

SHORT REPORT

Open Access



# Interferon inducible pseudouridine modification in human mRNA by quantitative nanopore profiling

Sihao Huang<sup>1</sup>, Wen Zhang<sup>1</sup>, Christopher D. Katanski<sup>1</sup>, Devin Dersh<sup>2</sup>, Qing Dai<sup>3</sup>, Karen Lolans<sup>4</sup>, Jonathan Yewdell<sup>2</sup>, A. Murat Eren<sup>4</sup> and Tao Pan<sup>1\*</sup> 

\* Correspondence: [taopan@uchicago.edu](mailto:taopan@uchicago.edu)

<sup>1</sup>Department of Biochemistry & Molecular Biology, University of Chicago, Chicago, IL 60637, USA  
Full list of author information is available at the end of the article

## Abstract

Pseudouridine ( $\Psi$ ) is an abundant mRNA modification in mammalian transcriptome, but its functions have remained elusive due to the difficulty of transcriptome-wide mapping. We develop a nanopore native RNA sequencing method for quantitative  $\Psi$  prediction (NanoPsu) that utilizes native content training, machine learning modeling, and single-read linkage analysis. Biologically, we find interferon inducible  $\Psi$  modifications in interferon-stimulated gene transcripts which are consistent with a role of  $\Psi$  in enabling efficacy of mRNA vaccines.

**Keywords:** Pseudouridine, Nanopore sequencing, Interferon, Machine learning

## Background

Pseudouridine ( $\Psi$ ) is the second most abundant mRNA modification in the mammalian transcriptome as measured by quantitative mass spectrometry [1] and may exert many cellular functions. For example,  $\Psi$  incorporation in synthetic, transfected reporter mRNA increases translation [2] through decreased activation of the RNA-dependent protein kinase (PKR) [3]. The innate immune evading property of  $\Psi$  (and its methylated derivative N<sup>1</sup>-methyl- $\Psi$ ) in mRNA is essential to the remarkable immunogenicity of successful COVID-19 mRNA vaccines [4].

Functional exploration and mechanistic investigation of mRNA  $\Psi$  modification requires appropriate mapping methods. Illumina sequencing of  $\Psi$  in mRNA relies on chemical RNA treatments that induce stop, mutation, or deletion signatures in cDNA synthesis [1, 5–8]. Many computational methods have been developed to map mRNA  $\Psi$  sites [9–20]. However, mRNA  $\Psi$  mapping is inconsistent among these studies, in part due to the high false positives and negatives generated by the chemical treatments. The read-length limitation of Illumina sequencing also narrows the possibility to examine  $\Psi$  usage in mRNA splice isoforms and the linkage of multiple  $\Psi$  sites in single molecules.



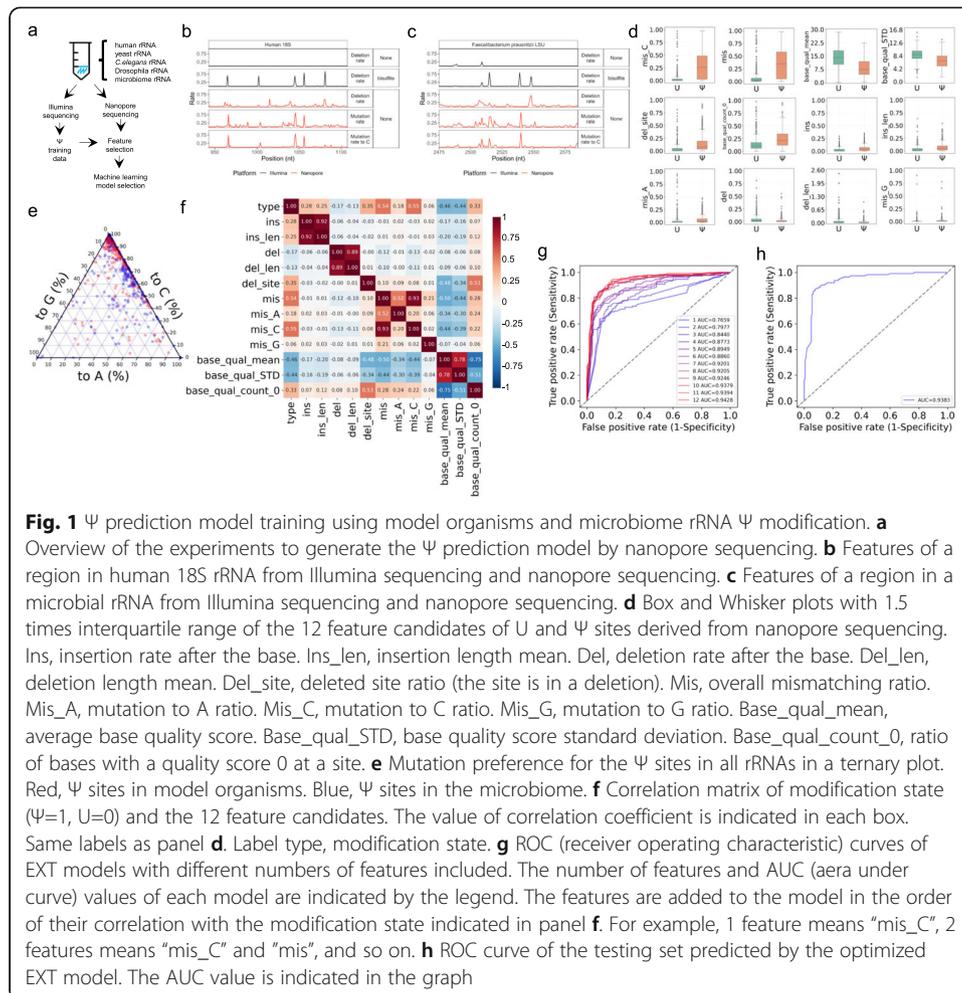
© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

The emergence of nanopore sequencing enables direct interrogation of RNA modifications [21–23]. Additionally, nanopore sequencing can extend to the full length of the mRNA [24], revealing all modified sites in single RNA isoforms [25]. Both signal strength and dwell time have been used to identify  $\Psi$  [26]. Recently, a nanopore direct RNA sequencing method, nanoRMS was developed by Novoa and co-workers that employs characteristic base-calling “error” signatures in the nanopore data for  $\Psi$  mapping [27]. NanoRMS identified new  $\Psi$  sites in mitochondrial rRNA, small nuclear RNA, small nucleolar RNA, and mRNA under normal and stress conditions in yeast and further, predicted stoichiometry via supervised learning. Although nanoRMS prediction of  $\Psi$  site incorporation using a threshold for base mismatch frequency is straightforward, it is unclear whether this approach can be applied to the mammalian transcriptomes, which are much larger than yeast, can contain introns and occur in multiple isoforms. For example, the standard Tombo software for nanopore data analysis is ineffective with spliced reads. Also, even though nanoRMS collects features from single reads, the single read features were averaged before  $\Psi$  prediction, erasing single molecule  $\Psi$  site incorporation information.

## Results and discussion

The key to nanopore identification of RNA modification is to generate training data from known modification sites. Modification training data generation, however, requires reads from long transcripts with distinct sequence contexts at the modification site. To maximize our ability to obtain nanopore training data from as many distinct  $\Psi$  sites as possible, we generated a mixture of rRNAs from human, yeast, *Caenorhabditis elegans*, *Drosophila*, and from human fecal bacteria (Fig. 1a). We Illumina sequenced half of the mixture after fragmentation, using the bisulfite reaction [8] to map rRNA  $\Psi$  sites, providing a total of 2142  $\Psi$  sites (Additional file 1: Fig. S1a, Additional file 2: Table S1). In Illumina sequencing of the bisulfite method,  $\Psi$  sites are found by RT deletions which enable identification and quantitative assessment of closely spaced rRNA  $\Psi$  sites; these sites are more difficult to assess using the more commonly used carbodiimide method that identifies  $\Psi$  sites by RT stops. Sequencing the remaining sample via direct RNA nanopore sequencing, we found that 640 of these  $\Psi$  sites passed our filter of 20 read coverage for further analysis (Additional file 1: Fig. S1b). The lower number of  $\Psi$  sites in nanopore sequencing was in part derived from the 3' bias of the nanopore sequencing library design where all reads start from the 3' end of the rRNA. These 640 sites were combined with 689 randomly chosen unmodified U sites as the training data set (Additional file 1: Fig. S1c). The modified and unmodified U sites contained 236 of the 256 NN( $\Psi$ /U)NN different 5mer contexts.

We next extracted nanopore signal features suitable for  $\Psi$  identification. NanoRMS [27] found that  $\Psi$  had negligible effect on nanopore current signals, which means it is hard to directly identify  $\Psi$  from the current squiggles like m<sup>6</sup>A [23, 25, 28]. However, distinct features could be found for  $\Psi$  identification. For instance, like NanoRMS, we found that apparent mutation to C is a prominent signature for  $\Psi$  modification, with apparent deletion also significant for some  $\Psi$  sites (Figs. 1b, c). In total, we examined 12 features of base calling errors and found that  $\Psi$  sites tend to have lower base quality mean values and standard deviation (Fig. 1d). The ternary plot of mutation signatures confirmed  $\Psi$  sites having a strong preference to be read as a C but not A or G (Fig.



1e). The significance of these features was quantified in the correlation matrix of all features and the modification labels (Fig. 1f). We made an extremely randomized trees (EXT) model to carry out  $\Psi$  probability prediction for each U site. To decide the combination of features included in the model, we added one feature at a time in the order of their correlation strength with the modification label. This revealed that the performance of  $\Psi$  calling maximized when all 12 features were included (Fig. 1g). Using the optimized parameters of our EXT model, its performance was evaluated by the testing set with an area under curve (AUC) of 0.9383 (Fig. 1h, Additional file 1: Fig. S1d). We named our method nanopore investigation of pseudouridines or “NanoPsu.”

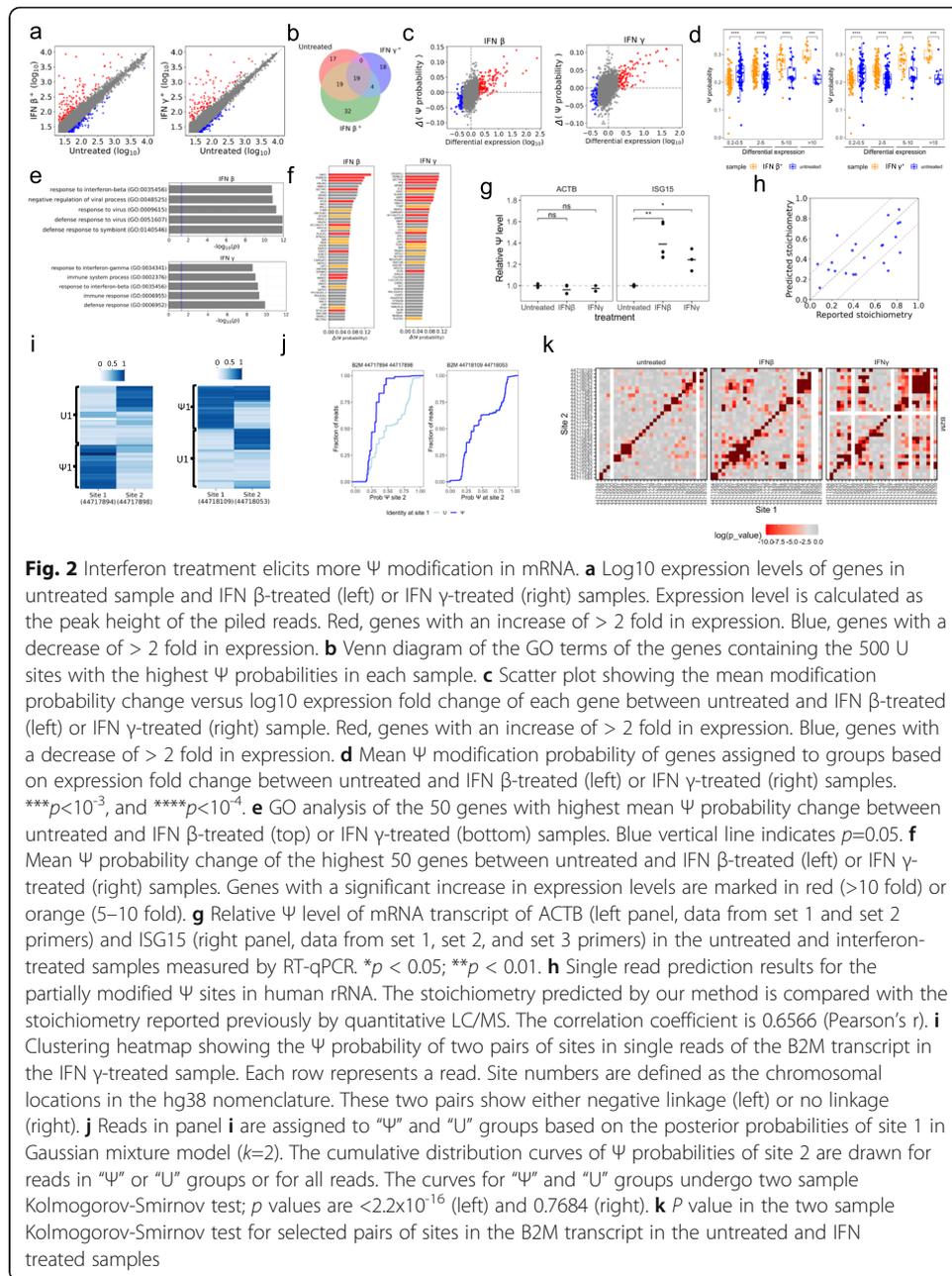
Interferons (IFNs), cytokines produced by nearly all cell types during viral and other microbial infections, play crucial roles regulating immune response [29]. mRNA vaccines incorporate  $\Psi$  or  $m^1\Psi$  to evade host cell foreign RNA sensing and enhance mRNA translation. Do endogenous mRNAs also use the same strategy through  $\Psi$  modification? IFNs can induce the expression of more than a thousand interferon stimulated gene (ISG) transcripts. ISGs include protein kinase R (PKR) which phosphorylates eIF2 $\alpha$  to reduce global translation. It is well established that  $\Psi$ -modified reporter mRNA activates PKR much less than the same unmodified mRNA and is translated at much higher levels [3]. We therefore hypothesize that

ISG transcripts may have elevated levels of  $\Psi$  modification to enhance translation in the presence of PKR.

We tested this hypothesis by treating cells with either IFN- $\gamma$  or IFN- $\beta$  followed by nanopore sequencing (Additional file 1: Fig. S2a). IFN treatments worked well as determined by upregulation of surface MHC class I (Additional file 1: Fig. S2b). The mRNA expression levels of the biological replicates were highly correlated (Additional file 1: Fig. S2c). For improved coverage, we combined the nanopore data from the biological replicates for downstream analysis (Additional file 1: Fig. S2d). We found strongly up-regulated mRNA transcripts upon IFN treatment that belong to the ISG genes with the expected gene ontology of interferon signaling pathway and viral defense (Fig. 2a, Additional file 1: Fig. S3a). These results indicate the feasibility of using nanopore sequencing to study the interferon response transcriptome.

We used the EXT model to predict  $\Psi$  modification probabilities in the transcriptome. In total, ~2.6 million U sites were analyzed in each transcriptome (Additional file 1: Fig. S3b). We found a “RA $\Psi$ U” motif and the previous revealed [30] “GU $\Psi$ C” motif among top  $\Psi$  sites in the untreated sample (Additional file 1: Fig. S3c). The  $\Psi$  sites belonging to “median” or higher groups in the previous study [30] showed significantly higher predicted  $\Psi$  probabilities than other U sites in the untreated sample (Additional file 1: Fig. S3d), indicating that our method provides valid prediction of  $\Psi$ . For the 500 sites with the highest probability of  $\Psi$  modification, the three samples shared some but also had distinct sites (Additional file 1: Fig. S3e). However, IFN treated samples had a wider range of GO terms than the untreated sample (Fig. 2b), suggesting that  $\Psi$  modification becomes more widespread to transcripts belonging to more diverse cellular processes. Going beyond the top 500 probable  $\Psi$  sites, globally the upregulated gene transcripts had higher average modification probabilities for IFN treated vs. untreated samples, with the magnitude of increase strongly correlated with the expression fold change (Figs. 2c, d, Additional file 1: Fig. S3f). Increased  $\Psi$  modification probability in a mRNA transcript could be attributed to increased number of  $\Psi$  sites and/or increased modification fraction of modified sites. The top 50 genes with highest increase in  $\Psi$  modification probability were related to the interferon pathway and anti-viral response (Fig. 2e), they included 88.5% of all genes with >10-fold increase and 60.9% of all genes with >5-fold increase in mRNA expression (Fig. 2f). These results are consistent with increased  $\Psi$  modification in the transcriptome upon interferon treatment enhancing ISG function.

We used a RT-qPCR method to validate the increased  $\Psi$  level in the ISG transcripts. Our method takes advantage of the standard  $\Psi$  detection method using N-cyclohexyl-N'-(2-morpholinoethyl) carbodiimide (CMC). The  $\Psi$ -CMC adduct introduces a RT stop in cDNA synthesis which reduces the amount of cDNA product compared to the control reaction without CMC. The differential amount of the cDNA product can then be precisely measured using real-time qPCR (scheme in Additional file 1: Fig. S3g). We first showed that the actin mRNA did not change its abundance nor its  $\Psi$  level, making it an appropriate internal control for comparing the  $\Psi$  levels of the ISG transcripts (Fig. 2g, Additional file 1: Fig. S3h, left panels).  $\Psi$  level increase in the ISG15 mRNA upon interferon treatment was validated upon normalization of its expression level and to the actin mRNA within the same sample (Fig. 2g, Additional file 1: Fig. S3h, right panels).



We performed single-read analysis for quantitative  $\Psi$  stoichiometry prediction and investigation of linking modification states of  $\Psi$  sites in single molecules of a mRNA transcript. Our training set contained the data points from previously reported [31], 100% modified human rRNA  $\Psi$  sites and randomly selected unmodified U sites. We used the same set of features and the same EXT algorithm, while we replaced all the "rate" features (like "mismatch ratio") with "indicator" features (like "mismatch-or-not") (Additional file 1: Fig. S4a). We tested the  $\Psi$  stoichiometry prediction from single reads on 22 partially modified  $\Psi$  sites (5–85%) and found that the predicted stoichiometry matched well with the previous reported stoichiometry obtained by LC/MS [31] (Fig. 2h).

A new aspect of our method is the ability to perform single-read analysis that links occurrence of multiple  $\Psi$  sites in individual mRNA transcripts. We examined whether

pairs of  $\Psi$  site modifications are linked either positively or negatively, meaning whether the modification state of site 2 is affected by the modification state of site 1 and vice versa. We selected 31 positions in the B2M transcript (which encodes the common small subunit of MHC class I molecules) and checked pairwise linkage by two sample Kolmogorov-Smirnov test. In most cases, the maximum distance  $D$  value from two sample K-S test was small (Additional file 1: Fig. S4b), which is consistent with the presence of  $\Psi$  at site 1 being independent of  $\Psi$  at site 2, these two sites are not linked. An example of a specific unlinked pair is shown in Fig. 2i, j (right panels). A few pairs of sites had high  $D$  values, but most of those were immediately adjacent  $\Psi$  sites. The pairs of  $\Psi$  sites with negative linkage tend to avoid each other in the same mRNA molecule (Additional file 1: Fig. S4c). An example of a specific negatively linked pair is shown in Fig. 2i, j (left panels) where simultaneous  $\Psi$  occurrence at both sites is very rare. This result indicates that the modification of  $\Psi$  at two sites in single molecule transcripts is negatively related for some and completely independent for others. Upon IFN treatment, the linkage between some sites in the B2M transcript became more prominent (Fig. 2k), suggesting that IFN-induced  $\Psi$  installation has stronger co-dependency.

## Conclusions

In summary, we generated a supervised-learning-based protocol to predict  $\Psi$  modification in the human transcriptome and analyzed  $\Psi$  on single reads which allows for the evaluation of stoichiometry and linkage between distal  $\Psi$  sites in the same mRNA molecule. Human genome contains 13 confirmed and putative  $\Psi$  installation enzymes [32], suggesting that  $\Psi$  installation is a highly robust and dynamic process in human cells. How these enzymes coordinate or antagonize their activities remains to be determined. We found a biological response of  $\Psi$  modification change in endogenous mRNA upon IFN treatment which is consistent with  $\Psi$  playing a role in IFN signaling pathway and viral defense.

## Methods

### Stool sample collection and total RNA extraction

Stool specimens were self-collected by 1 female volunteer using a commercial “toilet hat” stool specimen collection kit (Fisherbrand Commode Specimen Collection System; Thermo Fisher Scientific, 02-544-208). Specimens were immediately transported to the laboratory (<1h) and thoroughly homogenized. A 100 mg of stool was transferred into a cryovial using a sterile spatula, and 700  $\mu$ l RNAlater Stabilization solution was added. Specimens were stored at  $-80^{\circ}\text{C}$  until extraction.

RNAlater was first removed from stool sample by centrifugation at 17,200 rcf for 10 min at  $4^{\circ}\text{C}$ . Pelleted material was lysed in 400  $\mu$ L of 0.3M NaOAc/HOAc, 10mM EDTA, and pH 4.8 with an equal volume of acetate-saturated phenol to chloroform pH 4.5 (Invitrogen, AM9722). After addition of 1.0 mm glass lysing beads (Bio-Spec Products, 11079110) in a 1:1 ratio (bead to sample weight), samples were placed in a reciprocating bead beater (Mini-Beadbeater-16, Bio-Spec Products) for two 1-min intervals on maximum intensity. Samples were centrifuged at 17,200 rcf for 15 min at  $4^{\circ}\text{C}$  before re-extraction and isopropanol precipitation of total RNA. Pellets were washed with 75%

ethanol before resuspension in an acid-buffered elution buffer (10mM NaOAc, 1mM EDTA, pH 4.8).

#### **rRNA mixture sample preparation**

A mixture of human HEK293T, yeast BY4741 strain, *Drosophila* S2 cells, and *C. elegans* whole animal and stool microbiome total RNA was made by mixing 1 µg RNA from each model organism sample and 8 µg total RNA from a stool microbiome sample. ZYMO RNA Clean & Concentrator-5 (R1013) kit was used on this mixture to remove all small RNAs <200nt. The final sample was eluted with 20 µl RNase-Free H<sub>2</sub>O. The mixture was split into two halves. One half was used for Illumina sequencing (see below). For nanopore sequencing, the other half was polyadenylated by yeast Poly(A) Polymerase (ThermoFisher 74225Z25KU) by incubation with 0.48 mM ATP, 20 U/µL Poly(A) Polymerase, and 1x Poly(A) Polymerase Reaction Buffer at 37°C for 15 min. The product was size selected using ZYMO RNA Clean & Concentrator-5 (R1013) kit, and RNA molecules >200nt were retained. The sample was eluted with 20 µL RNase-free H<sub>2</sub>O. Then, ~500 ng of this rRNA mixture was used for nanopore direct RNA seq library preparation and nanopore direct RNA sequencing described below.

#### **rRNA mixture Illumina sequencing and mapping**

For Illumina sequencing, bisulfite treatment was performed as described previously [8]. Ψ modification was identified through the deletion at the Ψ site in the sequencing data. Raw reads were demultiplexed via a 4nt barcode on read 2 using je suite [33] with the following parameters: je demultiplex F1=#read1 F2=\$read2 BF=\$barcode\_key BPOS=BOTH BM=READ\_2 LEN=6:4 O=\$output. Only read 2 from paired-end reads were mapped with bowtie2 (version: bowtie2-2.3.3.1-linux-x86\_64) [34] using the following parameters: bowtie2 -x \$reference -U \$read2 -S \$output -q -p 10 --local --no-unal. Reads were mapped to either a set of rRNA from model organisms or a set of bacterial rRNA reads: rfam family RF02541 (bacterial large subunit) and RF00177 (bacterial small subunit). SAM files from bacterial rRNAs were processed with a custom python script to count the total number of reads mapping to each sequence. Only sequences with >1000 reads were processed further. Model organism rRNA sequences from human (NCBI: NR\_003286.4, NR\_003287.4), yeast (RNACentral: URS00005F2C2D\_559292, URS000061F377\_4932), *C. elegans* (RNACentral: URS00005A42AA\_6239, URS00008C9AB9\_6239), and *Drosophila* (RNACentral: URS000030AF9A\_7227, URS000008C6A9\_7227) to form a reference genome for bowtie mapping. Bowtie2 output “sam” files were converted to sorted bam files with samtools [35]. IGV was used to calculate deletion rates with the following parameters: igvtools [36] count -z 5 -w 1 -e 250 --bases \$input \$output \$reference. Custom python scripts were used to reformat the “wig” file.

#### **Nanopore direct RNA seq library preparation and sequencing**

The library preparation followed the protocol of Direct RNA Sequencing Kit (SQK-RNA002) provided by Oxford Nanopore Technology. Briefly, ~500 ng of Poly(A)<sup>+</sup> RNA sample was used for each run. Each single run contained one biological replicate of one

sample. The RT Adaptor (RTA) was ligated to the 3' end of Poly(A)<sup>+</sup> RNA by T4 DNA ligase (NEB M0202S) and then reverse transcribed by SuperScript III Reverse Transcriptase (ThermoFisher 12574018). The RNA was purified by 1.8x RNAClean XP beads (72  $\mu$ L) (Beckman Coulter A63987) and then the RNA Adaptor (RMX) was ligated to the 3' end of Poly(A)<sup>+</sup> RNA using T4 DNA ligase (NEB M0202S) and then the RNA was purified with 1x RNAClean XP beads (40  $\mu$ L). The sample was eluted with 21  $\mu$ L Elution Buffer. Then, the sample was loaded onto a R9.4.1 flow cell (FLO-MIN106D) in a MinION sequencer. Each flow cell was sequenced for 72 h.

### Nanopore data pre-processing

All raw fast5 files generated during sequencing were uploaded to Midway2 cluster for the following steps. Reads were base called by guppy base caller (version 3.2.2+9fe0a78) with `min_qscore 7`. The reads were aligned to by minimap2 (version 2.18-r1015) [37] with parameters `-ax splice -uf -k14`. The rRNA mixture reads are aligned to the same reference as the rRNA Illumina seq data described above. The human mRNA reads are aligned to the hg38 human genome reference (GRCh38.p13). The mapped reads were piled up to the reference chromosomes by samtools (v1.11). The "error" features were extracted from the mpileup files by customized python scripts ([https://github.com/sihaohuanguc/Nanopore\\_psU](https://github.com/sihaohuanguc/Nanopore_psU)).

### Model training

For nanopore seq data of rRNA, all sites mapped to "T" in the reference with >20 coverage made up the data pool. 640  $\Psi$  sites revealed by Illumina sequencing and 689 randomly selected U sites from the data pool made up the model training dataset. The dataset was divided into 60% training set, 20% validation set, and 20% testing set. The  $\Psi$  modification prediction models were generated by training set and validated with the validation set by extremely randomized trees (EXT) models with 1–12 features and customized parameters. Then, the models were applied to predict  $\Psi$  modification probabilities of the testing set and evaluated by AUC of ROC (receiver operating characteristic) curves derived from the predicted probabilities of the testing set. The final model used EXT algorithm (`n_estimators=200`, `criterion="gini"`, `max_depth=None`, `min_samples_split=2`) with 12 features.

### HeLa cell culture and interferon treatment

HeLa cells (ATCC, authenticated and tested for mycoplasma contamination) were cultured in the presence of 500 U/mL human interferon gamma (IFN  $\gamma$ , Peprotech), 500 U/mL human interferon beta (IFN  $\beta$ , Peprotech), or left untreated, with biological duplicates for each. Cells were incubated for 24 h, and an aliquot of each was processed for flow cytometry. Cells were washed into a flow cytometry staining buffer (FBS-containing RPMI and Hanks' Balanced Salt Solution) containing the anti-pan-MHC-I antibody W6/32 (BioXcell) conjugated with AlexaFluor 647 (Invitrogen). Cells were then washed 3 $\times$  and analyzed by a Fortessa X-20 (BD Biosciences) to determine upregulation of MHC class I. The rest of the cells were used for RNA extraction via the RNeasy Mini kit (Qiagen) following the manufacturer's protocol. RNA was eluted in pure water and quantified by Nanodrop (Thermo). PolyA<sup>+</sup> RNA from 50  $\mu$ g total RNA of each sample

was extracted by Promega PolyATtract® mRNA Isolation Systems Z5310. Each sample was eluted with 15  $\mu$ L H<sub>2</sub>O.

### Prediction of $\Psi$ in HeLa samples

The raw data of two replicates for the untreated, IFN  $\gamma$ -treated, and IFN  $\beta$ -treated samples were merged after aligned to the hg38 human genome reference (GRCh38.p13). The merged samples were down sampled so that they have almost the same number of reads and are directly comparable. The  $\Psi$  modification probabilities of all sites mapped to “T” in the reference with over 20 coverage were evaluated by the EXT model generated with the rRNA mixture sample. The coverage independence of  $\Psi$  probability was examined by down sampling all sites of the samples to similar coverages (expectation = 30) using different random seeds. We found that the change in mean  $\Psi$  probability of the transcripts maintained the same after down sampling. The coverage completeness of the transcripts was checked by counting the U sites predicted in the samples [38]. For the untreated sample, the U sites within 5'UTR, CDS, and 3'UTR represented 2.43%, 42.84%, and 54.73% of all U sites, respectively. The gene information was provided by the comprehensive gene annotation file (gencode.v37.annotation.gff3) in the GENCODE database (<https://www.gencodegenes.org>) [39]. The gene ontology (GO) analysis was performed using the Gene Ontology Resource (<http://geneontology.org>) [40, 41]. The sequence logo plots were generated by MEME (<https://meme-suite.org/meme/tools/meme>) [42].

### CMC-mediated RT-qPCR (CRP) validation of $\Psi$ level in mRNA transcripts

#### Primer design

qPCR primers were designed using NCBI Primer-BLAST tool (<https://www.ncbi.nlm.nih.gov/tools/primer-blast/>). Two to 3 sets of primers were selected to cover the 3' end, middle, and 5' end region of the whole transcript. qPCR was performed with Taq-Man style fluorescent probes. Probes for each PCR primer pair were designed using IDT PrimerQuest tool (<https://www.idtdna.com/pages/tools/primerquest>) and examined using NCBI nucleotide BLAST. Primers and probes were purchased from IDT. Actin (NM\_001101.5) and ISG15 (NM\_005101.4) transcripts were selected for  $\Psi$  validation. Below is the list of the sequences of qPCR primers and probes.

ISG15 primer1-Forward: GTGGACAAATGCGACGAACC

ISG15 primer1-Reverse: ATTTCCGGCCCTTGATCCTG

ISG15 probe1: 5'-/56-FAM/TCC TGG TGA/ZEN/GGA ATA ACA AGG GCC/3IABkFQ/-3'

ISG15 primer2-Forward: GCGCAGATCACCCAGAAGAT

ISG15 primer2-Reverse: GTTCGTCGCATTTGTCCACC

ISG15 probe2: 5'-/56-FAM/TTC CAG CAG/ZEN/CGT CTG GCT GT/3IABkFQ/-3'

ISG15 primer3-Forward: CAGCGAACTCATCTTTGCCAG

ISG15 primer3-Reverse: GACACCTGGAATTCGTTGCC

ISG15 probe3: 5'-/56-FAM/TGG GAC CTG/ZEN/ACG GTG AAG ATG C/3IABkFQ/-3'

ACTB primer1-Forward: ACAGGAAGTCCCTTGCCATC

ACTB primer1-Reverse: CAGTGTACAGGTAAGCCCTGG

ACTB probe1:5'-/56-FAM/ACA CGA AAG/ZEN/CAATGCTATCACCTCCC/  
31ABkFQ/-3'

ACTB primer2-Forward: AGATGTGGATCAGCAAGCAGG

ACTB primer2-Reverse: GGGGGATGCTCGCTCCA

ACTB probe2: 5'-/56-FAM/TCG TCC ACC/ZEN/GCA AAT GCT TCT AGG/  
31ABkFQ/-3'

#### CMC-mediated RT-qPCR (CRP) experiment

CMC [N-cyclohexyl-N'-(2-morpholinoethyl) carbodiimide] treatment was done as previously described [43]. 1.5 µg of untreated, IFNβ-treated, and IFNγ-treated total RNA in 12 µl was mixed with 24 µl TEU buffer (50 mM Tris-HCl (pH 8.3), 4 mM EDTA, 7 M urea) in microcentrifuge tubes. Four microliters of freshly made 1 M CMC (Sigma, C1011) in TEU buffer or 4 µl TEU buffer was added to each sample for +CMC or -CMC treatment, respectively. The sample mixture in 40 µl 0.7× TEU was incubated at 37°C for 1 h. The mixture was diluted to 200 µl with 160 µl of 50 mM KOAc (pH 7) and 200 mM KCl. One microliter of 5 µg/µl glycogen and 550 µl ethanol were added to the mixture to precipitate RNA at -80°C for >2 h. The mixture was then centrifuged at highest speed (17000×g) for 30 min. The RNA precipitate was mixed with 500 µl 75% ethanol and kept at -80°C for >2 h followed by centrifugation at 17000×g for 30 min. The washing step was repeated once. The RNA precipitate was mixed with 50 µl of 50 mM Na<sub>2</sub>CO<sub>3</sub> and 2 mM EDTA (pH 10.4) and incubated at 37°C for 6 h to remove CMC-U/CMC-G adducts. The RNA was purified using Zymo RNA Clean and Concentrator column (Zymo, R1014) with in-column DNase treatment by following the manufacturer's manual. The RNA was eluted in 11 µl sterile H<sub>2</sub>O. The concentration of the ±CMC treated RNA was measured using Nanodrop, and equal amount (~300 ng) of total RNA was used for RT-qPCR experiment.

Eleven microliters of 300 ng ±CMC-treated total RNA from untreated/IFNβ/IFNγ samples were mixed with 1 µl 50 µM 5'T<sub>22</sub>VN (V=A,C,G, N=A,C,G,T) primer (IDT) and 1 µl 10 mM dNTP mix. The mixtures were incubated at 65°C in thermal cycler for 5 mins followed by incubation at room temperature for 3 min. The PCR tubes were kept on ice until the addition of the SuperScript IV RT mix. 7 µl RT mix was prepared for each sample by combining 4 µl 5× SSIV Buffer, 1 µl 100 mM DTT, 1 µl RNaseOUT RNase inhibitor, and 1 µl SSIV reverse transcriptase. 7 µl RT mix was added to each PCR tube. The tubes were incubated at 55°C in thermal cycler for 1.5 h. The PCR tubes were then incubated at 80°C for 10 min followed by incubation on ice immediately to deactivate RT. 45 µl sterile H<sub>2</sub>O was added to each tube to dilute the RT mixture to 65 µl, and 2 µl was used for qPCR reaction.

qPCR reaction was performed in 10 µl consisting of 5 µl 2× PrimeTime Gene Expression Master Mix (IDT, 1055772), 2 µl RT mix, and 3 µl primer and probe mix. Three microliters of primer and probe mix (1.5 µM each PCR primer and 0.6 µM probe) was first added into each well of 384-well plate or 96-well plate. RT mix of each sample and 2× PrimeTime Gene Expression Master Mix were mixed at 2:5 ratio to make master mix based on the number of qPCR reactions for each sample. Seven microliters of the template and PrimeTime master mix were then added to each well. The plate was spun on a swing bucket plate centrifuge at 3000 RPM for 2 min. qPCR reaction was

performed on Bio-Rad CFX384 or CFX96 qPCR machine for 40 cycles.  $C_q/C_T$  values were obtained for follow-up data analysis.

Relative  $\Psi$  levels for ISG15 transcript was calculated using ACTB-1 as internal reference. First, we obtained  $\Delta C_q(-) = C_q(\text{ISG15,-CMC}) - C_q(\text{ACTB,-CMC})$ , and  $\Delta C_q(+) = C_q(\text{ISG15,+CMC}) - C_q(\text{ACTB,+CMC})$ ; then, we obtained  $\Delta\Delta C_q(\text{ISG15}) = \Delta C_q(+) - \Delta C_q(-)$ . The relative  $\Psi$  level is represented as  $2^{\Delta\Delta C_q(\text{ISG15})}$ .

### Single read $\Psi$ prediction model training

The 100% modified human rRNA sites were reported in a previously work measured by quantitative LC/MS [31]. A basic assumption was that all reads in our human rRNA sample would have  $\Psi$  at the reported 100% modified sites and U at the reported completely unmodified sites. The dataset for training contained 25 100%  $\Psi$  sites with 49,437 data points and 26 randomly selected U sites with 50,922 data points. The dataset was divided into 60% training set, 20% validation set, and 20% testing set. Features were extracted from each base in each read. The features describing the ratios in bulk prediction model were replaced with features indicating the mismatching and indel states of the base. The  $\Psi$  modification prediction models were generated by training set and validated with the validation set using the EXT algorithm ( $n_{\text{estimators}}=200$ , criterion= "gini", max\_depth=None, min\_samples\_split=2) with 10 features, which are insertion\_ot\_not, insertion\_length, deletion\_or\_not, deletion\_length, deleted\_site\_or\_not, mismatch\_or\_not, mutate\_to\_A, mutate\_to\_C, mutate\_to\_G, base quality score. The AUC value for the prediction of testing set was 0.8269. To further evaluate the model,  $\Psi$  modification probabilities of data points from 22 previously reported [31], partially modified human rRNA  $\Psi$  sites (modification fraction from 5 to 85%) were predicted. The base was viewed as  $\Psi$  when the probability was larger than 0.5 and as U when the probability was less than 0.5. The stoichiometry of each site was calculated as the number of predicted  $\Psi$  bases divided by the coverage of the site.

### Single read $\Psi$ analysis in HeLa samples

The  $\Psi$  probabilities of all U residues in selected genes were predicted with the protocol above. To investigate the linkage of multiple  $\Psi$  on single reads, each read was indexed so that the U data points with the same read index were from the same read.  $\Psi$  probabilities of residues of a certain site were fitted by Gaussian mixture model (GMM) with 2 components. The sites with  $\text{abs}(\mu_1 - \mu_2) > 0.5$  and  $\lambda_1$  and  $\lambda_2 > 0.05$  were selected for following analysis. When doing pair wise linkage analysis, the reads were assigned into " $\Psi$ " and "U" groups when it had >95% posterior probability for one population in the GMM for site 1. To evaluate whether there was a difference in the  $\Psi$  probabilities distribution of site 2 upon the presence or absence of  $\Psi$  at site 1, two sample Kolmogorov-Smirnov test was performed on the  $\Psi$  probabilities cumulative distribution curves of site 2 in the " $\Psi$ " and "U" groups with an output of the maximum distance  $D$  value and  $p$  value. The R library to do a two-sample Kolmogorov-Smirnov test was from GitHub ([https://rdr.io/github/happyrabbit/DataScienceR/man/pairwise\\_ks\\_test.html](https://rdr.io/github/happyrabbit/DataScienceR/man/pairwise_ks_test.html)).

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-021-02557-y>.

**Additional file 1.** Supplementary figures. Supplementary figures S1–S4.

**Additional file 2.**  $\Psi$  sites by Illumina seq. It's the list of  $\Psi$  sites in the rRNA mixture sample identified by our Illumina sequencing experiment.

**Additional file 3.** Review history.

#### Acknowledgements

We thank Jordan Brown and Dr. Heng-Chi Lee for the *C. elegans* total RNA.

#### Peer review information

Barbara Cheifet was the primary editor of this article and managed the editorial process and peer review in collaboration with the rest of the editorial team.

#### Review history

The review history is available as Additional file 3.

Authors' information

Twitter handle: @merenbey (A. Murat Eren)

#### Authors' contributions

S.H. performed all nanopore experiments and analyzed all nanopore data with guidance from A.M.E., W.Z., and C.D.K. designed the validation by RT-qPCR method. W.Z. performed the validation experiment. C.D.K. analyzed Illumina sequencing data and helped with nanopore data analysis. D.D. and J.Y. designed interferon experiment, and D.D. performed interferon experiment. W.Z. and Q.D. built the Illumina sequencing libraries. K.L. isolated total RNA from stools. S.H. and T.P. conceived the project, designed the experiments, and wrote the paper. The author(s) read and approved the final manuscript.

#### Funding

This work was supported by the grant from the NIH (RM1 HG008935 to T.P.). D.D. and J.W.Y. are supported by the Division of Intramural Research, NIAID. A.M.E. and K.L. were supported by an NIH NIDDK grant (RC2 DK122394).

#### Availability of data and materials

The datasets generated and analyzed during the current study are available in the NCBI GEO database under the accession GSE180656 [44]. The scripts for "NanoPsu"  $\Psi$  prediction package are available on GitHub [45] ([https://github.com/sihaohuanguc/Nanopore\\_psu](https://github.com/sihaohuanguc/Nanopore_psu), GNU General Public License v2.0) or Zenodo [46] (DOI: 10.5281/zenodo.5711328).

#### Declarations

##### Ethics approval and consent to participate

This study was reviewed and approved by the University of Chicago Ethics Committee and by the University of Chicago Institutional Review Board (IRB18-1539). Written and informed consent was obtained for all participants.

##### Consent for publication

N/A

##### Competing interests

The authors declare that they have no competing interests.

##### Author details

<sup>1</sup>Department of Biochemistry & Molecular Biology, University of Chicago, Chicago, IL 60637, USA. <sup>2</sup>Laboratory of Viral Diseases, National Institute of Allergy and Infectious Diseases, NIH, Bethesda, MD 20892, USA. <sup>3</sup>Department of Chemistry, University of Chicago, Chicago, IL 60637, USA. <sup>4</sup>Department of Medicine, University of Chicago, Chicago, IL 60637, USA.

Received: 9 August 2021 Accepted: 23 November 2021

Published online: 06 December 2021

#### References

1. Li X, Zhu P, Ma S, Song J, Bai J, Sun F, et al. Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. *Nat Chem Biol*. 2015;11(8):592–7. <https://doi.org/10.1038/nchembio.1836>.
2. Karikó K, Muramatsu H, Welsh FA, Ludwig J, Kato H, Akira S, et al. Incorporation of pseudouridine into mRNA yields superior nonimmunogenic vector with increased translational capacity and biological stability. *Mol Ther*. 2008;16(11):1833–40. <https://doi.org/10.1038/mt.2008.200>.
3. Anderson BR, Muramatsu H, Nallagatla SR, Bevilacqua PC, Sansing LH, Weissman D, et al. Incorporation of pseudouridine into mRNA enhances translation by diminishing PKR activation. *Nucleic Acids Res*. 2010;38(17):5884–92. <https://doi.org/10.1093/nar/gkq347>.
4. Jackson LA, Anderson EJ, Roupheal NG, Roberts PC, Makhene M, Coler RN, et al. An mRNA vaccine against SARS-CoV-2—preliminary report. *New England J Med*. 2020;383(20):1920–31. <https://doi.org/10.1056/NEJMoa2022483>.
5. Carlile TM, Rojas-Duran MF, Zinshteyn B, Shin H, Bartoli KM, Gilbert WV. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature*. 2014;515(7525):143–6. <https://doi.org/10.1038/nature13802>.

6. Schwartz S, Bernstein DA, Mumbach MR, Jovanovic M, Herbst RH, León-Ricardo BX, et al. Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell*. 2014;159(1):148–62. <https://doi.org/10.1016/j.cell.2014.08.028>.
7. Zhou KI, Clark WC, Pan DW, Eckwahl MJ, Dai Q, Pan T. Pseudouridines have context-dependent mutation and stop rates in high-throughput sequencing. *RNA Biol*. 2018;15(7):892–900. <https://doi.org/10.1080/15476286.2018.1462654>.
8. Khoddami V, Yerra A, Mosbrugger TL, Fleming AM, Burrows CJ, Cairns BR. Transcriptome-wide profiling of multiple RNA modifications simultaneously at single-base resolution. *Proc Natl Acad Sci*. 2019;116(14):6784–9. <https://doi.org/10.1073/pnas.1817334116>.
9. Li F, Guo X, Jin P, Chen J, Xiang D, Song J, et al. Porpoise: a new approach for accurate prediction of RNA pseudouridine sites. *Brief Bioinform*. 2021;22(6). <https://doi.org/10.1093/bib/bbab245>.
10. Salem DH, Acevedo D, Daulatabad SV, Mir Q, Janga SC. Penguin: a tool for predicting pseudouridine sites in direct RNA nanopore sequencing data. *bioRxiv*. 2021. <https://doi.org/10.1101/2021.03.31.437901>.
11. Li Y-H, Zhang G, Cui Q. PPU: a web server to predict PUS-specific pseudouridine sites. *Bioinformatics*. 2015;31(20):3362–4. <https://doi.org/10.1093/bioinformatics/btv366>.
12. Chen W, Tang H, Ye J, Lin H, Chou K-C. iRNA-PseU: identifying RNA pseudouridine sites. *Mol Ther Nucleic Acids*. 2016;5:e332. <https://doi.org/10.1038/mtna.2016.37>.
13. He J, Fang T, Zhang Z, Huang B, Zhu X, Xiong Y. PseUI: pseudouridine sites identification based on RNA sequence information. *BMC Bioinform*. 2018;19(1):1–11. <https://doi.org/10.1186/s12859-018-2321-0>.
14. Tahir M, Tayara H, Chong KT. iPseU-CNN: identifying RNA pseudouridine sites using convolutional neural networks. *Mol Ther Nucleic Acids*. 2019;16:463–70. <https://doi.org/10.1016/j.omtn.2019.03.010>.
15. Liu K, Chen W, Lin H. XG-PseU: an eXtreme gradient boosting based method for identifying pseudouridine sites. *Mol Genet Genom*. 2020;295(1):13–21. <https://doi.org/10.1007/s00438-019-01600-9>.
16. Bi Y, Jin D, Jia C. EnsemPseU: identifying pseudouridine sites with an ensemble approach. *IEEE Access*. 2020;8:79376–82. <https://doi.org/10.1109/ACCESS.2020.2989469>.
17. Lv Z, Zhang J, Ding H, Zou Q. RF-PseU: a random forest predictor for RNA pseudouridine sites. *Front Bioeng Biotechnol*. 2020;8:134. <https://doi.org/10.3389/fbioe.2020.00134>.
18. Khan SM, He F, Wang D, Chen Y, Xu D. MU-PseUDeep: a deep learning method for prediction of pseudouridine sites. *Comput Struct Biotechnol J*. 2020;18:1877–83. <https://doi.org/10.1016/j.csbj.2020.07.010>.
19. Song B, Tang Y, Wei Z, Liu G, Su J, Meng J, et al. PIANO: a web server for pseudouridine-site ( $\Psi$ ) identification and functional annotation. *Front Genet*. 2020;11:88. <https://doi.org/10.3389/fgene.2020.00088>.
20. Song B, Chen K, Tang Y, Ma J, Meng J, Wei Z. PSI-MOUSE: predicting mouse pseudouridine sites from sequence and genome-derived features. *Evol Bioinform*. 2020;16:1176934320925752. <https://doi.org/10.1177/1176934320925752>.
21. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods*. 2018;15(3):201–6. <https://doi.org/10.1038/nmeth.4577>.
22. Liu H, Begik O, Lucas MC, Ramirez JM, Mason CE, Wiener D, et al. Accurate detection of m<sup>6</sup>A RNA modifications in native RNA sequences. *Nat Comm*. 2019;10(1):1–9. <https://doi.org/10.1038/s41467-019-11713-9>.
23. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, et al. Nanopore native RNA sequencing of a human poly (A) transcriptome. *Nat Methods*. 2019;16(12):1297–305. <https://doi.org/10.1038/s41592-019-0617-2>.
24. Drexler HL, Choquet K, Churchman LS. Splicing kinetics and coordination revealed by direct nascent RNA sequencing through nanopores. *Mol Cell*. 2020;77:985–98. e988. <https://doi.org/10.1016/j.molcel.2019.11.017>.
25. Lorenz DA, Sathé S, Einstein JM, Yeo GW. Direct RNA sequencing enables m<sup>6</sup>A detection in endogenous transcript isoforms at base-specific resolution. *RNA*. 2020;26(1):19–28. <https://doi.org/10.1261/ma.072785.119>.
26. Fleming AM, Mathewson NJ, Howpay Manage SA, Burrows CJ. Nanopore dwell time analysis permits sequencing and conformational assignment of pseudouridine in SARS-CoV-2. *ACS Central Sci*. 2021;7(10):1707–17. <https://doi.org/10.1021/acscentsci.1c00788>.
27. Begik O, Lucas MC, Prysacz LP, Ramirez JM, Medina R, Milenkovic I, et al. Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat Biotechnol*. 2021;39(10):1–14. <https://doi.org/10.1038/s41587-021-00915-6>.
28. Jenjaroenpun P, Wongsurawat T, Wadley TD, Wassenaar TM, Liu J, Dai Q, et al. Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Res*. 2021;49(2):e7. <https://doi.org/10.1093/nar/gkaa620>.
29. Lee AJ, Ashkar AA. The dual nature of type I and type II interferons. *Front Immunol*. 2018;9:2061. <https://doi.org/10.3389/fimmu.2018.02061>.
30. Safra M, Nir R, Farouq D, Slutskin IV, Schwartz S. TRUB1 is the predominant pseudouridine synthase acting on mammalian mRNA via a predictable and conserved code. *Genome Res*. 2017;27(3):393–406. <https://doi.org/10.1101/gr.207613.116>.
31. Taoka M, Nobe Y, Yamaki Y, Sato K, Ishikawa H, Izumikawa K, et al. Landscape of the complete RNA chemical modifications in the human 80S ribosome. *Nucleic Acids Res*. 2018;46(18):9289–98. <https://doi.org/10.1093/nar/gky811>.
32. Borhardt EK, Martinez NM, Gilbert WW. Regulation and function of RNA pseudouridylation in human cells. *Ann Rev Genet*. 2020;54(1):309–36. <https://doi.org/10.1146/annurev-genet-112618-043830>.
33. Girardot C, Scholtalbers J, Sauer S, Su S-Y, Furlong EE. Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. *BMC Bioinform*. 2016;17(1):1–6. <https://doi.org/10.1186/s12859-016-1284-2>.
34. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature methods*. 2012;9(4):357–9. <https://doi.org/10.1038/nmeth.1923>.
35. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9. <https://doi.org/10.1093/bioinformatics/btp352>.
36. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24–6. <https://doi.org/10.1038/nbt.1754>.
37. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094–100. <https://doi.org/10.1093/bioinformatics/bty191>.
38. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–2. <https://doi.org/10.1093/bioinformatics/btq033>.

39. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, et al. GENCODE 2021. *Nucleic Acids Res.* 2021; 49(D1):D916–23. <https://doi.org/10.1093/nar/gkaa1087>.
40. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. *Nat Genet.* 2000;25(1):25–9. <https://doi.org/10.1038/75556>.
41. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research.* 2021;49(D1): D325–34. <https://doi.org/10.1093/nar/gkaa1113>.
42. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. *Nucleic Acids Res.* 2015;43(W1):W39–49. <https://doi.org/10.1093/nar/gkv416>.
43. Zhang W, Eckwahl MJ, Zhou KI, Pan T. Sensitive and quantitative probing of pseudouridine modification in mRNA and long noncoding RNA. *Rna.* 2019;25(9):1218–25. <https://doi.org/10.1261/rna.072124.119>.
44. Huang S, Zhang W, Katanski CD, Dersh D, Dai Q, Lolans K, Yewdell J, Eran AM, Pan T. Interferon inducible pseudouridine modification in human mRNA by quantitative nanopore profiling. GSE180656. *Gene Expression Omnibus.* <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE180656> (2021).
45. Huang S, Zhang W, Katanski CD, Dersh D, Dai Q, Lolans K, Yewdell J, Eran AM, Pan T. Nanopore\_psU. Github. [https://github.com/sihaohuanguc/Nanopore\\_psU](https://github.com/sihaohuanguc/Nanopore_psU) (2021)
46. Huang S, Zhang W, Katanski CD, Dersh D, Dai Q, Lolans K, Yewdell J, Eran AM, Pan T. Nanopore\_psU. <https://zenodo.org/record/5711328#.YZaoBy1h2Tc> (2021)

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

