-wide, without chemical derivatization of the input RNA and/or conversion to cDNA. This method is sensitive for detecting differences in the positional occupancy of psi across cell types, which can inform our understanding of the impact of psi 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, we find that conserved psi positions on mRNAs produce higher levels of protein than expected, suggesting a role in translation regulation. Interestingly, we identify cell type-specific sites of psi modification in ubiquitously expressed genes. Finally, we characterize transcripts with multiple psi modifications and found that these psi sites can be conserved or cell type-specific, including examples of multiple psi modifications within the same motif. Our data suggest that psi modifications contribute to regulating translation 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 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 that outlines psi positions and occupancy levels 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. 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 chemical 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. Nanopore DRS employs a motor protein sequencing adapter that allows for the voltage-driven translocation of RNA through a biological nanopore. A characteristic disruption of the ionic current is associated with each canonical nucleotide during basecalling allowing for the conversion of a strands current signal to a corresponding local

sequence. Because an RNA strand 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 levels(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. 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 in a method called Mod-*p* ID to assign psi-sites across the transcriptome(13, 23). U-to-C basecalling errors are sequence-dependent, and therefore, cannot be used for the absolute quantification of psi occupancy or the comparison of modification profiles of distinct transcripts(24). However, Mod- *p* ID can quantitatively determine both the magnitude and directionality of changes based on the relative psi occupancy of the same site across cell lines or biological conditions. Furthermore, DRS has the potential to detect multiple modifications on a single transcript(25).

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 and measure the correlation of psi modification with corresponding protein expression using mass spectrometry (MS). Finally, we detect transcripts bearing multiple modifications and show that double modifications can occur within the same motif. Our differential approach provides insight into specialization in RNA modifications across cell types transcriptome-wide.

Materials & Methods

Accessing publicly available data sets. Nanopore DRS raw fast5s for all IVTs were sourced from NIH NCBI SRA BioProject accession PRJNA947135(26). Nanopore DRS raw fast5 files 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(27).

Cell culture. HeLa, HepG2, A549, and NTERA-2 cells were cultured in DMEM (Gibco, 10566024); Jurkat cells were cultured with RPMI 1640 (Gibco™ 11875093); Jurkat cells were cultured in RPMI 1640 (Cat No. 11-875-093) and SH-SY5Y cells were cultured in 1:1 EMEM: F12 (Fisher Scientific, 50-983-283, Cytiva, SH3002601). All media was supplemented with 10% Fetal Bovine Serum (Fisher Scientific, FB12999102) and 1% Penicillin-Streptomycin (Lonza,17602E). Cells were cultured at 37°C with 5% CO_2 in 10 cm tissue culture dishes and allowed to reach ~80% confluence.

Total RNA extraction and poly(A) selection. Total RNA extraction from cells and Poly(A) selection was performed using a previously established protocol(13). 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(13).

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). Jurkat and NTERA cells were sequenced on the PromethION while all other cell lines were sequenced on the MinION.

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 S1**). 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 spectrophotometer. 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**.

Protein Extraction and LC-MS/MS. Cell lysis buffer was prepared using 8M urea (Fisher, Cat no: U15-500), 40mM Tris-HCl, pH 7.0 (Fisher, Cat no: BP1756-100), and Pierce Protease Inhibitor Tablets (Thermo Scientific, Cat no: PIA32963). Cells from six cell lines—A549 (3.492 x 10^7 cells), HeLa (3.240 x 10^7 cells), HepG2 (2.946 x 10^7 cells), Jurkat (7.070 x 10^7 cells), SH-SY5Y (7.236 x 10^7 cells), and NTERA-2 (1.392 x 10^7 cells)—were lysed using this buffer. The lysates were incubated on ice for 30 minutes and then centrifuged at 20,000 x g for 15 minutes at 4°C. Supernatants were collected, and protein concentrations were measured using the Qubit Protein Assay Kit (Invitrogen by Thermo Fisher Scientific, Cat no: Q33211).

100 µg of protein per sample from three biological replicates per group was taken for proteomics analysis. Samples were reduced with 10 mM TCEP at 56°C for 60 mins and alkylated using 20 mM iodoacetamide at RT for 30 mins. Before digestion, proteins were precipitated with acetone.

The protein pellet was re-suspended in 100 mM Triethylamonium bicarbonate (TEAB) and digested with MS-grade trypsin (Pierce TM) overnight at 37°C. Peptide digest was quantified using Thermo Scientific™ Pierce™ Quantitative Colorimetric Peptide Assay (Product No. 23275). An equal amount of peptide was labeled with isobaric TMT reagents Thermo Scientific™ TMT10plex (Pierce; Rockford, IL, USA) according to manufacturer protocols (Thermo Scientific TM). A small fraction of the sample from each isobaric tag channel was pooled and analyzed on LC-MS to obtain the median for all the reporter ion intensities. TMT-labeled peptides were pooled and cleaned with Pierce™ C18 spin columns (Thermo Scientific™) and were re-suspended in 2% acetonitrile (ACN) and 0.1% formic acid (FA). 1 µg of multiplexed sample was loaded onto in-house pull tip 75µm x 20 cm C18 ReproSil-Pur 120 1.9 µm (Dr. Maisch) LC Column, then separated with a Thermo RSLC Ultimate 3000 (Thermo Scientific™) with a 160 min gradient of 2–35% solvent B (0.1% FA in ACN) at 200 nL/min and 25oC with a 220 min total run time. Eluted peptides were analyzed by a Thermo Orbitrap Q Exactive (Thermo Scientific™) mass spectrometer in a data-dependent acquisition mode. A full survey scan MS (from m/z 350–1500) was acquired in the Orbitrap with a resolution of 70,000. The AGC target for MS1 was set as 3 × 106, and the ion filling time was set as 50 ms. The most intense 15 ions with charge states 2-6 were isolated and fragmented using HCD fragmentation with 33 % normalized collision energy and detected at a mass resolution of 35,000 at 200 m/z. The AGC target for MS/MS was set as 1 × 105, and the ion filling time was set as 120 ms. Dynamic exclusion was set for 30 s with a 0.7 m/z isolation window.

Proteomics Analysis. Raw LC-MS/MS data were loaded into MaxQuant(28) v2.4.2.0 and the peptides were identified with the built-in Andromeda search engine using a FASTA file containing all entries from the SwissProt Homo sapiens (human) database (20,218 proteins) supplemented with common contaminants.

For all searches, carbamidomethylated cysteine was set as a fixed modification and oxidation of methionine and N-terminal protein acetylation as variable modifications. Trypsin/P was specified as the proteolytic enzyme. Precursor tolerance was set to ±10 ppm, and fragment ion tolerance to ±20ppm. For modified peptides, we used the default cutoffs of at least 40 for the Andromeda score and 6 for the delta score. The false discovery rate was set at <1% for peptide spectrum matches and protein group identification employing a target-decoy approach.

Reporter ion isotropic distributions from the TMT product data sheet were input into the software to correct for labeling. The peptides were filtered to have <0.05 PEP (posterior error probability) and all the contaminants were excluded. We filtered the peptides to be present (Reporter ion intensity > 0) in all replicates of each cell line. The median reporter ion intensities were then calculated to account for the relative lengths of the proteins for quantification. The MaxQuant output can be found in **Supplementary Table S2**. Normalized protein abundance and corresponding mRNA TPMs can be found in **Supplementary Table S3.**

Bootstrapping. Because we compared MinION and PromethION runs, which generate substantially different numbers of reads, we performed computational resampling to generate ten *in silico* 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 from 30 direct reads to 10 direct reads at a query 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.

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 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; **Supplementary Table S4**. The higher number of reads observed in NTERA and Jurkat is attributed to the use of the PromethION for sequencing one or all the replicates for these cell lines. Following basecalling and genomic alignment, we mapped putative psi positions across the transcriptome for each cell line using our previously described Mod-*p* ID pipeline(13), which compares our direct RNA libraries to a reference *in vitro* transcribed (IVT) library(13, 26). 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 comprising merged unmodified transcriptomes of different cell lines can enrich data sets where paired IVTs do not have sufficient coverage(26). We used stringent filtering criteria to reduce the possibility of false positives in psi identification: 1) IVT control has ≥ 10 reads, 2) IVT error at the U site 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 in the direct library to be ≥ 10%. We identified putative psi sites for each cell line (**Figure 1c; Supplementary Table S5**). To find the relative ratio of putative psi sites to total uridines for a cell line, we calculated the number of putative psi sites detected by Mod-*p* ID and normalized by the number of aligned genomic uridines for each cell line. The values obtained from this normalization are 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 psi sites to total uridines across cell lines, while HepG2 and SH-SY5Y cells had the lowest (**Fig 1c**).

mRNA 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 S6**). 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 each 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(29); for *PUS7,* we searched for the motif UNUAR(30). We computed the proportion of Ψ positions with the motif divided by the total number of positions with the motif expressed in the transcriptome and deduced from gene models of the genome,

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

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

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 for both PUS7 and TRUB1 (**Figure 1e**). These findings are consistent with our high, observed global psi levels 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. When comparing the proportion of psi within each of these motifs to the respective enzyme mRNA expression levels, we found no correlation between PUS mRNA expression levels in each 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 S4**).

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 S7**), 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 *PUS*(24, 27), CMC-mediated Illumina sequencing methods, and RNA bisulfite labeling methods(5–7, 10, 31) produced a list of 17 orthogonally validated and conserved psi sites (**Figure 2a; Supplementary Table S5**). Of these, we mapped 11 sites to the CDS and 6 to the 3' UTR (**Supplementary Figure S2, Supplementary Table S8**). Interestingly, the *TRUB1* motif (GUUCN) predominantly appeared in 14 of our 17 highly conserved sites, and psi modification at those sites was additionally validated by a *TRUB1* siRNA knockdown experiment (27). 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 the standard deviation (SD) for each psi position to compare across all six cell lines (**Figure 2a**). The conserved psi site located on *NIP7* (chr16:69342144) has the smallest SD 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(27). Conversely, 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(27).

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; Supplementary Figure S2**). *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 TPMs) but one of the lowest psi positional occupancies at 42%. Conversely, 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 79%, and Jurkat cells have the lowest expression and psi occupancy at 37±4 and 32%, respectively (**Figure 2a; Supplementary Figure S2**).

Protein expression of transcripts with conserved psi sites

To assess post-transcriptional regulation by psi, we performed a joint mRNA and protein analysis for the conserved positions. This was an ideal group for which to perform this analysis for because these positions were found within highly expressed targets, and the levels of psi (based on U-to-C error) were high. We performed quantitative mass spectrometry on all 6 cell lines, labeling each with a unique isobaric tandem-mass-tag (TMT) (see Methods; **Supplementary Table S2**).

We first analyzed all measured transcripts, agnostic to the presence of psi, to visualize the full proteome distribution for each of the 6 cell lines (**Figure 2b; Supplementary Table S3**).

Overall, the mRNA levels are weakly correlated with corresponding protein abundance. Additionally, the genes represented by NTERA cells had lower corresponding protein expression than the other 5 cell lines.

Next, to assess post-transcriptional regulation by psi we applied several stringent criteria to qualify for analysis: at least one genomic uridine position on a given transcript must have 30 reads and contain a conserved psi site with a minimum of 10% direct U-to-C mismatch in all six cell lines. Filtering with these criteria resulted in 68 unique transcripts. We further filtered these targets to include only those with a protein reading in each of three replicates resulting in 22 targets, with psi found primarily in the CDS (**Figure 2b, Supplementary Figure S3, Supplementary Table S3**). Interestingly, ~68% of the 68 psi-bearing mRNA targets did not have protein detected despite high levels of mRNA expression (>30 TPMs) compared to 87% of non-psi-bearing transcripts that did not have protein detected despite high levels of mRNA expression (>30 TPMs). This suggests that psi-bearing mRNA targets may have higher protein expression levels because they are more frequently exceeding the minimum threshold of detection. To confirm this observation, we compared the relative lengths of the proteins detected and found that the median protein lengths of the modified and unmodified groups are similar. We then ascertained whether 32% (proteins found in the psi-bearing group) is statistically different from 13% (proteins found in the non-psi bearing group) by performing a simulation in which 68 genes were randomly sampled from a dataset containing gene names of mRNA transcripts with more >30 reads. A total of 10 million randomly selected sets of 68 were generated and cross referenced to a dataset containing genes for which proteins were identified via mass spectrometry in all 3 replicates of the 6 cell lines. The 95% confidence interval for the mean and standard deviation of this distribution is [13.869;13.874] and [4.148; 4.153] which suggests that the empirically observed difference of 32% and 13% between the modified and unmodified populations was not due to sample size imbalance and supports the observation that pseudouridylation impacts protein expression.

Consistent with this observation, we noted an increase in protein expression for psi-bearing transcripts of similar expression levels compared to the protein expression of non-psi-bearing transcripts. This is visualized by the average position of psi transcripts on the y-axis (TPMs) in the scatter plot of **Figure 2b** (middle plot), yet higher-than-average protein count values along the x-axis of the same scatter plot. To account for the sample size imbalance between the modified (68) and unmodified (3400) transcripts, a down sampling simulation was performed. Both datasets were pooled and 10 million sets of 68 randomly sampled genes were generated. Each random set of 68 genes was subsequently cross referenced to a comprehensive dataset containing all transcripts with protein detected in all 3 replicates of 6 cell lines, irrespective of modification status. The portion of transcripts for which a protein was detected in each set was then calculated. The resulting 95% confidence interval for the mean and standard deviation of this distribution is [13.871; 13.877] and [4.148; 4.153], respectively. This suggests that the empirically observed 32% of proteins belonging to modified transcripts which met the detection threshold, compared to 13% for unmodified transcripts, is not due to random chance and the difference between the two groups is statistically significant.

We also observed two separate populations for transcripts with a conserved psi, specifically a population of targets with high mRNA levels and relatively lower protein expression than

expected based on the mRNA levels (**Figure 2b**, middle). We observed that the targets in the top population were mainly encoding ribosomal proteins. To test the hypothesis that mRNAs encoding ribosomal proteins typically have high mRNA expression level yet lower protein counts, we plotted all mRNAs encoding ribosomal proteins including those without psi on them (**Figure 2b**, right) and observed a tight population that has a significantly higher trendline than the global correlation trendline for all mRNAs.

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(29) and found that the stem and loop structure on mRNA within the TRUB1 motif is sufficient for mRNA pseudouridylation. Thus, we aimed to assess whether these structural features were present on sites where psi is conserved across cell lines. Using Mfold(32), we took a 21-mer sequence flanking the conserved and validated psi positions within a TRUB1 motif (14 targets) and predicted the RNA secondary structure (**Figure 2c,d; Supplementary Table S9**). The 21-mer sequence was chosen because the predictions are more accurate with shorter sequences(32). Additionally, we randomly selected 14 control sites for each target, a total of 196 sites with the same motif but bearing no psi modification (**Figure 2e, f; Supplementary Table S9**).

In comparing the structural characteristics of these groups, we observed that the control sites exhibit shorter loop lengths on average compared to the psi sites ($p = 0.0369 < 0.05$). While the stem lengths were not significantly different between control sites and psi sites ($p = 0.9723$), the 5mer motif for psi sites tends to be closer to the 5' end of the stem than in the control sites ($p = 0.0016 < 0.05$). Conversely, the 5mer motif for controls is closer to the 3' end of the stem than psi targets, and this difference is significant ($p < 0.0001$; **Figure 2g**).

We randomly selected 14 control sites from the 196 total controls to directly compare with the 14 psi targets (**Figure 2d,f)**. Interestingly, some control sites displayed structural variations distinct from those observed in psi targets. For instance, SPR68 exhibits its 5-mer motif beyond the 5' end, while NACA positions its first nucleotide of the 5-mer at the 3' end of the stem, with four nucleotides extending behind the stem. In contrast, all psi targets exhibit the 5-mer motif positioned within the stem, loop, or both.

Importantly, all psi targets displayed their U/psi positioned within the loop of hairpin structures for the 21-mer structures, whereas only 7 of the 14 control sites had U/psi positioned within the loop. Additionally, 3 of the 14 control sites contained internal loops, a feature not observed in psi targets. We also compared the U/psi position within 31-mer and 41-mer structures for the 14 psi targets and 14 controls. For psi targets, 12 out of 14 have their psi sites on the loop for 31-mer and 41-mer structures (**Figure 2d**; **Supplementary Table S9, Supplementary Figure S4-5**). Therefore, longer sequences do not significantly change the preference for loop locating of psi sites. As for the U controls, 3 U/psi sites are located on the loop for the 31-mer structures and 4 for the 41-mer structures. Both 31-mer and 41-mer structures have 8 U/psi sites located on the stem. In summary, for all lengths of structures, U/psi sites prefer to be located on the loop of psi targets. Conversely, the control sites appear in the loop less often.

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 were interested to see whether psi sites were present only in specific cell lines, even if 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 S10**). 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), meaning all identified sites had <40% U-to-C mismatch, indicating <50% psi occupancy. We compared all aligned reads' raw ion current traces at the psi position, revealing that psi presence affects the ion current distributions (**Figure 3b**).

The unique absence of a psi site for a single cell line may implicate cell type-specific functionality. To study this, we had to change our selection criteria from what was used to analyze "housekeeping" psi targets (i.e., 30 direct reads and >30% U-to-C error in every cell line). 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 S11**). 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. SH-SY5Y demonstrates the highest occurrence of uniquely absent sites, with only 6 psi sites identified. 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 confirmed to be psi by two orthogonal methods(5, 10).

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, 7 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),

458 *SLC25A1* (chr22:19177964), and *IPO7* (chr11:9445686). Additionally, we identified 14 putative
459 psi sites using Mod-*p* ID which have >30% U-to-C errors (**Figure 3e**). For all sites of unique
460 modification presence, we ruled out the presence of a single-nucleotide variant in these cell lines
461 by Sanger sequencing on the corresponding gDNA (**Figure 3f; Supplementary Figure S4**). We
462 wanted to confirm that these sites deviated from the expected ionic current, which indicates a
463 probable modification. We observed that the pan-IVT trace and the other five cell lines have
464 similar ionic current trace distributions and a visible deviation of the cell line distribution with
465 the modification (**Figure 3g**).
466
467 **Type II hypermodification of transcripts can be variable or shared across cell lines.**
468 Type II psi hypermodification has been previously defined as transcripts with more than one psi
469 modification(13). To assess the type II hypermodification status across cell types, we filtered to
470 require ten direct reads, and the paired-IVT had ten reads at the query position (**Supplementary**
471 **Table S12**). We detected up to 5 modified positions per transcript for A549 cells, HeLa cells,
472 HepG2 cells, Jurkat cells, and NTERA cells. We detected the presence of many transcripts with
473 two modifications in all cell lines. A549 cells, NTERA cells, and Jurkat cells consistently had the
474 highest detected type II hypermodified transcripts. In contrast, SH-SY5Y cells had the lowest
475 number of hypermodified transcripts as we only detected up to 3 positions on one transcript
476 (**Figure 4a**).
477
478 *PSAP* encodes the prosaposin gene and is one example of a transcript with multiple positions
479 orthogonally confirmed (chr10:71816374(7, 10), 71816664(5–7, 10, 31), 71817048(10), and
480 71819612(10)) and detected at varying positional occupancy with our Mod-*p* ID method (**Figure**
481 **4b**). The encoded protein is essential for the proper functioning of the lysosome. Of the positions
482 identified, chr10:71819612 is in the CDS, while the other three are in the 3' UTR. None of these
483 confirmed sites have a known PUS motif. All 4 psi locations are present in HeLa cells, 3 in A549
484 cells, 2 in NTERA, 1 in HepG2, and 0 in Jurkat and SH-SY5Y. The positional occupancies for
485 detected psi sites are relatively low, with a maximum of 33% in HepG2 at chr10:71819612.
486 Mod-*p* ID also identified a 5th putative position at chr10:71816662 in HeLa and NTERA cells.
487 This additional putative psi site falls within the same motif of chr10:71816664, a putative double
488 modification (**Figure 4b**).
489
490 *LAPTM4B* (chr8:97776038, 97776040) is another example of a double modification, where
491 chr8:97776038 has a TRUB1 motif, GUUCU (**Figure 4c**), which has been orthogonally
492 confirmed with both PRAISE(10) and CeU-seq(6). We confirmed with Sanger sequencing that
493 neither position was an SNV (**Supplementary Figure S4**). CMC and bisulfite-based methods for
494 psi detection cannot detect two neighboring modifications. Thus, there are no known examples
495 of double modifications. Double modifications can also potentially disrupt current traces in
496 direct RNA sequencing. To explore further, we performed a strand analysis of all six cell lines'
497 combined reads (pan-direct) to see if the double modification occurred simultaneously in both
498 positions on a given read. A read must span both positions and only contain a U or C to be
499 included in the analysis. We separated the 1,823 reads into four groups: i) no modification (1,619
500 reads), ii) U-to-C error in the first position and an unmodified second position (76 reads), iii) an
501 unmodified first position and U-to-C error in the second position (124 reads), and iv) double
502 modification (4 reads). We wanted to visualize the ionic current distribution at these two
503 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 motif, but the mechanism remains unknown (**Figure 4d**).

Discussion

Our recently developed technique, Mod-*p* ID, enables psi site mapping and differential analysis of psi occupancy levels across the human transcriptome. 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 normalized to the number of reads (**Figure 1**). Many of the detected sites that do not fall within a known motif are orthogonally validated by other biochemical methods; therefore, these are believed to be psi. Other sites that have not been discovered could either be more easily detectable by nanopore sequencing or we are detecting another unknown uridine modification. For this reason, we are careful to call these positions "putative psi".

Next, looking at conserved psi-containing housekeeping genes of high mRNA expression levels we find that most sites are substrates for the TRUB1 enzyme (>80%), 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 standard deviation to quantify this. The conserved psi sites with the most similar relative occupancy 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(33), 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 the largest differences in occupancy, *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(29)(**Figure 2**). We observed deviations in the length of the loop and stem for sites that did not have a detected psi to support this hypothesis.

Functional studies of psi have shown the modification can influence translation efficiency and induce changes in the translation of the mRNA itself(34–36). However, little is known about the direct impact on the proteins, such as abundance or amino acid substitution. To assess protein abundance in psi modified transcripts, we limited our analysis to mRNAs that have a site with high psi occupancy in all 6 cell lines. When we compared transcripts corresponding to a conserved psi site against unmodified transcripts, we found that the protein levels are significantly higher for psi containing mRNAs, supporting the hypothesis that psi regulates translation, in this case, leading to higher protein expression than expected for high psi occupancy transcripts. This result is complementary to previous studies suggesting that psi containing transcripts have higher translation efficiency based on polysome profiling(25). The isoform specific contributions towards this mechanism are unknown because in the present

550 study, psi is assigned to a position using a genomic alignment rather than an alignment to
551 mRNA. Future studies will clarify the mechanism by which psi modulates translation.
552
553 We also observed positions susceptible to pseudouridylation in specific cell lines while
554 remaining unmodified in others (**Figure 3**). We consistently observe lower positional occupancy,
555 typically ranging from 10% to 50%, at psi sites detected in two, three, four, or five distinct cell
556 lines. This makes it possible that the modification fell below the detection limit for some cell
557 lines. We carefully checked for cell line specific SNVs that could be misinterpreted as a
558 modification for these sites and excluded these from the analysis. To understand the reason for
559 the U-to-C basecalling error, we analyzed the raw signal data features to detect anomalies in
560 signal patterns. The presence of a modification causes a visible shift of the ionic current intensity
561 distribution from that of the corresponding canonical nucleotide at a given position(15). We
562 observe this shift even at a lower relative positional occupancy level for the absence of a
563 modification site in a single cell line (**Figure 3f**). Conversely, the unique presence of a putative
564 psi position in a cell line can suggest its functional importance and regulatory role in that cell
565 type. *DRG2* (chr17:18107480) is a cell type-specific psi modified by PUS7 detected in HeLa
566 cells. DRG2 helps to regulate cell growth and differentiation. Disruption in *DRG2* expression in
567 HeLa cells has been shown to alter microtubule dynamics, resulting in cell proliferation
568 changes(37).
569
570 Hypermodification type II on a single transcript varies between cell lines (**Figure 4**). While one
571 cell line may have psi sites at multiple positions on the transcript, another may exhibit
572 hypermodification at different positions on the same transcript. A special case of type II
573 hypermodification is a double modification where two modifications occur within a motif so that
574 both modifications will exist in the nanopore at the same time. For this type of biological pore,
575 approximately 11 nucleotides will fit within the region interacting with the pore, and
576 modifications that are close to each other can potentially impact the detected signal of the other.
577 It is generally accepted that mRNA modifications are scarcer than modifications in rRNAs and
578 tRNAs. However, Illumina-based technologies cannot discern neighboring modifications, so
579 there was no previous evidence that satellite modifications occur on mRNAs. We found a few
580 rare examples of these modifications. For example, we observed double modifications in
581 *LAPTM4B* in A549, HeLa, and HepG2 cell lines at chr8:97776038 and chr8:97776040 (**Figure**
582 **4**). The visible differences in the KDE of merged ionic traces may indicate that the detection of
583 double modification is not a false positive. chr8:97776038 has been orthogonally confirmed(6,
584 10). TRUB1 siRNA knockdown in HeLa cells showed the U-to-C error disappear in both
585 chr8:97776038 and chr8:97776040 compared to both the native HeLa mRNA and scrambled
586 control. We ruled out the presence of an SNV in all cell lines for these sites by Sanger
587 sequencing on the corresponding gDNA (**Supplementary Figure S1**). Overall, these studies
588 advance our knowledge of site-specific psi distribution across the human body and the biological
589 impact of individual mRNA modifications.
590
591 **Code Availability**
592 All code used in this work is publicly available at
593 https://github.com/RouhanifardLab/PanHumanPsiProfiling
594
595 **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
Proteomics data is available on MassIVE (MSV000095673) and the ProteomeXchange (PXD055082).

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). We thank the Mass Spectrometry facility associated with the Chemistry and Chemical Biology Department at Northeastern University, which provided TMT-labeling and LC-MS/MS services.

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., M.T., Y.L., M.M, and O.F. analyzed the data with guidance from S.H.R., M.J. and M.W. C.A.M. and M.M. wrote the paper with guidance from S.H.R.

FIGURES

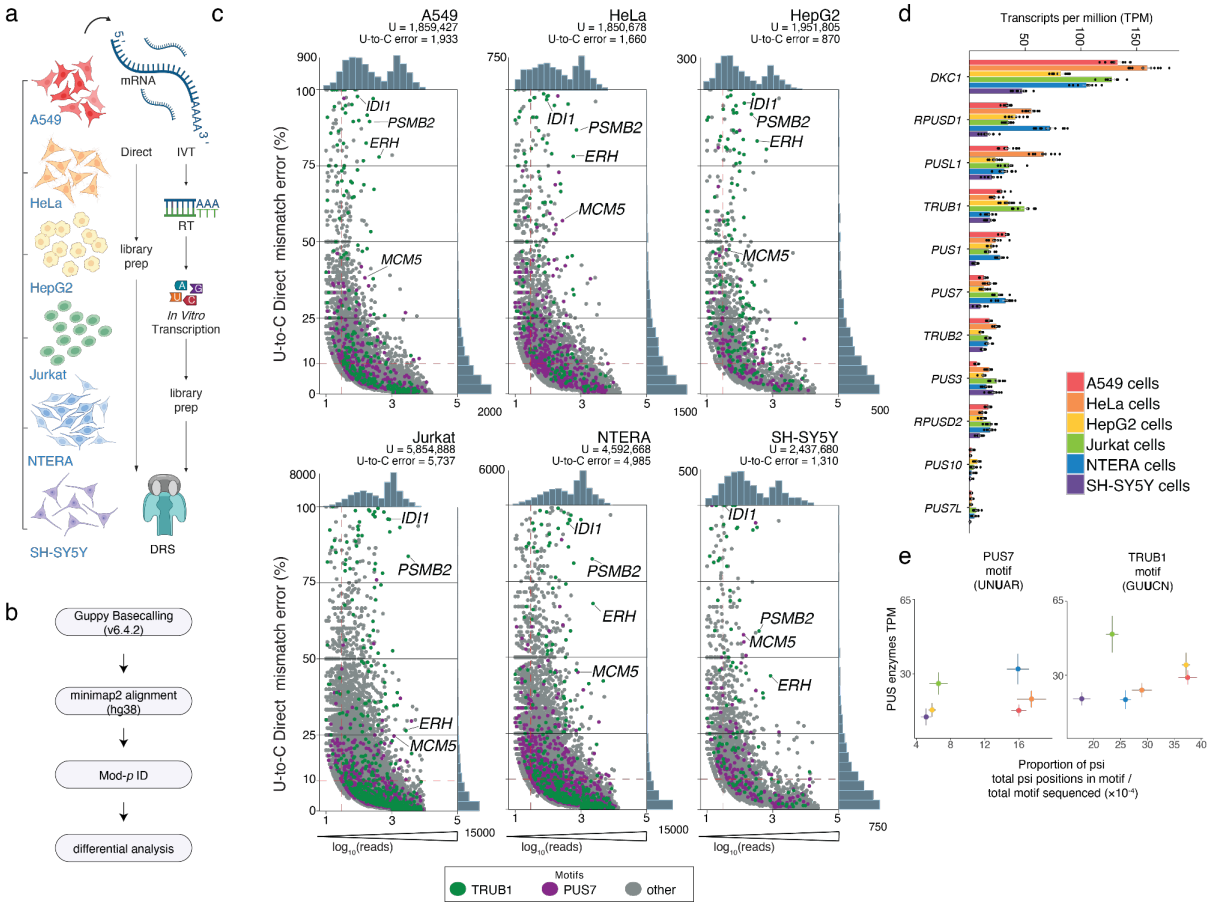

Figure 1. Individual cell line putative psi identification by Mod-*p* ID and PUS expression analysis. **a)** Experimental summary of A549, HepG2, HeLa, Jurkat, NTERA, and SH-SY5Y direct and IVT library prep and DRS. **b)** Overview of computational pipeline following DRS. **c)** Number of total direct 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 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 detected by Mod-*p* ID in a *TRUB1* or *PUS7* motif divided by the total number of *TRUB1* or *PUS7* motifs observed versus the TPMs of the PUS enzyme and standard error of the mean from computational resampling.

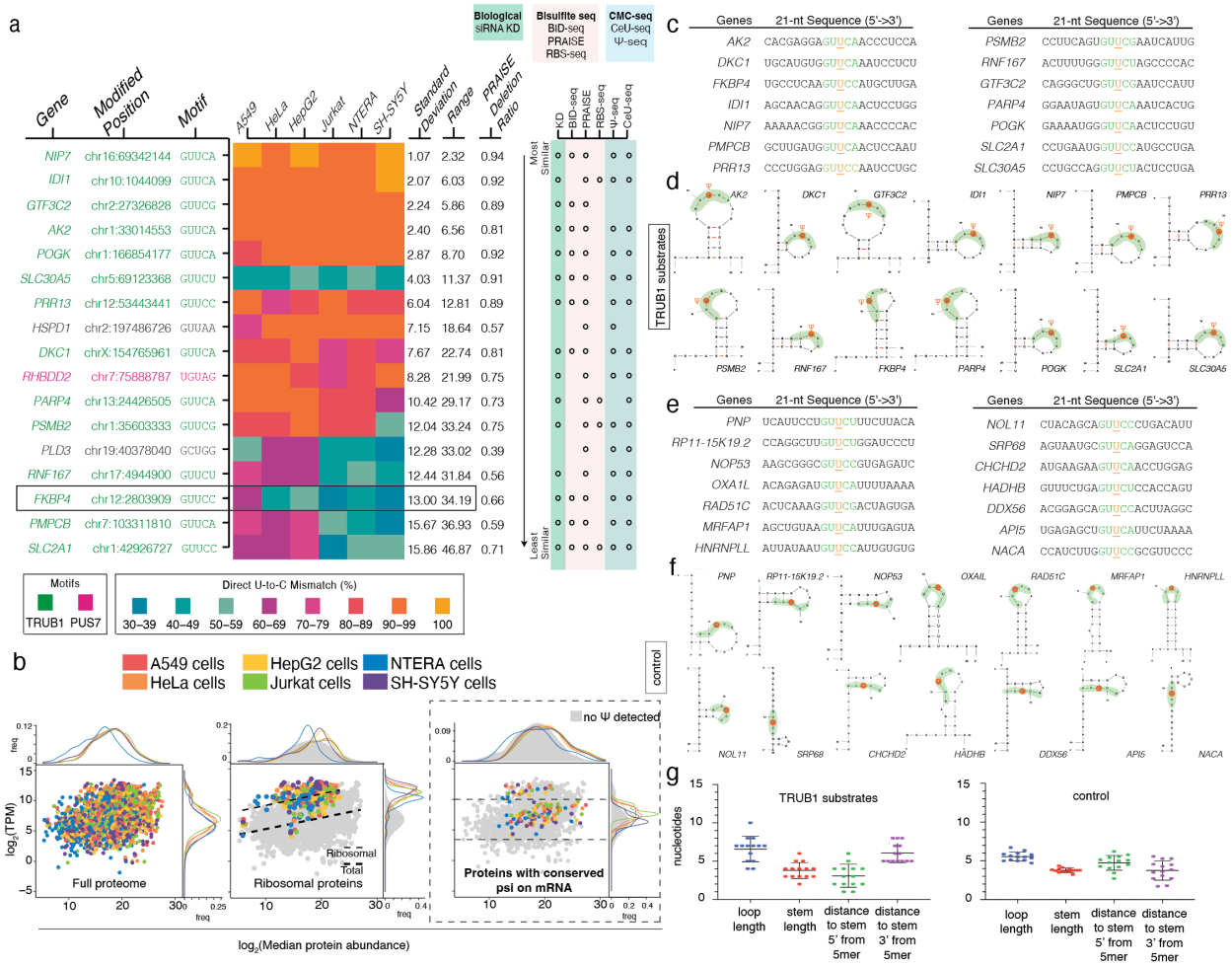

Figure 2. Conserved psi sites across 6 human cell lines. **a**) Heatmap of relative psi occupancy for sites identified by Mod-p ID[17] with ≥30 reads and ≥30% U-to-C mismatch for each cell line. Colors indicate that the psi site has a TRUB1 (green) or PUS7 (pink) motif. Standard deviation and range are reported across cell types. For the same position, the PRAISE(10) deletion ratio is reported. (right) Orthogonal methods confirming the psi site including biological (green), bisulfite based (pink) and CMC-based detection (blue). **b**) Distribution of transcripts (in TPMs) with their corresponding median protein abundance for 6 cell types (left). mRNAs encoding ribosomal proteins are plotted with their corresponding protein expression (middle). The TPM of transcripts with no modification detected in any of the 6 cell lines was plotted against their median protein abundance in grey. The TPM of transcripts with modifications detected in all 6 cell lines was plotted against median protein abundance and overlaid onto the unmodified population (shown in colors for each cell line; right). **c**) 21-nt sequence of the conserved TRUB1 psi sites. **d**) Predictive secondary structure modeling of canonical 21-nt sequences of the conserved TRUB1 psi sites. **e**) 21-nt sequence of randomly selected unmodified uridine positions **f**) Predictive secondary structure modeling of canonical 21-nt sequences of the unmodified uridines **g**) Comparison of sequence properties between 14 TRUB1 substrates and 14 control sequences. The bars show the median value with standard error.

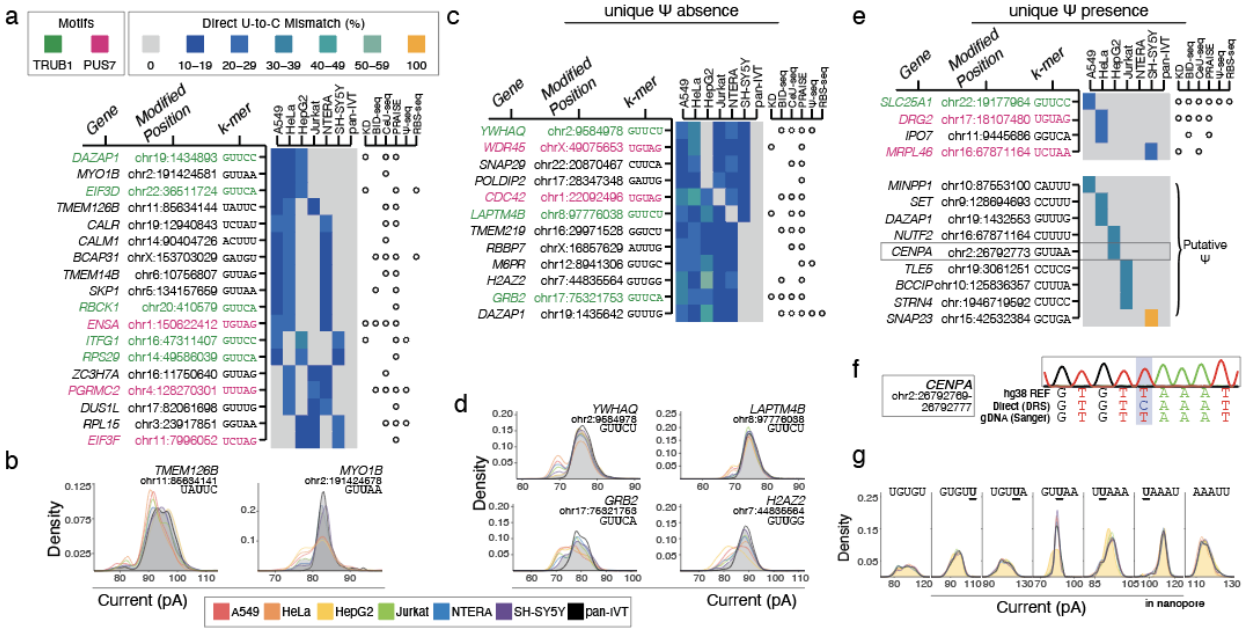

Figure 3. Cell type-specific expression of psi. **a**) Heatmap of relative psi occupancy for sites detected in 3 cell lines identified by Mod-*p* ID and confirmed by orthogonal methods. Colors indicate that psi site has a TRUB1 (green) or PUS7 (pink) motif. **b**) KDE ionic current traces of cell-type 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) 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.

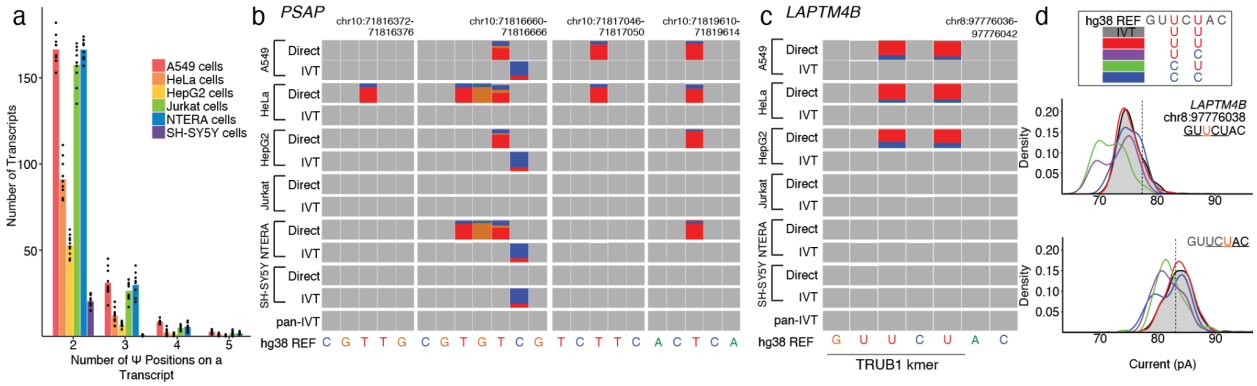

Figure 4 Type II hypermodification analysis. **a)** Number of hypermodified type II transcripts with 2, 3, 4, and 5 putative psi sites. **b)** Variable type II hypermodification across cell lines for the prosaposin gene, *PSAP*. Location on the transcript of three putative psi sites and IGV snapshots of each position's motif 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 motif on *LAPTM4B*. IGV snapshot of each position's motif for the six-cell line's direct and IVT library and pan-IVT. The first detected modification is found in *TRUB1* motif. **d)** Merged cell line KDE ionic current traces 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.

697 **REFERENCES**
698

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

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

707 3. Kierzek,E., Malgowska,M., Lisowiec,J., Turner,D.H., Gdaniec,Z. and Kierzek,R. (2014) The
708 contribution of pseudouridine to stabilities and structure of RNAs. *Nucleic Acids Res.*,
709 **42**, 3492–3501.

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

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

716 6. Li,X., Zhu,P., Ma,S., Song,J., Bai,J., Sun,F. and Yi,C. (2015) Chemical pulldown reveals
717 dynamic pseudouridylation of the mammalian transcriptome. *Nat. Chem. Biol.*, **11**, 592–
718 597.

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

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

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

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

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

734 12. Workman,R.E., Tang,A.D., Tang,P.S., Jain,M., Tyson,J.R., Razaghi,R., Zuzarte,P.C.,
735 Gilpatrick,T., Payne,A., Quick,J., *et al.* (2019) Nanopore native RNA sequencing of a
736 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. (2024) Multicellular, IVT-derived, unmodified human transcriptome for nanopore-direct RNA analysis. *GigaByte*, **2024**, gigabyte129.

24. Makhamreh,A., Tavakoli,S., Fallahi,A., Kang,X., Gamper,H., Nabizadehmashhadtoroghi,M., Jain,M., Hou,Y.-M., Rouhanifard,S.H. and Wanunu,M. (2024) Nanopore signal deviations from pseudouridine modifications in RNA are sequence-specific: quantification requires dedicated synthetic controls. *Sci. Rep.*, **14**, 22457.

25. 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.

26. 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.

27. 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. PREPRINT: NOT PEER REVIEWED

28. Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**, 1367–1372.

29. 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.

30. 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.

31. 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.

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

33. 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.

34. Eyler,D.E., Franco,M.K., Batool,Z., Wu,M.Z., Dubuke,M.L., Dobosz-Bartoszek,M., Jones,J.D., Polikanov,Y.S., Roy,B. and Koutmou,K.S. (2019) Pseudouridinylation of mRNA coding sequences alters translation. *Proc. Natl. Acad. Sci. U. S. A.*, **116**, 23068–23074.

35. Karikó,K., Muramatsu,H., Welsh,F.A., Ludwig,J., Kato,H., Akira,S. and Weissman,D. (2008) Incorporation of pseudouridine into mRNA yields superior nonimmunogenic vector with increased translational capacity and biological stability. *Mol. Ther.*, **16**, 1833–1840.

36. Karijolich,J. and Yu,Y.-T. (2011) Converting nonsense codons into sense codons by targeted pseudouridylation. *Nature*, **474**, 395–398.

37. 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.